Information Quality Strategy - an Empirical Investigation of the Relationship Between Information Quality Improvements and Organizational Outcomes

THESIS

Submitted in partial fulfillment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

by

Arun T Sundararaman

ID NO. 2006PHXF002

Under the Supervision of

Dr. Peter Z Yeh

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

PILANI (RAJASTHAN) INDIA

2012

Abstract

The topic of Data Quality in the field of Information Management has been extensively researched; within this field, relationship between Data Quality and organization outcomes has been relatively less explored. Data Quality plays an important role in decision making and has a potential to impact the quality of business outcomes resulting from decisions based on analysis of data. Data Quality research as it relates to the role of Data Quality in Decision Support Systems has been relatively less explored. Thus, research analyzing the impact of DQ on decision making is limited. Even the limited studies in this subject have been focused on select set of factors (e.g. timeliness or accuracy or consistency etc.) that contribute to overall data quality (which is essential), but, have not approached Data Quality comprehensively covering all applicable factors. Also, conventionally, the study of Data Quality has not considered the need to adopt different measurement approaches based on context of the decision tasks e.g. by industry or by function. For example, "data accuracy" as a factor assumes higher significance for "fraud detection" related decision in a financial industry, whereas the same factor may assume relatively lesser significance for "tactical" decisions in a telecom industry. Past literature acknowledges that DQ is multi-dimensional, yet, the measurement of DQ has not been multi-dimensional.

In this work, the author presents an approach that addresses these problems by introducing a framework for measurement for Data Quality as it pertains to Decision Support Systems (DSS) and links this measure to the quality of business outcomes. This framework approaches Data Quality measurement both comprehensively and in the context of business decisions, through the concept of **decision categories**. This research sets forth the hypothesis of existence of relationship between data quality and quality of business outcome; further, this work introduces a model to measure data quality based on factor that influence the quality of such outcomes. As part of this work, a new research model was proposed which is evolved from "Methodology for the Quality Assessment of Financial Data". Based on the suggestions of TIQM Methodology, 4 new DQ factors applicable for DSS (granularity, relevance of dimensions, relevance of measures and aggregation) were identified as part of this work. Empirical study designed for the purpose of this study comprised 3 stages viz.,

- Stage 1 - Calibration of the model by industry experts (assign weightage for DQ factors for each decision category)

- Stage 2 – DQ measurement using the model by end users (measurement of confidence in DQ)
- Stage 3 - Validation of model outcomes (independent assessment of DQ and comparison of the same with the model output).

. Experts in Banking and end of users of DSS in Banks were actively engaged for providing data for the empirical study. Data for this study was collected by administering 3 different questionnaire in sequence, reaching out to experts and practitioners in the Banking industry; data analysis revealed existence of relationship between data quality and business outcome; findings further established that within an organization, different measures of data quality exists for different business requirements. To validate the model outcomes, as the final step (Stage 3), different users of DSS were approached independently to obtain their assessment levels of data quality which served as the gold standard. Outputs produced by the framework were compared to the above gold standard to determine the validity of the framework.

Key outcomes of the research work:

1. A contextual framework for comprehensive and context-aware measurement of DQ in Decision Support Systems
2. A conceptual model for researching the relationship between DQ aspects and organizational outcomes
3. Findings (from empirical study) on relationship between DQ and organization outcomes, in specific context of Decision Support Systems.

Dedication

This Dissertation is dedicated to my parents and elders in our family who have always had great pride and passion in seeing members of the family excelling in education. I would like to thank everyone in my family for the unrelenting support extended by understanding and accommodating their requirements to the odd extra hours and efforts I had to devote to this work.

I would also like to thank my friends, colleagues and well wishers for their continuous encouragement and support to this work. Years back, they provided me the initial inspiration to undertake this Course and provided the much needed confidence that I can do this. Of course, I tried doing this at my own pace – as my focus was in going through the learning experience in doing this research work and not necessarily having to do this racing against time. This approach provided me the opportunity to interact with several learned scholars in different countries and learn from their experiences. I would like to thank everyone (Professors in different Universities or Authors or Experts in the field of Information Management research) who have shared their work with me or provided me guidance, review comments and/or feedback to pursue my research work.

Getting specific, I don't have words or expressions to record my sincere thanks for my Guide, Dr. Peter Z Yeh, who has always provided excellent guidance, coaching and thoughts throughout the research work. I would like to record my sincere thanks to Vandana Naveen and Naveen Sriraman for their support and encouragement in the data collection process which was extremely critical for this work. I would also like to thank the Professors and Dean at Birla Institute of Technology and Science, Pilani, India for their timely guidance and directions on this research.

Table of contents

ABSTRACT **2**

DEDICATION **4**

TABLE OF CONTENTS **5**

LIST OF TABLES **8**

LIST OF DIAGRAMS **9**

CHAPTER 1 – INTRODUCTION **10**

Background of the Study 12

Statement of the problem 15

Research Questions 17

Nature of the Study 17

Scope of the Study 18

Organization of chapters 19

CHAPTER 2 – LITERATURE SURVEY **22**

Introduction 22

Study of Data Quality Factors 22

Study of Data Quality Assessment frameworks 36

Study of Data Quality in Decision Support Systems 44

Study of Data Quality linked to business outcome 48

Gaps in existing research and research objectives 54

CHAPTER 3 – RESEARCH METHODOLOGY **60**

Introduction 60

Research philosophy 60

Research framework 61

Selection of DQ factors 63

Research Methodology 76

Research approach – DQ Assessment Framework 78

Research problem and hypothesis 85

CHAPTER 4 – EMPIRICAL STUDY DESIGN **87**

Introduction 87

Study Design 87

Sampling methodology 91

Study set-up activities 97

Survey Instruments Design 109
Empirical Study Design: Setup Questionnaire *110*
Empirical Study Design: Stage 1 – Model Calibration *112*
Empirical Study Design: Stage 2 – Measure DQ *114*
Empirical Study Design: Stage 3 - Validation *114*
Statistical methods and instruments used in the study 114
Kappa Statistics *115*
Chi-Squared Test *119*
Friedman Tests *120*

CHAPTER 5 – RESULTS AND ANALYSIS 123

Introduction 123
Results of empirical study 123
Survey Results 127
Analysis of results 133

CHAPTER 6 – FINDINGS AND RECOMMENDATIONS 136

Introduction 136
Research findings 136
Uniqueness of this work 138
Recommended use of findings 141

CHAPTER 7 – DISCUSSIONS 143

Introduction 143
DQ – an area of enduring research 143
DQ in relation to Banking 144
DQ related to unstructured data 146
Decision support Vs. decision making 147

CHAPTER 8 – CONCLUSIONS AND FUTURE WORK 149

Conclusions 149
Limitations of current work 151
Recommendations for future work 152

REFERENCES 154

APPENDICES 166

Appendix 1 :: Preliminary Survey Instrument 166
Appendix 2 :: Survey Instrument 2 171
Appendix 3 :: Survey Instrument 3A 180

APPENDIX 4 :: SURVEY INSTRUMENT 3B 187
APPENDIX 5 :: DATA QUALITY FACTORS – DEFINITIONS 196
APPENDIX 6 :: LIST OF PUBLICATIONS AND PRESENTATIONS 198
3rd International Conference on Trendz in Information Sciences *198*
5th India Software Engineering Conference *199*
8th National Conference on Medical Informatics *200*
APPENDIX 7 :: PROFILE OF SUPERVISOR / RESEARCH GUIDE 201
Research Interests *201*
Education *201*
Professional Experience *202*
Teaching Experience *205*
Publications *205*
Journal and Book Chapter 205
Conference, Workshop, and Symposium 205
Thesis and Technical Report 208
Other 208
Professional Services and Activities *208*
Program Committee 208
Reviewer 209
Member 209
Awards and Honors *209*
APPENDIX 8 :: PROFILE OF AUTHOR 210
Research Interests *210*
Education and Professional Membership *210*
Professional Experience *210*
Publications and professional development activities *211*

List of Tables

TABLE I. DEFINITION OF TERMS AND ACRONYMS 20
TABLE II. TOPICS AND METHODS OF DQ RESEARCH 24
TABLE III. COMPARISON OF EXISTING INFORMATION QUALITY FRAMEWORKS AND DQ FACTORS 28
TABLE IV. FREQUENTLY REFERRED DQ FACTORS 32
TABLE V. EXISTING DQ ASSESSMENT METHODOLOGIES 38
TABLE VI. STUDY OF SELECTED DQ FACTORS 56
TABLE VII. RECOMMENDED AREAS OF FURTHER RESEARCH FROM LITERATURE 61
TABLE VIII. DQ FRAMEWORK FROM LITERATURE 63
TABLE IX. COMPARISON OF INFORMATION QUALITY FRAMEWORKS 65
TABLE X. DQ FRAMEWORKS CONSIDERED FOR THE RESEARCH WORK 69
TABLE XI. DQ FACTORS FROM NAUMANN AND ROLKER FRAME WORK 72
TABLE XII. DQ FACTORS CONSIDERED FOR THE RESEARCH WORK 75
TABLE XIII. QAFD STEPS AND THEIR ADOPTION FOR THIS WORK 76
TABLE XIV. BANKS COSNIDERED FOR DATA COLLECTION 93
TABLE XV. DECISION CATEGORIES IDENTIFIED 102
TABLE XVI. HYPOTHESIS OF THE RESEARCH WORK 102
TABLE XVII. DECISION CATEGORIES – DQ FACTORS MATRIX 105
TABLE XVIII. OBSERVATIONS FROM MAPPING EXERCISE 106
TABLE XIX. STAGES OF DATA COLLECTION 111
TABLE XX. KAPPA STATISTICS STUDY SET UP 118
TABLE XXI. DQ FACTORS FOR DSS AND THEIR DEFINITIONS 123
TABLE XXII. SETUP DQ SCORES 126
TABLE XXIII. DQ MEASUREMENT REQUIREMENTS 126
TABLE XXIV. DQ SCORE COMPUTED USING THE FRAMEWORK FOR CREDIT DECISION CATEGORY 128
TABLE XXV. DQ SCORE COMPUTED USING THE FRAMEWORK FOR BUSINESS PROMOTION DECISION CATEGORY 128
TABLE XXVI. DQ SCORE COMPUTED USING THE FRAMEWORK FOR PRODUCT DECISION CATEGORY 129
TABLE XXVII. DQ SCORE COMPUTED USING THE FRAMEWORK FOR TACTICAL DECISION CATEGORY 129
TABLE XXVIII. DQ SCORE COMPUTED USING THE FRAMEWORK FOR RELATIONSHIP DECISION CATEGORY 130
TABLE XXIX. DQ SCORE COMPUTED USING THE FRAMEWORK FOR REGULATORY DECISION CATEGORY 130
TABLE XXX. SUMMARY OF SURVEY RESULTS 131
TABLE XXXI. CHI-SQUARE TEST RESULTS 132
TABLE XXXII. KAPPA STATISTICAL ANALYSIS 134
TABLE XXXIII. ADDITIONAL TEST RESULTS 135
TABLE XXXIV. UNIQUENESS OF THIS WORK 138

List of Diagrams

DIAGRAM I. DQ Requirements analysis process 26
DIAGRAM II. Main issues in data quality 35
DIAGRAM III. Organization of SQuaRE series of ISO standards 36
DIAGRAM IV. Phases of QAFD 40
DIAGRAM V. DQ closed loop 42
DIAGRAM VI. Business oriented DQ metrics 51
DIAGRAM VII. Dr. slone's Research model - study of information quality 54
DIAGRAM VIII. Step 1: Identify questions 78
DIAGRAM IX. Step 2: identify decision categories 79
DIAGRAM X. Step 3: validation by experts 79
DIAGRAM XI. Step 4: DQ factor-Decision category matrix 80
DIAGRAM XII. Step 5: Calibration 80
DIAGRAM XIII. Step 6: DQ Measurement 81
DIAGRAM XIV. Step 7: validate DQ score 82
DIAGRAM XV. DQ Assessment framework 84
DIAGRAM XVI. research problem and hypothesis 88
DIAGRAM XVII. design of empirical study 90
DIAGRAM XVIII. pre study set-up steps 97
DIAGRAM XIX. study results data analysis - Chi-Square test 120
DIAGRAM XX. uniqueness of this research work 141

Chapter 1 – Introduction

Data Quality (DQ) has serious consequences, of far reaching significance, for the efficiency and effectiveness of organizations and businesses (Batini et. al., 2006). To appreciate the magnitude and significance of DQ, it is worth referring to the report of the Data Warehousing Institute (Eckerson, W., 2002), which estimates that DQ problems cost U.S. businesses more than 600 billion dollars a year. There are several examples to substantiate the above estimates, from the corporate world that signify the magnitude of potential impact to business due to poor DQ. A few examples are listed below:

- DQ problems in 2008 at British Gas had multiple impacts – caused a project to fail, lost revenues estimated around £180M and resulted in degraded relationships with customers.
- In December 2011, poor data quality issues cost the Irish National Lottery € 5 MN.
- Banking group Royal Bank of Scotland reported disastrous data quality issues in June 2012 and has attributed an initial loss of £125 million and may lose more in future. (the incident was triggered due to daily update through a batch data processing, resulting in a significant backlog of daily data and information processing)

People often tend to consider DQ as synonymous merely with data accuracy. However, DQ is more than simply data accuracy. There are other critical dimensions that characterize DQ. From a research perspective, DQ has been the subject of study under different disciplines such as statistics, management and computer science. Historically, study of DQ traces back to statisticians that were the first to investigate some of the problems related to DQ and were followed by researchers in management who were focused on how to control data manufacturing systems in order to detect, eliminate or control DQ problems, With the advent of computer systems to manage businesses and store data in electronic form, study of DQ was widely adopted in the domain of computer science Computer scientists considered the problem of defining, measuring and improving quality of electronic data stored in databases, transactional systems , data captured through web application or data stored data warehouses for decision support.

Past research in the area of Information Quality (IQ) point out that Data and Information Quality can be conceived as a multi-dimensional concept with varying attributes depending on individual researcher's viewpoint. Most commonly, the term "Data Quality" is described as data that is "Fit-for-use", which implies that it is **relative**, as data considered appropriate for one use may not possess sufficient attributes for another use.

DQ is subjective in nature and therefore if assessed independent of the business objective for which the data is intended to be put to use, it is likely to lead to incorrect results. Just as every metric is defined to provide a definite perspective of any selected subject or as every tool is designed to be used for specific purposes, DQ needs to be defined and approached with purpose in mind; the quality of business outcomes can be a common and relevant purpose in the study of Decision Support Systems. This necessitates development of DQ Assessment Methods that assess and measure DQ. In the following sections, this work proposes a framework for assessment of DQ in Data Warehouse Systems (also referring to Decision Support Systems or any form of MIS / EIS).

Despite a decade of research and practice, only piece-meal techniques are available for measuring, analyzing, and improving DQ in organizations. There are several issues that remain unresolved with respect to DQ and the relationship between DQ and organizational outcomes. Broad objectives of this research work is to establish if evidence exists that the relationship between DQ and organization outcomes is systematically measurable and to identify how various aspects of DQ for a DSS are related to categories of organization outcomes or decision tasks (defined in this work as '***decision categories***'). Please refer discussions in Chapter 2 (page no. 56) for examples on how past research in this subject has dealt only with select set of DQ factors. As a result, organizations are unable to develop comprehensive measures of the underlying DQ and their impact on business outcomes. Research in the area of DQ measurement / assessment, have so far been focused on a select set of DQ factors (e.g. timeliness or correctness); to state differently, the approaches to date have been ***inputs driven***. The limitation with these approaches is the lack of focus on quality of business outcomes and the associated research challenge lies in the need to develop a

framework that is focused on quality of business outcomes that result from the quality of decisions that are derived from use of Information from DSS i.e. development of a model to measure DQ from the quality of business outcomes.

Our research was designed to address these challenges. We developed a model that is ***outcome focused*** as opposed to the existing *inputs driven* assessment methodologies; we developed a framework for measurement of DQ that recognizes the business context for which the data is being put to use (in decision making).

There are several issues that remain unresolved with respect to data quality and the relationship between information quality improvements and organizational outcomes. More and more companies are recognizing that data is a key organizational resource, and all kinds of business data are used increasingly in strategic information systems in decision support. The Data Warehousing Institute estimates that data quality problems cost American businesses more than $600 billion a year. The ability of an organization to make accurate strategic decisions is greatly weakened when the data warehouse contains inaccurate data.
Broadly speaking objectives of this research work is to establish if evidence exists that the relationship between the quality of information and organization outcomes is systematically measurable and to identify how various aspects of information quality for a Decision Support System(such as soundness, usability, reliability and usefulness) are related to categories of organization outcomes (strategic or tactical).

Background of the Study

Several research projects have attempted addressing the problem of assessing information quality criteria. Existing literature in the study of DQ can be logically grouped into 4 categories viz., general study of DQ factors, framework / methodology for assessment / measurement of DQ, study of DQ in specialized systems (e.g. Datawarehouse or CRM) and study of DQ in the context of business decisions / outcomes. This chapter provides a brief overview of representative works from each of these categories below and readers are requested to refer Chapter 2 for more in-depth discussion.

General Study of DQ Factors: The data quality literature provides a thorough classification of data quality dimensions; however, there are a number of discrepancies in the definition of most dimensions due to the contextual nature of quality. Wang et. al., 1995 present an information quality assessment methodology called AIMQ which is designed to help organizations to assess the status of their organizational information quality and monitor their IQ improvements over time. Dr. Slone's work "Information Quality Strategy: an empirical investigation of the relationship between Information Quality improvements and organizational outcomes" (Slone, J., 2006) is largely based on this methodology.

It has long been recognized that data is best described or characterized via multiple attributes, or dimensions. For example, in 1999, Indushobha et. al., identified four dimensions of data quality: accuracy, completeness, consistency, and timeliness (Indushobha et. al., 1999). In 2002, Lee et al., 2002 analyzed the various attributes of DQ from the perspective of the people that use data. They identified a full set of DQ dimensions, adding believability, value added, interpretability, accessibility, and others to the original four attributes. The said work (Lee et al., 2002) proceeded to classify the updated dimensions into four broad categories: intrinsic, contextual, and representational and accessibility. For instance, 'data accuracy' belongs to intrinsic category; completeness and timeliness to contextual; consistency to representational and availability to accessibility.

Framework / Methodology for Assessment / Measurement of DQ: In their recently published work (Heinrich et. al., 2009) on data quality measurement, Heinrich and Kaiser have emphasized the need for adequate measurement (of data quality) since quantifying data quality is essential for planning quality measures in an economic manner. In their work they analyze how data quality can be quantified with respect to particular dimensions and have designed new metrics for the dimensions correctness and timeliness.

Study of DQ in Specialized Systems: Data quality issues are common in data warehouses and administrators are concerned about the usability of this decision environment. Much of the earlier research on data quality has been focused on accuracy (please Chapter 2, page no. 56), however, much more research is needed on other dimensions of Data Quality. One of the

key questions raised in the past has been “How does one determine which quality attributes are appropriate for a given application domain”.

In view of this, a need was identified for the present research as follows:

- To identify data quality attributes appropriate for Decision Support systems
- To evaluate relationship between information quality and organization outcomes, in specific context of DS Systems.

Previous research in this area has been modeled on four specific aspects of information quality (soundness, dependability, usefulness and usability) and two categories of organization outcomes (strategic benefit and transaction benefit).

Study of DQ in the Context of Business Decisions / Outcomes: The relationship between information and decision-making is a complex one and has been the subject of extensive research spanning several decades. Extensive research has been conducted in the area of data quality such as dimensions that constitute data quality and frameworks to measure data quality. Despite these research advances, there has thus far been very little understanding from either a theoretical or practical perspective of the relationship between information quality improvement activities and organizational outcomes. A review of relevant literature revealed few examples of research addressing information quality strategy. Those that were identified were written from a variety of perspectives. However, the field of research related to relationship between data quality and organization outcomes has been found to be limited. It is proposed to extend the existing research work conducted by Dr. Slone (Slone, J., 2006). Elements of this research are discussed on several occasions in subsequent paragraphs of this Thesis. The area of research involved conducting empirical study and statistical analysis of the relationship between information quality in Decision Support Systems and organization outcomes.

This work is aimed at developing a framework to measure DQ in decision support systems, based on the quality of outcomes from business decisions made using such data. The research approach is aimed to be in line with the 2 dimensions of research (topic and method), as listed in TABLE II. in Chapter 2 (page no.24)

TABLE III. in Chapter 2 (page no. 28) summarizes 5 widely accepted DQ Frameworks collated from the various earlier works of DQ research. While varied in their approach and application, the frameworks share a number of characteristics regarding their classifications of the dimensions of quality.

Statement of the problem

The work of Wang et. al, 1995, that introduced a path breaking framework approach for DQ research, had laid down a few principle findings among which is the fundamental need for an overall DQ metric. It is pertinent to note that even in a recent work (Sadiq et. al., 2011) this theme has been reiterated.

The issue of DQ has been existent ever since data existed and the impact of poor DQ on decision making and organizational outcomes has always remained an area of concern and interest. Nature of data and complexity of its sources has been changing rapidly – e.g. data has traversed from structured to unstructured patterns and the volume of data has been growing in leaps and bounds. In modern IT world and economies, for business decision making, corporations depend on a wide variety of external channels of information, in addition to disparate internal source systems. These developments (in nature and sources of data required for business decisions) have increased the complexity of DQ definition and its measurement techniques. In order for the research community to adequately respond to the changing landscape of DQ challenges, a unified framework for DQ research is needed ((Sadiq et. al., 2011).

A review of the literature revealed limited research addressing DQ measurement in a comprehensive manner; those that were based on this topic approached DQ with limited subset of factors comprising DQ. Therefore, a need was identified for the current research work to evolve a framework for measurement of DQ and for linking such a measurement to organizational outcomes.

The author has identified the following areas of gaps in the existing work:

First, literature review of existing work on DQ revealed that work so far has only focused on study of DQ criteria from conventional data requirements (for transaction processing systems or web based applications). They do not take into consideration the unique nature of data required for DSS where data is mostly organized in multi-dimensional form (dimensions and measures) for easy referencing and effective decision making.

Second, Lack of context for use: Data is intended to be used for different business decisions in varied decision making situations. Examples include decisions for inventory reorder or identification of profitable products or decision to cross-sell products to customers or decision on target client segment for launching a campaign etc. DQ needs vary based on nature and criticality of decisions being taken (Singh and Singh, 2010 and Indushobha et.al., 1999). As such DQ measurement should be sensitive to this context of business decisions. However, past research has not captured the context of purpose (put to use and related outcome) while studying the DQ criteria.

Third, relative importance of DQ criteria: The definition of an aggregate DQ measure remains an open issue and the most common approach towards this aggregate DQ measure is to consider all the different DQ dimensions (or factors) and combine them by using weighted sum (Helfert et. al., 2009). DQ is comprised of multiple DQ factors and each of those factors influences the overall DQ measurement differently for different decision scenarios (Keeton et. al., 2010). For example, degree of impact of *accuracy* as a DQ criteria on *regulatory decisions* is different than that on *pricing decisions*. Past research on DQ does not recognize this feature and the current DQ Assessment techniques do not include appropriate factors to consider this phenomenon. .

Fourth, participation of users in the measurement of DQ is essential for an effective DQ assessment mechanism (Slone, J., 2006; Sintchenko et. al., 2007 and European Commission, 2007). Existing DQ measurement approaches do not consider inputs of the end users of data

(such as users' confidence in data or users' experience with decision quality etc.,) in such assessment process.

Research Questions

The objectives of this research work is to answer the following research questions, derived from the problem statement discussed in the preceding paragraphs and the gaps identified in existing research works listed above.

- How can the relationship or association between business outcome and DQ be measured?
- Is this association influenced by the relative importance of DQ factors?
- Is this association further strengthened by confidence of users of DSS on the quality of data contained in the underlying DSS?
- How can the relative importance (weightage) of the DQ factors and confidence in the data be measured?
- How can the end users (of data for business decision making) be involved in the DQ assessment process?

Nature of the Study

The main objective of this research work is to develop a model to measure DQ from a business outcome perspective and define a framework for such measurement. Since the problem is general in nature and proposed solution is conceptual in nature (related to the selected topic), this work follows the *fundamental research* philosophy. The work involved conducting empirical study (data collection and statistical analysis) described in subsequent chapters to prove the theory postulated in subsequent chapters, through *empirical study*.

The conceptual DQ measurement model was developed initially by identifying a list of DQ criteria and potential decision categories and mapping those DQ criteria to the decision categories based on their applicability for selected decision category. The model was then refined by adding 2 computational variables i.e. *extent of impact* (weightage) and *probability of impact* (confidence factor). The initial mappings and the added variables led a weighted average DQ score that represents an objective measurement of DQ of DSS.

This DQ measurement model may be administered separately for different business functions to arrive at relative weighted scores. For example, the weighted scores for a DW System that is used largely for Strategic decision is expected to be quite different from those that are designed for tactical decision support. However, within the scope of a business function and the Decision Support Systems that support decision making for the said function, this model is expected to give pointers towards the relative ranking of DQ Criteria that needs to be focused for DQ improvement and improve quality of decision making.

Scope of the Study

In order to relate the DQ factors to outcomes, it is imperative that we identify the key decisions that are typically taken by users with the use of the Decision Support Systems. Of course, these decisions vary by different dimensions such as Industry, type or size of the Organization, Function within the Organization, role and level of the decision maker in the organization etc. As such, a comprehensive list of decisions applicable to all the above scenarios would become too wide open for the purpose of this study. At the same time, it will be essential that the framework that we develop distinguish the variables depending on the above dimensions and the constants, irrespective of the above scenarios and thus yield itself to a maximum degree of reusability at all times. Therefore, it was considered that the framework be subjected to deep analysis for a specific function within a selected Industry and analyze applicability of DQ factors and their impact on quality of outcomes, relevant to the selected function / Industry. Accordingly, the research work was carried out to conduct empirical study, on the basis of the proposed model and measurement framework, in retail banking industry.

Organization of chapters

The following chapters of this thesis work are organized in a manner to elaborate each of the above broad topics in the above sequence. Chapter 2 “Literature Survey” provides a much deeper view of the background of the study, research work done so far in this area, findings from such study and gaps identified so far from such literature review. Chapter 3 “Research Methodology” deals at length the evolution of the research framework and the research methodology followed for the work. Further, this chapter narrates the detailed steps comprising the DQ Assessment framework introduced through this work. Chapter 4 “Empirical Study Design” explains in detail the methodology adopted for the research work and the empirical study that were designed to test the hypothesis. This chapter introduces the data collection mechanism, and the manner in which the survey questionnaire were designed and administered. Further, this chapter goes on to describe the data analysis procedure adopted. Chapter 5 “Results and analysis” deals with analysis of the results from the survey, key findings from the analysis, evaluation of the hypothesis in light of the above findings. Chapter 6 “Findings and recommendations” summarizes the key findings from the research, their relevance in the field of research and recommended application of these findings. Chapter 7 “Discussions” sets the context of this research work for future directions. Chapter 8 “Conclusions and future work” discusses conclusions drawn from the current research, limitations, if any, from the Study and recommends directions for future work.

TABLE I. DEFINITION OF TERMS AND ACRONYMS

Term / Acronym	*Definition*
Business promotion decisions	Decisions related to enhancing or furthering business prospects with existing and/or potential clients (e.g. include a customer in mailing list for a new product launch)
Credit decisions	Decisions related to granting or otherwise of credit (or loans) to customers in a Bank
Decision category	Logical grouping of decisions, based on the business function, in the decision making process.
DQ	Data Quality
DSS	Decision Support Systems
DW	Data Warehouse
ISO/IEC 25012:2008	Software product Quality Requirements and Evaluation (SQuaRE) -- Data quality model (standards on DQ from International Organization for Standardization (ISO) and the International Electrotechnical Commission)
IQ	Information Quality
Product decisions	Decisions related to designing the structure and/or processes related to products or services being offered (e.g. rate of interest for tenure, eligibility criteria for different slabs of loans etc.)
Regulatory decisions	Decisions governed by statutory and other legal requirements and based on actual data in the enterprise. E.g. if the loan is in default for > 24 months and interest is not paid for the last 2 quarters, loan needs to be categorized as non-performing.

Term / Acronym	*Definition*
Relationship decisions	Customer Relationship Management related decisions. E.g. decision on classification of selected client.
Tactical decisions	Decisions related to operational or transactions aspects; e.g. whether or not to install ATM in a selected locality
TIQM	Total Information Quality Management

Chapter 2 – Literature Survey

Introduction

This chapter presents a survey of previous research done in the study of DQ, DQ Assessment and relationship between DQ and business outcomes; and in the context of such work presents the gaps in current published literature. The first section introduces existing frameworks and approaches in the general study of DQ and the research philosophy underlying the study. The second section covers various frameworks that exist in the research world for assessment and measurement of DQ. The third section covers details of the past work related to study of DQ and its assessment in Decision Support Systems and/or Data warehouse System(s). The next section covers in detail past literature as it relates to study of DQ as related to business outcome, although limited literature exists in this direction. The last section summarizes the above inputs and consolidates the gaps existing in current work that is the basis for this research work and serves as the background for rest of the chapters of this thesis.

Study of Data Quality Factors

Although there has been no consensus about the distinction between data quality and information quality, there is a tendency to use data quality to refer to technical issues and information quality to refer to nontechnical issues. In this work, no distinction has been made and term data quality has been used to refer to the full range of issues.

DQ has been the subject of research for many years. This section of the chapter explores the literature documenting such research, beginning with the theoretical roots forming the foundation of DQ theory, followed by a discussion of the predominant research focused on establishing a definition of DQ. The section continues with analysis of research examining factors that contribute to DQ.

The work of Wang et al.,1995 can be considered as the consolidated source of literature for the study of DQ. According to the said work, the database literature refers to DQ management as ensuring 1) syntactic correctness (e.g., constraints enforcement that prevents

"garbage data" from being entered into the database), and 2) semantic correctness (data in the database truthfully reflect the real world situation). This traditional approach of data quality management leads to techniques such as integrity constraints, schema integration, and concurrency control. Although critical to data quality, these techniques fail to address some issues that are important to users. Many databases are plagued with erroneous data or data that do not meet users' needs. To address these problems, Wang et al.,1995 introduced a framework for identifying and studying DQ issues. The framework consists of seven elements: management responsibilities, operation and assurance costs, research and development, production, distribution, personnel management, and legal function. The framework laid stress on few principle findings such as the clear need to develop techniques that help DQ and fundamental technical needs for an overall DQ metric.

A recent revisit of these principles by Madnick et. al., 2009 has revealed that most of the findings of the above work remain relevant for further research and work in today's context. However, a few extensions to the framework have been proposed. Per this cited work, "awareness of data and information quality issues has grown rapidly in light of the critical role played by the quality of information in our data-intensive, knowledge-based economy. Research in the past two decades has produced a large body of data quality knowledge and has expanded our ability to solve many data and information quality problems. This work introduced a framework to characterize the research along two dimensions: topics and methods. This work of Madnick et. al., 2009 focuses on 2 key dimensions of any DQ research viz., topic and method and presents 5 broad topics (19 sub-topics) and 14 methods covering the study of DQ research listed in TABLE II. . In the referred table, each element listed can span into a topic-method combination as a research framework. Each DQ research work may be categorized according to the topic addressed and method used or combinations of the same.

TABLE II. TOPICS AND METHODS OF DQ RESEARCH

Topics	***Methods***
1. Data Quality Impact a. Application area (e.g. CRM, KM etc.,) b. Performance, cost-benefit, Operations c. IT Management d. Organizational change, processes e. Strategy, policy 2. Database related technical solutions for DQ a. Data integration, Data warehouse b. Enterprise architecture, conceptual modeling c. Entity resolution, record linkage, corporate householding d. Monitoring, cleansing e. Lineage, provenance, source tagging f. Uncertainty 3. DQ in the context of computer science and IT a. Measurement, assessment b. Information Systems c. Networks d. Privacy	1. Action research 2. Artificial Intelligence 3. Case Study 4. Datamining 5. Design Science 6. Econometrics 7. Empirical 8. Experimental 9. Mathematical Modeling 10. Qualitative 11. Quantitative 12. Statistical analysis 13. System design, implementation 14. Theory and formal proofs

Topics	*Methods*
e. Protocols, standards f. Security 4. DQ in curation a. Curation – Standards and policies b. Curation – Technical Solutions	

In terms of the above summary of topics and methods, the current research work falls under the topics of measurement / assessment & DSS and empirical method i.e. 1.a, 7 and 3.a, 7 combinations. It also touches DW, although not from a technical solution perspective, but, from the perspective of measurement DQ for DW, since the objective of DW is decision support. On this topic, Madnick et. al., 2009 suggests that to manage data quality, an organization first needs to evaluate the quality of data in existing systems and processes.

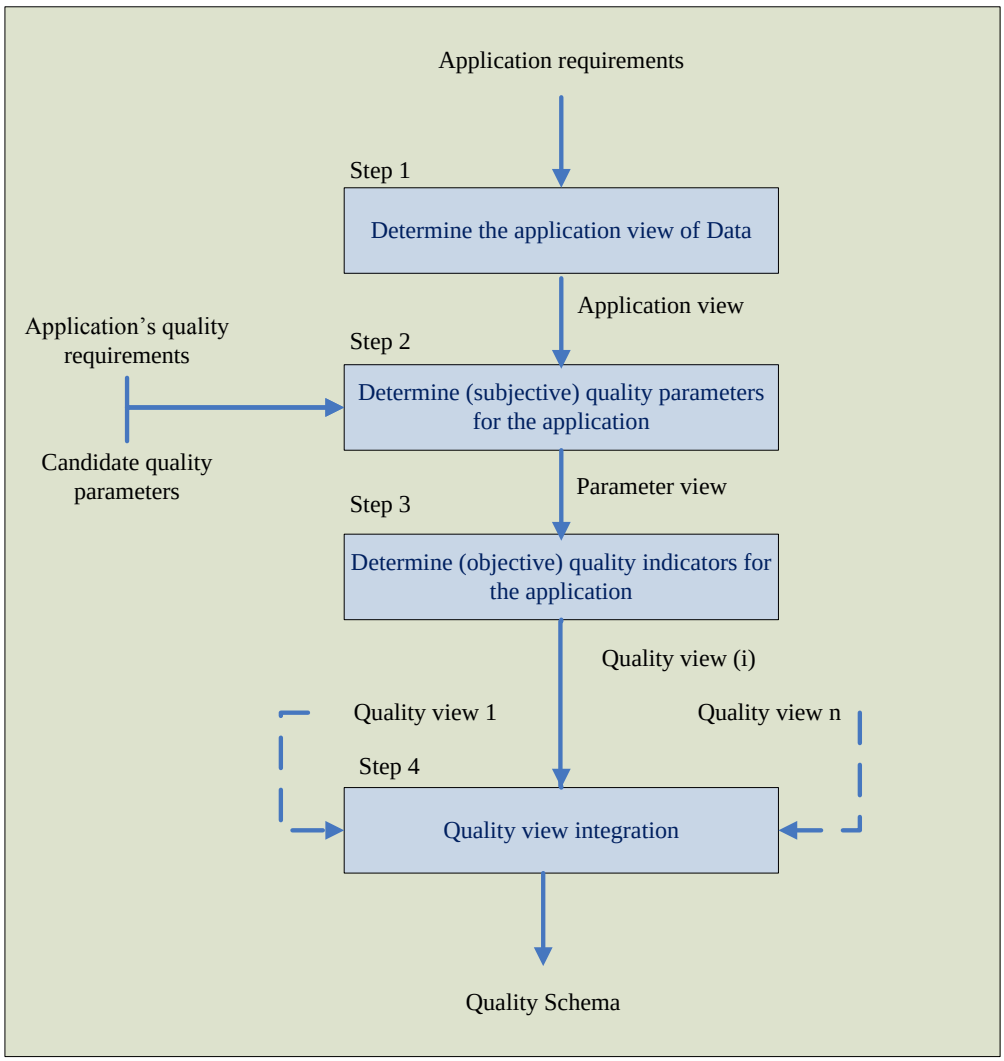
Relevance of the study of DQ has further been summarized by Roger et. al., 2010 in a recent published work where the authors have analyzed abstracts of 467 journal and conference articles published over the past ten years in DQ and IQ, with a view to provide a better understanding of this research area by identifying the core topics and themes. The said work has identified five core topics and fourteen core themes of data quality research and it is expected that these classifications can significantly improve our understanding of the body of literature in data and information quality. The five core topics suggested are:

1. Data quality assessment
2. Management of data quality
3. Quality of data in repositories
4. Data quality in networked data
5. Research Design and methodologies

In terms of the above work, the current research work follows the core topics of ‘data quality assessment’ and ‘ Research Design and methodologies’ and themes of ‘Information systems for measuring data quality’, ‘Data quality in data warehouses’, ‘contextual data quality’ and ‘research methods in data quality’

In a related study Wang et. al., 1995 suggest that it would be useful to tag data with quality indicators which are characteristics of the data and from these quality indicators, users can make their own judgment of the quality of the data for the specific application at hand. Further, the authors propose that DQ is multi-dimensional and argue that different users may have different data quality requirements, and different types of data may have different quality characteristics. The said work presented a process for analysis of DQ requirements, which is depicted in DIAGRAM I. below.

DIAGRAM I. DQ REQUIREMENTS ANALYSIS PROCESS



Moving from the basic definitions and construct of DQ research, this paragraph introduces published literature relevant to DQ factors. In their recent work Knight and Burn, 2005 summarize and compare DQ frameworks and in the process present different sets of DQ factors considered in different frameworks.

The below table (TABLE III.) presents a sub-set of the said frameworks and thus introduce the broad set of DQ factors from past research that was the basis for further work through this research effort.

TABLE III. COMPARISON OF EXISTING INFORMATION QUALITY FRAMEWORKS AND DQ FACTORS

Model	***Constituents***	***Constructs***		***Reference***
		Category	***Dimension***	
Conceptual Framework for Data Quality	4 Categories 16 Dimensions	Intrinsic IQ	Accuracy, Objectivity, Believability, Reputation	Wang & Strong 1996
		Accessibility IQ	Accessibility, Security	
		Contextual IQ	Relevancy, Value added, Timeliness, Completeness, Amount of Info	
		Representational IQ	Interpretability, Ease of understanding, Concise representation, Consistent Representation	
Extended ISO Model	6 Quality Characteristics 32 sub-characteristics	Functionality	Suitability, Accuracy, Interoperability, Compliance, Security, Traceability	Zeist & Hendriks 1996
		Reliability	Maturity, Recoverability, Availability, Degradability, Fault Tolerance	
		Efficiency	Time behaviour, Resource behaviour	

Model	*Constituents*	*Constructs*		*Reference*
		Category	*Dimension*	
		Usability	Understandability, Learnability, Operability, Luxury, Clarity, Helpfulness, Explicitness, Customisability, User friendliness	
		Maintainability	Analysability, Changeability, Stability, Testability, Manageability, Reusability	
		Portability	Adaptability, Conformance, Replaceability, Installability	
Semiotic-based Framework for Data Quality	4 Semiotic descriptions 4 goals of IQ 11 dimensions	Syntatic	Well-defined / formal syntax	Shanks & Corbitt 1999
		Semantic	Comprehensive, Unambiguous, Meaningful, Correct	
		Pragmatic	Timely, Concise, Easily Accessed, Reputable	
		Social	Understood, Awareness of Bias	
Classification of IQ Metadata Criteria	3 Assessment classes 22 IQ Criterion	Subject criteria	Believability, Concise representation, Interpretability, Relevancy, Reputation, Understandability, Value-Added	Naumann & Rolker 2000
		Object criteria	Completeness, Customer Support, Documentation, Objectivity, Price, Reliability, Security, Timeliness, Verifiability	

Model	*Constituents*	*Constructs*		*Reference*
		Category	***Dimension***	
		Process criteria	Accuracy, Amount of data, Availability, Consistent representation, Latency, Response time	
Mapping IQ dimension into PSP/IQ Model	2 Quality Types (Product & Service) 4 IQ Classifications 16 Dimensions	Soundness	Free-of-Error, Concise, Representation, Completeness, Consistent Representation	Kahn 2002
		Usefulness	Appropriate Amount, Relevancy, Understandability, Interpretability, Objectivity	
		Dependable	Timeliness, Security	
		Useable	Believability, Accessibility, Ease of Manipulation, Reputation, Value-Added	
Conceptual Framework for IQ	5 IQ Dimensions	Accuracy	Discrepancy, Timeliness, Source/Author, Bias/Intentionally False Information	Klein 2002
		Completeness	Lack of Depth, Technical Problems, Missing Desired Information, Incomplete When Compared with Other Sites, Lack of Breadth	
		Relevance	Irrelevant Hits When Searching, Bias, Too Broad, Purpose of Web Site	

Model	*Constituents*	*Constructs*		*Reference*
		Category	***Dimension***	
		Timeliness	Information is Not Current, Technical Problems, Publication Date is Unknown	
		Amount of data	Too Much Information, Too Little Information, Information Unavailable	

The below TABLE IV. provides a snapshot of the most common DQ factors and the frequency with which they are included in the above DQ Frameworks, together with a short definition of the DQ factor (Knight and Burn, 2005).

TABLE IV. FREQUENTLY REFERRED DQ FACTORS

S No	***DQ Factor***	***Frequency of reference in DQ Frameworks***	***Definition***
1	Accuracy	8	Extent to which data are correct, reliable and certified free of error
2	Consistency	7	Extent to which information is presented in the same format and compatible with previous data
3	Security	7	Extent to which access to information is restricted appropriately to maintain its security
4	Timeliness	5	extent to which information is not missing and is of sufficient breadth and depth for the task at hand
5	Completeness	5	extent to which information is not missing and is of sufficient breadth and depth for the task at hand
6	Concise	5	extent to which information is compactly represented without being overwhelming (i.e. brief in presentation, yet complete and to the point)
7	Reliability	5	extent to which information is correct and reliable
8	Accessibility	4	extent to which information is available, or easily and quickly retrievable
9	Availability	4	extent to which information is physically accessible
10	Objectivity	4	extent to which information is unbiased, unprejudiced and impartial
11	Relevancy	4	extent to which information is applicable and helpful for the task at hand

S No	*DQ Factor*	*Frequency of reference in DQ Frameworks*	*Definition*
12	Useability	4	extent to which information is clear and easily used
13	Understandability	5	extent to which data are clear without ambiguity and easily comprehended
14	Amount of data	3	extent to which the quantity or volume of available data is appropriate
15	Believability	3	extent to which information is regarded as true and credible
16	Navigation	3	extent to which data are easily found and linked to
17	Reputation	3	extent to which information is highly regarded in terms of source or content
18	Useful	3	extent to which information is applicable and helpful for the task at hand
19	Efficiency	3	extent to which data are able to quickly meet the information needs for the task at hand
20	Value-added	3	extent to which information is beneficial, provides advantages from its use

In order to define and measure the concept of DQ, it is not enough to identify the common elements of DQ Frameworks as individual entities in their own right. DQ needs to be assessed within the context of its generation and intended use. This is because the attributes of DQ can vary depending on the context in which the data is to be used (Knight and Burn, 2005).

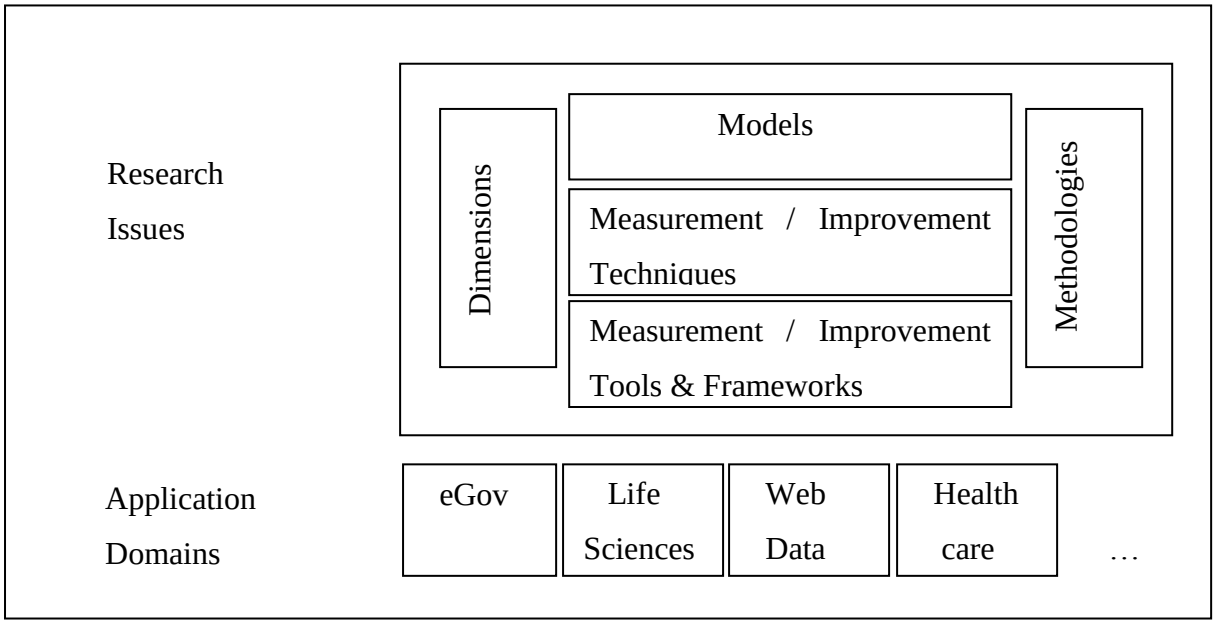
Helfert and Foley, 2009 argue that the frameworks, quality indicators and measurement systems still have limitations. Despite the large amount of literature, agreed criteria lists or measurement approaches are still missing and that as of today no widely accepted IQ framework with generic, generally applicable measurements is available.

To summarize, as per Batini et. al., 2009 the DQ literature provides a thorough classification of data quality dimensions. However, there are a number of discrepancies in the definition of most dimensions due to the contextual nature of quality. The six most important classifications of quality dimensions are provided by Wand and Wang, 1996; Wang and Strong, 1996; Redman, 1996; Jarke et al., 1995; Bovee et al., 2001; and Naumann, 2002. By analyzing these classifications, it is possible to define a basic set of DQ dimensions, including accuracy, completeness, consistency, and timeliness, which constitute the focus of the majority of authors. However, no general agreement exists either on which set of dimensions defines the quality of data.

Amicis and Barone, 2006 argue that the analysis of the dependencies among DQ dimensions is extremely important in the area of information quality in order to improve the quality level of a data set, reconstruct the cause-effect patterns on data quality dimensions, select the most important improvement activities, and more generally increase knowledge on dimensions and their relationship.

Batini and Scannapieco, 2006 have summarized the research topics related to DQ as depicted in the below diagram(DIAGRAM II.).

DIAGRAM II. MAIN ISSUES IN DATA QUALITY

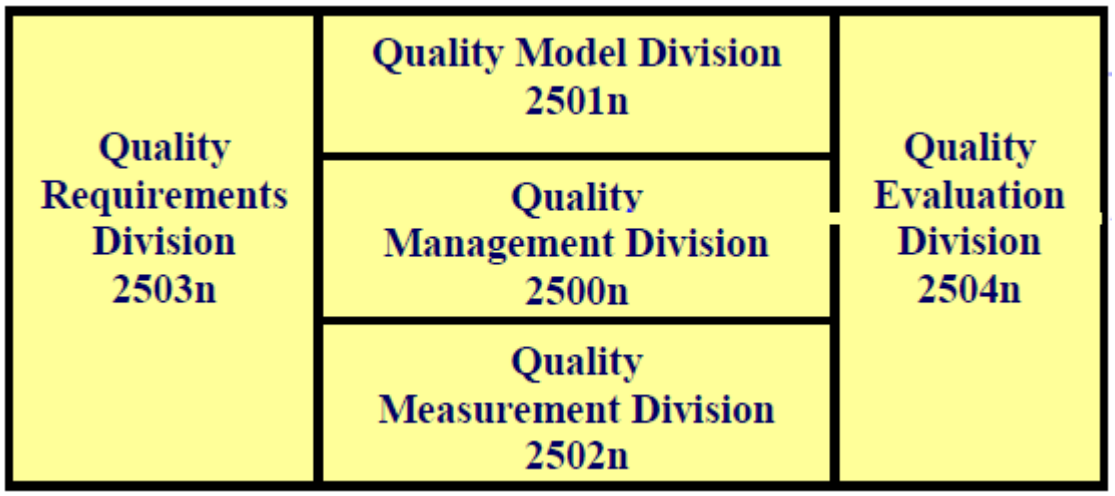


DQ activity starts with choosing dimensions to measure the level of quality of data. Models refer to databases to represent data and database structures. Techniques include algorithms, heuristics, knowledge based procedures and learning processes providing solution to a specific DQ problem or DQ activity. Methodologies provide guidelines to choose DQ measurement processes and when a set of coordinated tools is integrated, it is defined as a framework.

Study of Data Quality Assessment frameworks

Data Quality model is defined as "A defined set of characteristics, and of relationships between them, which provides a framework for specifying data quality requirements and evaluating data quality" (ISO/IEC, 2008). ISO/IEC 25012 (refer TABLE I. for definition) is a part of the SQuaRE series of standards. The SQuaRE series of standards consist of five divisions under the general title "Software product Quality Requirements and Evaluation". This is relevant for study of DQ as well and is represented in the below diagram.

DIAGRAM III. ORGANIZATION OF SQUARE SERIES OF ISO STANDARDS



To manage data quality, an organization first needs to evaluate the quality of data in existing systems and processes (Madnick et. al., 2009). Given the complexity of information systems and information product manufacturing processes, there are many challenges in obtaining accurate and cost-effective assessments of data quality. Research in this area develops techniques for systematic measurement of data quality within an organization or in a particular application context. The measurement can be done periodically or continuously.

Existing literature provides a wide range of techniques to assess and improve the quality of data, such as record linkage, business rules, and similarity measures. Recent research by Stvilia et. al., 2007 has focused on defining methodologies that help select, customize, and apply data quality assessment and improvement techniques. According to Batini et. al., 2009 DQ assessment methodology may be defined as a set of guidelines and techniques that, starting from input information describing a given application context, defines a rational

process to assess and improve the quality of data. In the cited work, Batini et. al., 2009 summarize different perspectives that can be used to analyze and compare DQ methodologies, listed below:

1. Phases and steps that compose the methodology (includes assessment / measurement)
2. The strategies and techniques that are adopted in the methodology for assessing and improving DQ levels
3. The dimensions and metrics that are chosen in the methodology to assess DQ levels
4. The types of costs that are associated with data quality issues
5. The types of data that are considered in the methodology
6. The types of information systems that use, modify, and manage the data that are considered in the methodology
7. The organizations involved in the processes that create or update the data that are considered in the methodology
8. The processes that create or update data
9. The services that are produced by the processes that are considered in the methodology

This research work explores study of DQ covering items 1, 3, 6 and 7 listed above.

This paragraph provides a summarized view of existing DQ assessment methodologies (based on the above perspectives). The table below lists methodologies from existing literature.

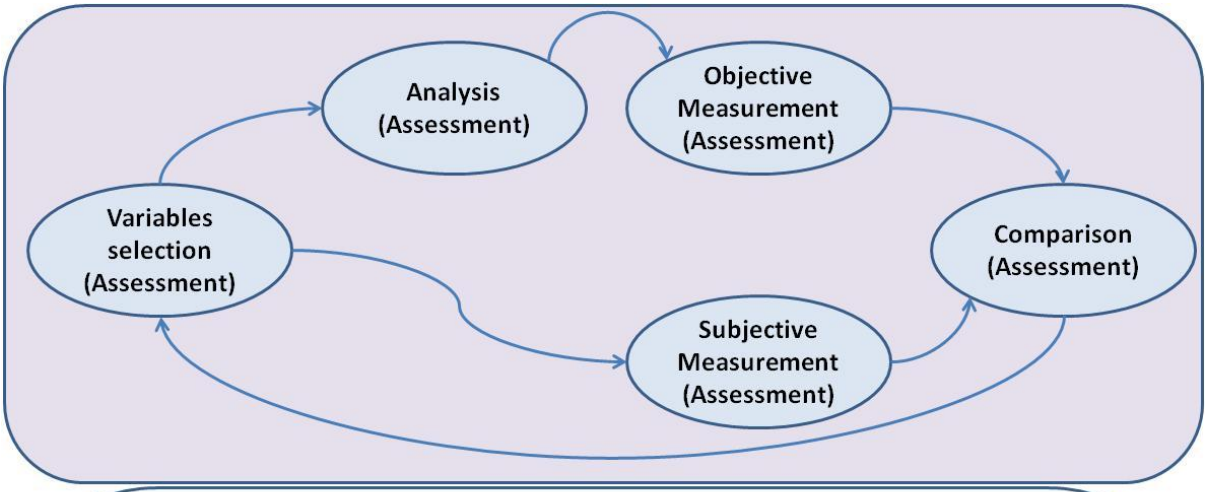
TABLE V. EXISTING DQ ASSESSMENT METHODOLOGIES

S No	***Acronym***	***Name of the Methodology***	***Author***	***Reference***
1	TDQM	Total Data Quality Management	Wang	Wang et. al. 1998
2	DWQ	The Datawarehouse Quality Methodology	Jeusfeld	Jeusfeld et. al.,, 1998
3	TIQM	Total Information Quality Management	English	English, L., 1999
4	AIMQ	A Methodology for Information Quality Assessment	Lee	Lee et. al., 2002
5	CIHI	Canadian Institute for Health Information Methodology	Long and Seko	Long. J et. al., 2005
6	DQA	Data Quality Assessment	Pipino.L, Lee.Y and Wang.R. Y.	Pipino, L. et. al., 2002
7	IQM	Information Quality Management	Eppler	Eppler, M., 2002
8	ITSAT	ITSAT Methodology	Falorsi, P., Pallara, S., Pavone, A., Alessandroni, A., Massella, E., and Scannapieco, M.	Falorsi et. al., 2003
9	AMEQ	Activity-based Measuring and Evaluating of product information Quality (AMEQ) methodology	Su and Jin	Su, Y and Jin, Z 2004
10	COLDQ	Loshin Methodology (Cost-effect Of Low Data Quality	Loshin	Loshin, 2004

S No	*Acronym*	*Name of the Methodology*	*Author*	*Reference*
11	DaQuinCIS	Data Quality in Cooperative Information Systems	Scannapieco, M., Pernici, B., and Pierce, E	Scannapieco et. al., 2005
12	QAFD	Methodology for the Quality Assessment of Financial Data	Amicis, De F. and Batini, C.	Amicis and Batini, 2004
13	CDQ	Comprehensive methodology for Data Quality management	Batini, C., Cabitza,F., Cappiello,C. and Francalanci, C.	Batini et. al., 2008

Of the different methodologies listed above, QAFD methodology will be discussed in detail. The QAFD methodology combines quantitative objective, and qualitative subjective assessments to identify quality issues and select the appropriate quality improvement actions. Context-dependent indices, data quality rules, measurements, and strategies for quantitative and qualitative assessments are defined. The phases and steps involved in QAFD are depicted in the below Diagram and the paragraph thereafter.

DIAGRAM IV. PHASES OF QAFD



Phase 1 : **Variables selection**

Input → Financial variables and financial context

Data Analysis: identification and description of the most relevant financial registry variables, classification of the variables along 3 classes: qualitative / categorical, quantitative / numerical and dates

Output → Classification and groups of primary variables

Phase 2 : **Analysis**

Input → Classification and groups of primary variables, data quality analysis techniques, business rules, process analysis

DQ Requirements Analysis: Analysis of DQ dimensions, syntactic and semantic accuracy, internal and external consistency, completeness, currency, timeliness and uniqueness

Output → Identification of errors

Phase 3 : **Objective measurement**

Input → Identification of errors

Measurement of quality : Definition of indices for the evaluation and quantification of the global data quality level

Output → Quantitative objective assessment

Phase 4 : **Qualitative subjective assessment**

Input → Business and data quality expertise, measurement and classification and groups of primary variables

Measurement of quality: Business expert assessment, financial operator assessment, data quality expert assessment, comparison of subjective assessments

Output → Qualitative subjective assessment

Phase 5 : **Comparison**

Input → Qualitative objective dimensions, qualitative subjective dimensions

Measurement of quality: Calculation of discrepancies

Output → Assessment and improvement

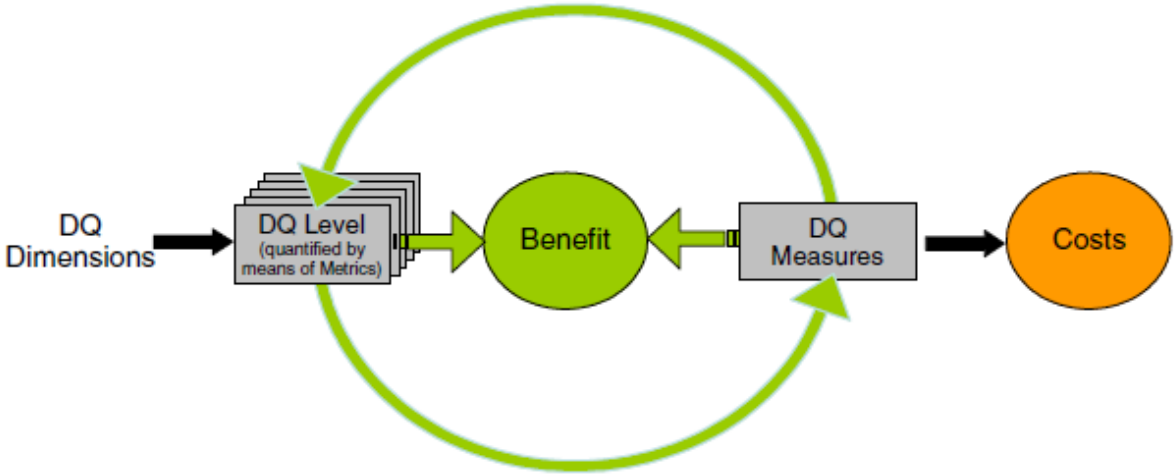
1. First, the methodology selects the most relevant DQ variables. Selection is usually based on knowledge from previous assessments, according to their practical effectiveness. Variables are grouped in categories of "related issues" that are characterized by the same risk, business, and descriptive factors.
2. The second phase aims at discovering the main causes of errors. The most relevant data quality dimensions are identified in this phase and data quality rules are produced. Data quality rules represent the dynamic semantic properties of variables that cannot be measured along quality dimensions.
3. In the third phase, the objective assessment is performed based on quantitative indexes.
4. The subjective assessment is performed in the fourth phase from three different perspectives; business experts, customers, and data quality experts. Each interviewee has to assess the quality level along each quality dimension. An overall assessment is obtained as the mean value of the subjective assessment of each class of experts.
5. Finally, objective and subjective assessments are compared in the fifth phase.

Recognizing that there is no widely recognized effective way on the evaluation of information quality, Gu Lin et. al., 2011 propose evaluation of enterprise information quality based on QFD. This work is based on the definition and analysis of information quality and suggests an improved House of Quality model for enterprise information quality evaluation.

Keeton et. al., 2010 suggest that quantifying DQ can help improve decision making while proposing a research agenda to explore DQ metrics in the Systems domain. Heinrich et. al., 2007 reason that the growing relevance of DQ has revealed the need for adequate measurement since quantifying DQ is essential for planning quality measures in an economic manner. Elaborating the theory of DQ measurement, the referred authors present an illustration of the closed loop of an economically oriented management of DQ. According to the authors, this loop can be influenced via DQ measures (e.g. data cleansing measures, buying external address data etc.). Taking measures improves the current level of DQ

(quantified by means of metrics). This leads to a corresponding economic benefit (e.g. enabling more effective customer contacts). Moreover, based on the level of DQ and taking into account benchmarks and thresholds, firms can decide on taking (further) measures or not.

DIAGRAM V. DQ CLOSED LOOP



On a related work, Heinrich et. al., 2009 observe that many DQ metrics are designed on an ad hoc basis to solve specific, practical problems and that they are often highly subjective. To enable a scientific foundation and an evaluation of the metrics, Heinrich et. al., 2009 propose six normative requirements (for DQ metrics) from literature. These requirements are listed below:

Normalization: An adequate normalization is necessary to assure that the values of the metric are comparable (for instance, to compare different levels of DQ over time (Pipino et al., 2002).

Interval Scale: To support both the monitoring of the DQ level over time and the economic evaluation of measures the metrics should be interval scaled. This means that the difference between two levels of DQ must be meaningful.

Interpretability: The metrics should be "easy to interpret by business users" and the values of the DQ metrics have to be comprehensible.

Aggregation: the metrics must allow aggregation of the quantified values on a given level to the next higher level.

Adaptivity: To quantify DQ in a goal-oriented way, the metrics need to be adaptable to the context of a particular application.

Feasibility: To ensure practicality, the metrics should be based on input parameters that are determinable. When defining metrics, methods to determine the input parameters shall be defined.

In a recently published work Ge et. al., 2011 have dealt with the difficulties associated with assessing information quality. Through their work, while they acknowledge that research provides several approaches to measure information quality and many case studies constantly illustrate the difficulties in assessing information quality, they reveal that even though a number of IQ assessment frameworks have been proposed, in practice, organizations are still facing difficulties when implementing these assessment frameworks. Through a wide ranging literature survey, the authors further reveal that most existing frameworks are too generic to be used for assessment purposes or merely remain at a theoretical stage. Hence, they emphasize the need to address the limitations of some IQ frameworks, and to develop a practical IQ model on the basis of valid and reliable measurements. Summarizing the findings of their work the authors conclude that IQ is a complex and multi-dimensional phenomenon, which has yet not been fully understood and that this nature of IQ causes challenges to measure IQ and may explain why current frameworks have their limitations.

In a similar study, Helfert and Foley, 2009 view that an analysis of related literature reveals that most approaches (to IQ frameworks) are context dependent, although the contextual dimension is usually not represented in IQ frameworks. The general approach to the study of IQ has offered numerous management approaches, IQ frameworks and list of IQ criteria. There are many IQ measures proposed for several application domains; however organizations still are challenged to apply IQ frameworks and measurement approaches within a specific context. With an aim to address this limitation and extend current work on IQ frameworks, the authors of the cited work propose a context-aware IQ framework that can be applied in various contexts.

The need for studying DQ as it pertains to different types and nature of Information Systems has been dealt with in various research works (Singh, 2010). For Example, Fehrenbacher and Helfert, 2008 conclude that technology is an influencing factor in DQ. On the same lines, Batini et. al., 2009 suggest that different methodologies of DQ assessment deal with different types of information systems viz., monolithic information systems, data warehouse, distributed information system, cooperative information system, web information system and P2P Information systems. In a recent study on related topic, Yaari et. al., 2011 examined the ways in which information consumers evaluate the quality of content in a collaborative-writing environment and the findings support the claim that quality is a subjective concept which depends on the user's unique point of view (specific to the type of information system).

However, past research literature (Ge and Helfer, 2006) states that the problem of insufficient DQ is widespread and is often cited as one of the key factors for decision making. While numerous decisions are failed because of the low quality information, many organizations and individuals are still ignoring the importance and necessity of DQ in decision making. Over the last decade many researchers have focused on decision making models and IQ. However research analyzing the impact of DQ on decision making is limited.

Detailing the association between DQ and decision making process and the need for context in such decision making process, Indushobha et. al., 1999, conducted an experiment to explore the consequences of providing information regarding quality of data used in decision making. Through the said work, they conclude that including information about the quality of the data can impact the decision making process and further find that what information should be included depends on various factors and what is complex to one class of users may not be to another.

DQ measurement should be done not just for the sake of measurement but also to enable and support effective decision making. Traditional methods for evaluating DQ dimensions do so objectively without considering contextual factors such as the decision-task and the decision-

maker's preferences. Quality of the data, therefore, is dependent on the purpose (task). The perceived quality of the data is influenced by the decision-task and that the same data may be viewed through two or more different quality lenses depending on the decision-maker and the decision-task it is used for. For example, a Bank's Branch Manager trying to place orders for brochures to existing customers may find an approximate number of account holders in the branch sufficiently accurate to decide the number of copies to order. The same Branch Manager will not consider this figure an accurate-enough representation of his/her customer base when requesting for confirmation of balance for the purpose of statutory audit. Decision-makers must have the ability to evaluate data quality based on the decision-task that the data is used for. It is therefore important to communicate data quality information to the decision-maker and offer the decision-maker the ability to gauge the quality of the data using task-dependent interpretations. In their paper Shankaranarayanan and Cai, 2006 have justified the need to permit decision-makers to incorporate contextual considerations in the process of evaluating DQ. This important issue has not been explicitly addressed by previous DQ research.

Through a recent work by Shankaranarayananet. al., 2008, the authors submit that one way of addressing data quality issues in decision-making is to provide decision makers with DQ metadata, data that describes the quality of the data, together with the data needed for the decision. Further, the quality of a decision depends on the quality of data used in that decision process and poor decision quality is one of the key problems attributed to poor data quality. Through this cited work, the authors investigate the impact of integrating DQ metadata into the decision process and the role of task complexity and task-related experience in this process. Results from this cited work (Shankaranarayanan et. al., 2008) indicate that when there is a well-established way to integrate DQ metadata into the structured decision task, providing DQ metadata to decision makers could increase the amount of data to be processed. As a result, the decision accuracy could decrease and the time-for-task will increase, depending on the task complexity and the experience of the decision maker. However, having identified the need for DQ metadata for data warehouses, the authors have summarized the issues with implementation of such a framework and suggest directions for future research. In the above referred work, the authors focus on the impact of the extra work

load added by the DQ metadata and seek to understand if and under what circumstances decision makers can accomplish a decision task effectively despite the additional DQ metadata they need to process. Using DQ metadata could negatively affect the performance of a decision task in two different ways: by decreasing the task accuracy and/or increasing the time-for-task. The lesson for designers from this is obvious: assess the complexity of the tasks prior to providing DQ metadata and provide support for integrating the DQ data into the decision process. This has implications for the design of decision support systems and interfaces. Only when the users are not overloaded by the DQ metadata can they harness its benefit. Thus, although this work introduced a theoretical framework, the need for further research works in the direction of refining the model and focus on implementation was underlined.

Underlying the relationship between DQ and decision making, Price et. al., 2008 and Price et. al., 2010 are of the view that reliance on incorrect, obsolete or unsuitable data or uncertainty regarding the quality of available data leads to less-effective decision making. The authors present a framework for understanding DQ and they feel that a comprehensive understanding of DQ is critical to understand how DQ impacts decision making. Through a framework introduced, the authors propose using "data quality tags" (information on quality of existing data) to the decision makers for improving quality of decision making. In a subsequent experiment, Price et. al., 2011 observe that using the above cited framework in the decision making process impacted decision time and consensus. On the same lines, Asproth, 2007 states that while designing a decision-support system, lack of attention to DQ can deteriorate dramatically the accuracy of the output results and introduces visualization aids into the system to help augment awareness of the underlying data quality in decision-support systems. Similarly, acknowledging that decision making in a BI (Business Intelligence) environment can be extremely challenging if the underlying data is of poor quality, Marshall et. al., 2010 studied the impact of underlying quality of data in a BI environment on the decision making process and the findings include specific dimensions of DQ (such as accuracy, consistency etc.) that impact quality of decisions from BI.

Another recent study, Singh and Singh, 2010, examines a very interesting aspect of DQ problem in data warehouses, i.e. identify the reasons for data deficiencies, non-availability or reach ability problems at all the aforementioned stages of data warehousing and to formulate descriptive classification of these causes. The premise of this study was that over the period of time many researchers have contributed to the data quality issues, but no research has collectively gathered all the causes of data quality problems at all the phases of data warehousing. However, a big gap that remained not addressed was linking of these DQ problems to outcomes from decisions from using the DW for decision support.

Reiterating the above aspect of criticality of DQ for decision making, with special reference to specialized application systems (e.g. Data warehouse projects), Idris 2011, suggests that DQ issues are best addressed in the early stages of data sources. According to the said work DQ and data source management is one of the key success factors for data warehouse project and that low quality of data fed into the data warehouse system is likely to lead to inaccurate results if such data is used in the decision making process. In a similar published work AlMabhouh, A et al., 2010 have underpinned the importance of DQ in the study of quality and success factors within a data warehouse. In their future work, the said authors intend to study how importance of different quality factors may differ across impacted groups such as top executives, users, project team members, internal IT specialists, vendors, and consultants.

The impact of DQ in DSS has been studied in different industries. For example, Hasan et. al., 2009; Hasan et. al., 2006 studied the impact of DQ in guideline based clinical decision support systems; the work concludes that poor quality of data in medical records and databases poses a risk in medical-decision making process and to address this propose a framework that explicitly models the nature of data, errors, and how guideline based clinical decisions support systems process information and produce guidance. This framework gives the decision-maker the ability to assess how uncertainty about DQ translates into the risk of negative medical consequences and determine which data elements are most critical for minimizing this risk. On similar lines, Portela et. al., 2010 emphasizes the criticality and sensitivity of DQ on healthcare provided based on integrated decision support system.

Woodall et. al., 2010 conclude that no individual existing technique for assessing DQ is wholly suitable to assess DQ for all types of requirements due to the varying nature of requirements over time and organizational needs; the requirements may be different for every organization and even the same organization over time. They further observe that while some of the DQ assessment techniques are geared towards specific application areas and are often not suitable in different applications, other techniques are more general and therefore do not always meet specific requirements.

In summary, past literature shows that poor quality of data in a warehouse adversely impacts the usability of the warehouse and managing DQ in a warehouse is very important (Shankaranarayanan, 2005). The needs of DQ definition and approach for DW and/or Decision Support Systems are unique and different. However, most of the research work focused on this specialized type of information system, have approached with metadata model (integrating / extending existing metadata in a warehouse with quality-related metadata, which has associated practical problems in implementation) and have not focused on implementation aspects of the framework.

Study of Data Quality linked to business outcome

DQ is relative in nature and the need for DQ varies in context for each business and organization or different parts of the organization, Shankaranarayanan and Yu Cai, 2006. It is therefore important to be able to customize any DQ assessment framework based on business context. The prioritization can act as a basis for weighting or for excluding DQ dimensions.

As seen from the previous paragraph, the key limitation of existing research (DQ research as it relates to decision support) is that it focuses merely on technical / architectural aspects of data quality. Addressing this limitation, Gustafsson et. al., 2006 present a framework for assessing DQ focusing on how data supports the business and that can be used as a compliment to software architecture analysis. In their work, the authors had customized the framework for an insurance company by weighting the dimensions of DQ and by relating DQ to the effect it impose on the enterprise's business.

Even et. al., 2007 emphasize that data consumers assess quality within specific business contexts or decision tasks. DQ has been studied from different technical, functional, and organizational perspectives and poor DQ was shown to cause major damages to organizational outcomes such as failures or profitability loss, as studied by Even et. al., 2009. The same data resource may have an acceptable level of quality for some contexts but this quality may be unacceptable for other contexts. However, existing DQ metrics are mostly derived impartially, disconnected from the specific contextual characteristics. They further argue for the need to revise DQ metrics and measurement techniques to incorporate and better reflect contextual assessment. Through the cited work the authors present new metrics for assessing DQ along commonly used dimensions - completeness, validity, accuracy, and currency. The metrics are driven by data utility, a conceptual measure of the business value that is associated with the data within a specific usage context. The suggested DQ measurement framework uses utility as a scaling factor for calculating quality measurements at different levels of data hierarchy.

In a related work Gustafsson et. al., 2006; Olson, 2003 argue that asking users is an effective way to gain knowledge on the data quality level. These works further state that it is how well the data supports the business that should be measured rather than just assessing the completeness and correctness of data values and to measure DQ on the basis of that definition, the users´ perspective is very important. While presenting a framework for assessing DQ focusing on how data supports the business, the following characteristics of such a framework have been laid out:

- Show how data quality can be improved by stating what actions that makes the biggest increase in data quality.
- Be based on the business need for data quality.
- Provide measures so that the change in data quality can be tracked over time.
- Prioritize the dimensions from the enterprise point of view.
- Compare the actual level of data quality with the needed level.
- Be time-efficient and point out where to focus the efforts, both for further investigations and for improvements.
- Take the users experience as a basis for measuring.

Based on their experiments, the authors conclude that the problems in data quality seem to arise not in the technical system design or human error, but rather the misunderstanding between users and developers.

Evidences in the literature establishing the relationship between the management of information quality and organizational outcomes has been limited and sparse with much of the evidence being anecdotal (Harold et. al., 2008). The authors view that banks provide an interesting focus on information (quality) problems in decision making and that poor information quality has a significant impact on banking (business outcomes).
This work presents a conceptual framework of the relationship between information quality and organizational outcomes (Net benefits), including empirical evidence regarding the validity of this framework. From a related work by Frank, 2008 an analysis to determine the effects of DQ on the quality of decisions reveals relations that are useful to consider (though in the limited context of analyzing the influences of DQ on the quality of a decision with an example of environmental engineering decisions).

DQ characterizes the whole business process rather than just the data present in the databases. Each step in the process, from data capture to processing for decision support, has an impact on the final quality of the data, Dewan et. al., 2008.

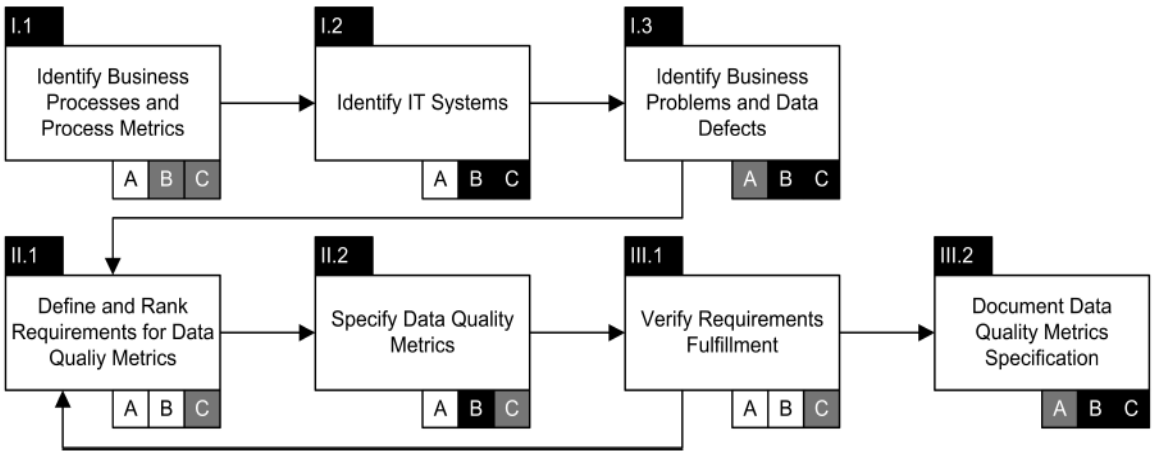
It would be very relevant at this stage to introduce the findings of a recent work, Helfert et. al., 2009 that confirms the relationship between DQ and business outcomes. The said work recommends the need to consider dependencies among DQ factors, and studies these dependencies in the context of a generic DQ assessment framework. The framework proposed through the present research (in Chapters 3 and 4) differs from this previous work in one significant way. Dependencies among DQ factors are studied in the context of specific business decisions. Hence, the resulting DQ score is a function of the decision of interest.

Further elucidating the relationship between DQ and business outcome, in a very recent work Alkharboush et. Al., 2010 emphasize that the assessment of DQ is a key success factor for organizational performance. It supports managers and executives to clearly identify and

present incomplete or inconsistent values in their information systems and as a result, minimizes and eliminates the risk associated with decisions based on poor data. Despite the importance of data quality assessment, limited research has been carried out on assessing the completeness and consistency of the data.

This paragraph presents findings from a recent work of Otto et. al., 2009 that seeks to present a method for the identification of business oriented data quality metrics. There have been numerous theoretical studies on the identification of DQ dimensions (Lee et. al., 2006; Batini et. al., 2006; Wende, 2007; Batini et. al., 2007; Caballero et. al., 2008). Some of these studies even include a definition of metrics for DQ measurement. However, the metrics proposed are either generic and do not include a description of possible measuring techniques, or they refer to certain domains or even single specific cases only. Also, the impact of DQ on companies' business process performance or on companies' capabilities in general has been examined by many experts. However, what has rarely been provided yet are concrete measurements of DQ or any attempts of quantification of any stated impact on business process performance. Through their work, the authors suggest the below steps for identification of business oriented DQ metrics:

DIAGRAM VI. BUSINESS ORIENTED DQ METRICS



In the above diagram, I.3 is the main activity of phase I and aims at identifying cause-effect chains between business problems and data defects (i.e. data defects that cause business problems). The top-down search direction (i.e. first identifying critical business problems and then indentifying causing data defects) has proven to be effective in the discussed cases, but indentifying potential business problems for already known data defects might be might be useful as well. Subject matter experts from both business and IT departments should be involved in collaborative focus group interviews to enable discussions with different perspectives. This important reference concludes that design and documentation of DQ metrics in real-world cases, the analysis of the identified cause-effect chains, and the derivation of generic cause-effect patterns between data defects (e.g. grouped by DQ dimensions) and business problems (e.g. grouped by commercial sectors, or supply chain reference models) constitute multiple areas for further research.

Underlying the importance of link between DQ and quality of decisions, Shankar and Watts, 2003 suggest that decision making is significantly affected by quality of data used in the decision task and so it is necessary to inform the decision makers of the quality of data and also involve them in gauging DQ as it relates to their decision.

In an important work recently, Slone, 2006 studied the empirical investigation of relationship between information quality and organization outcomes to conclude that the relationship between the quality of information and organizational outcomes is systematically measurable. The research work was based on the literature survey finding that researchers have suggested a relationship between the quality of information and the quality of decision-making, with a consequent relationship with organizational strategy; however, there has been very little research in which this relationship was investigated systematically. In the said Ph.D. Thesis, the author argues that despite existing advances in the study of DQ, there has thus far been very little understanding from either a theoretical or practical perspective of the relationship between information quality improvement activities and organizational outcomes.

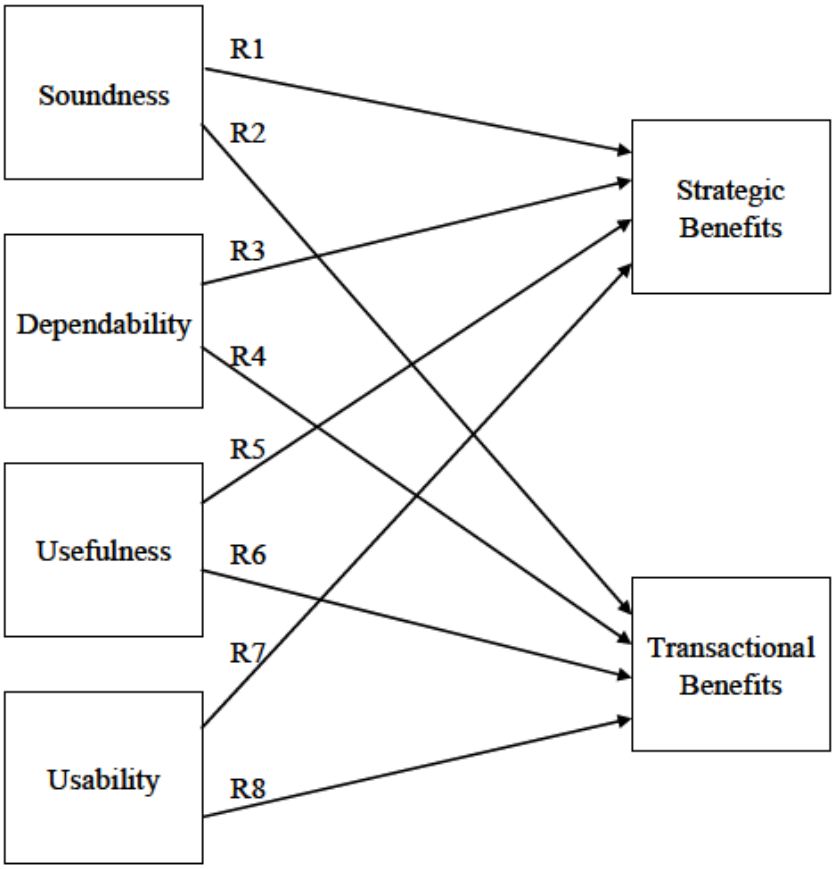
> *"A review of the literature revealed few examples of research addressing information quality strategy; those that were identified were written from a variety of perspectives with little or no commonality in approach or findings.*

> *Consequently, a need was identified for research providing a common conceptual framework for information quality strategy and for research evaluating the relationship between information quality and organizational outcomes"* (Slone, 2006).

The said research work was a quantitative analysis employing multiple regressions to explore the ability to predict organizational outcomes based on information quality. Through the said work, Dr. Slone developed a research model that identified four specific aspects of information quality (soundness, dependability, usefulness, and usability) and two categories of organizational outcome (strategic benefit and transactional benefit), constituting variables in the contextual model. That work focused on the four aspects listed above, because these had previously been demonstrated to represent sixteen dimensions of information quality (Kahn et. al., 2002; Lee et. al., 2002). Dr. Slone derived the research model as depicted below in DIAGRAM VII. , which focused on research of eight strategic relationships (R1 through R8). While underlying the need for further research that are likely to yield more meaningful results, Dr. Slone has recommended several lines of further research, listed below.

- First, research similar to this study, but using a different regression model or a different analytical approach was recommended. Such a study could build directly on the findings of this research.
- Researchers were encouraged to replicate this study using a different sampling frame
- Additional work on improving the instrument.
- Most notably, this research draws into question the appropriateness of the pursuit of increasingly simple metrics for information quality
- An additional question that was raised was related to general impressions of the survey instrument, since the participants very consistently reported being troubled by the redundancy of the survey items. An examination of the instrument seems warranted, with an eye toward providing a more streamlined and parsimonious instrument without unduly diminishing its ability to measure information quality.

DIAGRAM VII. DR. SLONE'S RESEARCH MODEL - STUDY OF INFORMATION QUALITY



Gaps in existing research and research objectives

This section discusses in detail the gaps identified in existing research based on detailed literature review. It lists the major gaps that are chosen for the current research work and at the end of this section, maps those gaps to the current research objectives.

DQ Research is highly interdisciplinary. Instead of this representing an obstacle, it should be considered as a challenge to studies in the area of Management, because of its relevance; after all, in the last decade it was already noted that "poor DQ can have a severe impact on the overall effectiveness of an organization" , Lima et. al., 2006. The evolution of Information Systems, of the forms of work in organizations and even of the environment in which we live, have added several elements of Complexity: it is difficult to measure and manage the information, especially when there are problems in knowing the information we work with. Knowledge and the decision support criteria used in these systems should be restructured in

order to facilitate decision making, which indicates a continuing concern with Information Management and its Quality.

Research of DQ measurement / assessment, has so far been "inputs driven" i.e. focused on identifying a set of DQ factors or analyzing select set of DQ factors (e.g. timeliness or correctness). Even a recent study, Helfert et. al., 2009 reiterates this statement and emphasizes the need for comprehensive DQ measure considering all relevant DQ factors.

> *"Several algorithms have been developed for a subset of dimensions, such as accuracy, completeness, consistency, and timeliness. The definition of an aggregate quality measure is still a much debated issue and existing contributions should be further analyzed and extended…. the most common approach used to obtain a data quality index is to consider all the measures associated with the different quality dimensions and combine them by using a weighed sum.."*

Thus, the study of a comprehensive assessment framework, covering all appropriate DQ factors remains an open area of research in the study of DQ, Alkharboush et. al., 2010. Batini et. al., 2006 emphasize the need for defining a comprehensive set of DQ measurement allowing objective assessment of the underlying DQ and appropriate assessment methods, while dealing with the problems associated with defining a reference set of data, quality dimensions and metrics.

For example, the below table shows how different works have focused only on select set of DQ factors and not in a comprehensive manner:

TABLE VI. STUDY OF SELECTED DQ FACTORS

Name of the work	***DQ Factor(s)***	***Year***	***Reference***
A Procedure to Develop Metrics for Currency and its Application in CRM	Currency	2009	Heinrich et. al., 2009
A novel data quality metric for timeliness Considering supplemental data	Timeliness	2009	Heinrich et. al., 2009
A conceptional approach to unify completeness, Consistency, and accuracy as quality dimensions of data values	Completeness, Consistency, and accuracy	2010	Kaiser, 2010
How to measure Data Quality – A Metrics based approach	Correctness and timeliness	2007	Heinrich et. al., 2007
Does the EU insurance mediation directive help to improve data quality? – A metric-based analysis	Correctness and completeness	2008	Heinrich et. al., 2008
Improving Data Quality: Consistency and Accuracy	Consistency and accuracy	2007	Cong et. al., 2007
Measuring Data Believability: a Provenance Approach.	Believability	2008	Prat et. al., 2008

To summarize, these approaches have resulted in 4 major gaps that form the focus areas for the current research work. These major gaps identified are:

- DQ factors relevant for DSS are not captured
- Lack of context (domain / industry/ function) in measurement of DQ
- Relative influence of DQ factor not considered ("how" a factor impacts DQ and "extent" to which it impacts are not considered)
- Confidence level of users in quality of data and decisions are not captured.

The next 4 paragraphs describe these major gaps and research objectives to meet those gaps. The subsequent 2 paragraphs substantiate (with reference to literature) existence of these gaps.

Gap 1: DQ Factor relevant for DSS: Most of the past research work lead to a set of DQ factor that are largely related to transaction processing systems. However, the DQ requirements for DSS are different from transactional systems because data for DSS are primarily used for decision making. A large number of quality issues relevant for data warehouse cannot be expressed with traditional models, Alkharboush et. al., 2010.

Research objective 1: Revisit the DQ factors (from the past research work) on their applicability for DSS and evaluate need to extend with new DQ factors as applicable to DSS.

Gap 2: Lack of context for use: Past research has not captured the context (i.e. how the data is used and its related outcome) while studying DQ factor. Data is intended to be used for different business decisions and as such any measure of DQ should be sensitive to this context of business decisions for which the data is being used and the quality of business outcomes based on such decisions.

Research objective 2: Evolve a DQ assessment / measurement framework that accounts for applicable DQ factors based on categories of business decisions.

Gap 3: Relative importance of DQ factor: In a set of 20+ DQ factor (comprising overall DQ score) not every factor impacts the overall score equally. The degree of impact varies again based on the decision category for which the data is consumed – e.g. the impact of the DQ factor of timeliness on tactical decisions is different than that on credit decisions. Past research on DQ does not recognize this varying degree of impact.

Research objective 3: Evolve a DQ assessment / measurement framework that considers appropriate weightage for DQ factor applicable for decision category.

Gap 4: Measurement of DQ should reflect the confidence of the users in the quality of data based on their experience with usage of the data.

Research objective 4: Incorporate user confidence (in the quality of underlying data used for business decision making) appropriate in measuring DQ, based on impact of DQ on business outcomes.

Published literature materials call out some/all of these gaps specifically and provide a direction for further research. For example, Knight, 2011 opines that systems information quality (DQ) investigative frameworks, thus far, lack a widely accepted model with which researchers can conceptualize the context of their study, and identify the important DQ characteristics to be examined and empirically tested. The result is a widely varied body of literature lacking a coherent and consistent approach to identifying and measuring systems DQ. Similarly, Gibson, 2010 observes that DQ research to date has approached the subject independent from actual users although the interdependency between the two is obvious.

It may be very interesting to note that Batini et. al., 2009 summarize open issues in the area of DQ assessment as below (reproduced):

"Further open problems in DQ methodologies concern:

1. *The identification of more precise statistical, probabilistic, and functional correlations among data quality and process quality, with a focus on the empirical validation of the models and the extension of the analysis to a wider set of dimensions and to specific types of business processes.*
2. *The validation of methodologies; Often, a methodology is proposed without any large-scale specific experimentation and with none or only a few, supporting tools. There is a lack of research on experiments to validate different methodological approaches and on the development of tools to make them feasible.*

3. *The extension of methodological guidelines to a wider set of dimensions, such as performance, availability, security, accessibility, and to dependencies among dimensions.*
4. *In Web information systems and in data warehouses, data are managed at different aggregation levels. Quality composition should be investigated to obtain aggregate quality information from the quality metrics associated with elementary data."*

Chapter 3 – Research Methodology

Introduction

This chapter presents the methodology used to conduct the research for this study. The first section describes the framework within which the research was conducted and describes how the existing frameworks were adopted and tailored for the problem on hand; the next section covers the details of how the framework was arrived at – selection of data quality factors and reasons behind why the model was designed in such a way. The next section describes the research methodology, covering how a relevant methodology from published literature was adopted and modified for this work and describes the proposed framework introduced through this work. The final section consolidates the research problem and presents the main hypothesis to be validated through the work.

Research philosophy

Existing DQ literature denotes that research on this topic spans multiple research paradigms, methodologies and approaches thus providing flexibility for the researcher to explore their research based on any of these methodologies or approaches, based on the research question (Slone, 2006). This has been described by Greene et. al., 2005 as pragmatic stance for an inclusive philosophical framework within which multiple assumptions and diverse methods can comfortably reside. With this pragmatic stance as the underlying backbone, this research was conducted from the perspective of the post-positivist paradigm.

Post-positivist research emphasizes 2 key characteristics (Ryan, 2006) i.e. research is broad (rather than specialized) and that theory and practice cannot be separate. This paradigm postulates that full understanding (about the research questions) can be reached based on experiment and observation. This paradigm employs empirical means and deductive logic in the quest for an objectively knowable truth. This research, therefore, was undertaken as an empirical study with the objective of finding affirmations in support of the set of hypotheses defined later in this chapter.

Research framework

An important approach for this research work was inspired from and was focused on addressing the gaps listed at the end of Chapter 2 and is also primarily based on further research areas identified in literature (Slone, 2006). A summary view of the limitations of the said study and how they are addressed through this work are summarized in the table below (and elaborated in rest of the chapter):

TABLE VII. RECOMMENDED AREAS OF FURTHER RESEARCH FROM LITERATURE

Suggested area of research	***Approach in this Research Work***
First, research similar to this study (Slone, 2006), but using a different regression model or a different analytical approach was recommended. Such a study could build directly on the findings of this research.	Results from the study are proposed to be subjected to correlation analysis using chi-square test
Researchers were encouraged to replicate this study using a different sampling frame	Sampling framework is proposed to be revamped and implemented to target different population subjects for different inputs i.e. stage 1 – data collection from industry experts that will be the basis for rest of the study, stage 2 – values for selected variable from select target population (actual users involved in decision making) and stage 3 – data for validation to be collected from different set of target population.
Additional work on improving the instrument.	Simplify the survey instrument and enhance the same to include DQ factors appropriate for this work

Suggested area of research	*Approach in this Research Work*
An additional question that was raised was related to general impressions of the survey instrument, since the participants very consistently reported being troubled by the redundancy of the survey items. An examination of the instrument seems warranted, with an eye toward providing a more streamlined and parsimonious instrument without unduly diminishing its ability to measure information quality.	Simplify the survey instrument and split the instrument to multiple instruments with questions and measurement factors appropriate for the target subjects.
Most notably, this research draws into question the appropriateness of the pursuit of increasingly simple metrics for information quality	Introduce and validate a model to measure DQ more comprehensively and enable the measurement in a context sensitive manner

The above approach gave rise to the below research objective, in addition to those listed in page no 57.

Research objective 5: Involve users that are actively involved in the decision making process, to validate the DQ assessment framework.

The various works discussed in Chapter 2 and the gaps identified therein provided an important direction for this research work i.e. development of a unified DQ measure encompassing appropriate DQ dimensions and focusing on business / organization outcomes. The key objective of this work is to develop a comprehensive DQ measurement framework which also addresses filling in the gaps from the past research and more specifically past work related to relationship between DQ and business outcomes (Slone, 2006). While dealing with this subject of DQ measurement Keeton et. al., 2010 are of the view that measurement can be either '*stand-alone IQ metrics*' or '*context-dependent IQ metrics*' and have dealt with research challenges in such measurements. **The focus of the current**

research work is to develop a context based, comprehensive DQ measurement framework.

The need for development of such a measurement framework is evident from several existing literature detailed at length in Chapter 2. Emphasizing the need for further research in the study of DQ and its influence on decisions, Jung, 2007 recommends that the area of study (whether DQ, especially contextual DQ, influences decision performance) is expected to extend a body of research examining the effects of factors that can be tied to human decision-making performance. In addition, in a recent work, Kaiser, 2010, argues that despite a magnitude of literature on different DQ dimensions, effort on defining these DQ dimensions in terms of metrics for measuring different dimensions of DQ has only been done recently and that what is still lacking is an approach towards a unified measure of DQ based on these metrics for different dimensions.

Selection of DQ factors

Knight et. al., 2005 have summarized 12 widely accepted DQ framework collated from published research work. These frameworks followed a construct of '*IQ category*' and '*IQ dimension*' for grouping DQ factors. Examples of '*IQ category*' are intrinsic or functionality etc. and that of '*IQ dimension*' are accuracy or completeness or believability etc. At the beginning of this Study, the author used these 12 frameworks (listed in TABLE VIII.) as reference for study and identification of DQ factors that are relevant for the research work.

TABLE VIII. DQ FRAMEWORK FROM LITERATURE

Sl. No.	***Name of the framework***	***Year***	***Author/ Reference***
1	A Conceptual Framework for Data Quality	1996	Wang and Strong, 1996
2	Extended ISO Model	1996	Zeist et. al., 1996
3	Applying a Quality Framework to Web Environment	1999	Alexander et. al, 1999

Sl. No.	*Name of the framework*	*Year*	*Author/ Reference*
4	IQ of Individual Web Site	1999	Katerattanakul et. al., 1999
5	Semiotic-based Framework for Data Quality	1999	Shanks et. al., 1999
6	Conceptual Framework for measuring IS Quality	2000	Dedeke, 2000
7	Classification of IQ Metadata Criteria	2000	Naumann et. al., 2000
8	Quality metrics for information retrieval on the WWW	2000	Zhu et. al., 2000
9	Adapted Extended ISO Model for Intranets	2001	Leung, 2001
10	Mapping IQ dimension into the PSP/IQ Model	2002	Kahn et. al., 2002
11	Conceptual Framework for IQ in the Website Context	2002	Eppler et. al., 2002
12	Conceptual Framework for IQ	2002	Klein, 2002

The author carried out an exercise of validation (of relevance of the framework and applicability of the DQ factors contained therein to the subject of DQ in decision making and DQ in DSS). This exercise involved studying each of these 12 frameworks with reference to the below aspects:

- Primary focus of the framework (e.g. framework for assessing DQ in decision support setting in DSS Vs. content aggregation in WWW or DQ concepts in general.
- New framework or adaptation of an existing framework (e.g. 'IQ of Individual Web Site' adapted 2 dimension from existing framework i.e. 'A Conceptual Framework for Data Quality'
- Repeating DQ factors that can be consolidated by referring to other frameworks. For example, in a WWW DQ related framework, in addition to

factors such as visual settings or attractiveness (which are relevant for DQ in decision making context) , if there are reference to factors such as accuracy or correctness, which are repeated across multiple frameworks, then such DQ factors are consolidated for the purpose of listing.

The above exercise revealed that while varied in their approach and application, the frameworks share some common characteristics regarding their classifications of the dimensions of quality and share a lot of common DQ factors. Based on the above exercise (of validation and consolidation), 6 frameworks that deal with DQ in a conceptual manner were taken up for further closer study and analysis. Since, the objective of this research work (refer page 17) is to address a few open questions (related to DQ assessment) that are general in nature, selection of these frameworks through the above exercise was considered appropriate.

Summary of these 6 DQ frameworks collated from the various earlier works of DQ research (Knight et. al., 2005) is presented in TABLE IX. .

TABLE IX. COMPARISON OF INFORMATION QUALITY FRAMEWORKS

Model	***Constituents***	***Constructs***		***Author & Year***
		Category	***Dimension***	
Conceptual Framework for Data Quality	4 Categories 16 Dimensions	Intrinsic IQ	Accuracy, Objectivity, Believability, Reputation	Wang & Strong 1996
		Accessibility IQ	Accessibility, Security	
		Contextual IQ	Relevancy, Value added, Timeliness, Completeness, Amount of Info	
		Representational IQ	Interpretability, Ease of understanding, Concise representation, Consistent Representation	
Extended ISO Model	6 Quality Characteristics	Functionality	Suitability, Accuracy, Interoperability,	Zeist & Hendriks

Model	*Constituents*	*Constructs*		*Author & Year*
		Category	*Dimension*	
	32 sub-characteristics		Compliance, Security, Traceability	1996
		Reliability	Maturity, Recoverability, Availability, Degradability, Fault Tolerance	
		Efficiency	Time behaviour, Resource behaviour	
		Usability	Understandability, Learnability, Operability, Luxury, Clarity, Helpfulness, Explicitness, Customisability, User friendliness	
		Maintainability	Analysability, Changeability, Stability, Testability, Manageability, Reusability	
		Portability	Adaptability, Conformance, Replaceability, Installability	
Semiotic-based Framework for Data Quality	4 Semiotic descriptions 4 goals of IQ 11 dimensions	Syntatic	Well-defined / formal syntax	Shanks & Corbitt 1999
		Semantic	Comprehensive, Unambiguous, Meaningful, Correct	
		Pragmatic	Timely, Concise, Easily Accessed, Reputable	

Model	*Constituents*	*Constructs*		*Author & Year*
		Category	*Dimension*	
		Social	Understood, Awareness of Bias	
Classification of IQ Metadata Criteria	3 Assessment classes 22 IQ Criterion	Subject criteria	Believability, Concise representation, Interpretability, Relevancy, Reputation, Understandability, Value-Added	Naumann & Rolker 2000
		Object criteria	Completeness, Customer Support, Documentation, Objectivity, Price, Reliability, Security, Timeliness, Verifiability	
		Process criteria	Accuracy, Amount of data, Availability, Consistent representation, Latency, Response time	
Mapping IQ dimension into PSP/IQ Model	2 Quality Types (product & Service) 4 IQ Classifications 16 Dimensions	Soundness	Free-of-Error, Concise, Representation, Completeness, Consistent Representation	Kahn 2002
		Usefulness	Appropriate Amount, Relevancy, Understandability, Interpretability, Objectivity	
		Dependable	Timeliness, Security	
		Useable	Believability, Accessibility, Ease of Manipulation,	

Model	*Constituents*	*Constructs*		*Author & Year*
		Category	*Dimension*	
			Reputation, Value-Added	
Conceptual Framework for IQ	5 IQ Dimensions	Accuracy	Discrepancy, Timeliness, Source/Author, Bias/Intentionally False Information	Klein 2002
		Completeness	Lack of Depth, Technical Problems, Missing Desired Information, Incomplete When Compared with Other Sites, Lack of Breadth	
		Relevance	Irrelevant Hits When Searching, Bias, Too Broad, Purpose of Web Site	
		Timeliness	Information is Not Current, Technical Problems, Publication Date is Unknown	
		Amount of data	Too Much Information, Too Little Information, Information Unavailable	

Of the 6 DQ Frameworks presented above, the IQ Metadata framework by Naumann and Rolker, 2000 provided the basis to analyze and determine the appropriate DQ categories and dimensions in a new, assessment-oriented way. This cited work consolidates DQ criteria and presents them in an assessment oriented way and additionally has provided assessment methods for each criterion. This cited work considers confidence measures for the

assessment methods. How each of these aspects contributed to evolve the proposed framework are discussed in detail in subsequent paragraphs of this work.

While the above framework was chosen to be the basis, an exercise was carried out to analyze the individual components of the framework, evaluate applicability of them in light of the current research work and introduce appropriate modifications. The following table deals with the existing framework components and how they are dealt with as part of this study:

TABLE X. DQ FRAMEWORKS CONSIDERED FOR THE RESEARCH WORK

Factor	***Approach in Naumann's framework***	***Approach in this Research Work***
Identification of DQ Criteria	22 DQ criteria were identified and categorized into 3 criteria viz., subject, object and process.	Objectively revisited the criteria (from the work of Naumann & Rolker 2000) on their applicability for Decision Support Systems; evaluated need to extend to new criteria as applicable to Decision Support Systems. Revisited the DQ classes. This evaluation was based on TIQM Methodology (explained in subsequent paragraph). Accordingly the following changes were introduced as part of the new framework that this work introduces: 1. Introduced 3 new DQ criteria relevant for

Factor	*Approach in Naumann's framework*	*Approach in this Research Work*
		Decision support systems. The new criteria focus on data quality as it pertains to relevance of dimensions of data, adequacy of business measures and appropriateness of aggregation supported by such systems. 2. Arrived at list of 23 DQ criteria for rest of the study
Identification of DQ classes	User, Source and Query Process as existing classification	1. Introduced the concept of decision category as influencing factor in the study of data quality. 2. Replaced the existing 3 DQ classes (subject criteria, object criteria and process criteria) with 6 decision categories (relevant for the Industry chosen for

Factor	*Approach in Naumann's framework*	*Approach in this Research Work*
		rest of the work).
Identification of assessment methods	Naumann's work suggests using 3 assessment methods viz., user experience, user sampling and continuous user assessment.	1. These methods were adopted and refined further to capture "Confidence in IQ assessment methods" as dealt with in Nuamann's work. 2. Survey instruments used in similar past work (Slone, 2006) were re-used and refined further to incorporate additional DQ criteria and classes (decision categories), besides capturing confidence in assessment.
Map DQ Criteria to Organization outcome	The existing work doesn't map DQ criteria to organization outcome to measure the impact and as such is a gap	Introduce a framework to identify the possible outcomes / decision from use of Decision Support Systems and mapping DQ criteria that are likely to impact the outcomes. E.g. accuracy as a criteria is likely to impact pricing decision.

TABLE XI. lists DQ factors that were the considered in the framework introduced by Naumann and Rolker, 2000and the suggest assessment method.

TABLE XI. DQ FACTORS FROM NAUMANN AND ROLKER FRAME WORK

Assessment Class	***IQ Criteria***	***Assessment Method***
Subject Criteria	Believability	User Experience
	Concise representation	User sampling
	Interpretability	User sampling
	Relevancy	Continuous user assessment
	Reputation	User experience
	Understandability	User sampling
	Value-added	Continuous user assessment
Object Criteria	Completeness	Parsing, sampling
	Customer Support	Parsing, contract
	Documentation	Parsing
	Objectivity	Expert Input
	Price	Contract
	Reliability	Continuous assessment
	Security	Parsing
	Timeliness	Parsing
	Verifiability	Expert Input
Process Criteria	Accuracy	Sampling, cleansing techniques
	Amount of data	Continuous assessment
	Availability	Continuous assessment
	Consistent representation	Parsing
	Latency	Continuous assessment
	Response Time	Continuous assessment

In this Section we discuss the need for extending the above DQ Criteria list to meet the DQ requirements specific to Decision Support Systems. Of the different DQ methodologies that exist in current literature, TIQM (Total Information Quality Management) (Batini et. al., 2009) needs specific mention in this context. The TIQM methodology has been designed to support data warehouse projects (English, 1999). Since the focus of this research is on DQ in DSS (used interchangeably with data warehouse projects) this methodology needs special mention. DQ dimensions (or factors) considered by TIQM methodology include Definition conformance (consistency), Completeness, Business rules conformance, Accuracy (to surrogate source), Accuracy (to reality), Precision, Non-duplication, equivalence of redundant data, Concurrency of redundant data, accessibility, timeliness, contextual clarity, Derivation integrity, Usability, Rightness (fact completeness), cost. In the next few paragraphs, some of the above factors that need special attention (in the context of their uniqueness to data warehouse projects and/or DSS) are narrated in detail and how this treatment has led to the need for introduction of new DQ factors in assessment of DQ has also been explained. DQ dimensions from TIQM that are given special treatment are contextual clarity (new DQ factor introduced: relevance of dimension), Rightness (fact completeness) (new DQ factor introduced: relevance of measures) and precision (new DQ factor introduced : granularity).

Granularity of data in a Decision Support System infers the level of details that it carries and thus refers to the depth of data available. High granularity refers to data that is at or near the transaction level. Data that is at the transaction level is usually referred to as atomic level data. Low granularity refers to data that is summarized or aggregated, usually from the atomic level data. Summarized data can be lightly summarized as in daily or weekly summaries or highly summarized data such as yearly averages and totals. The data model for a DSS should consider the functional requirements and appropriately arrive at the granularity. The design has to provide for a right balance i.e. very high granularity may impact query performance or very low granularity may impact ability to conduct intended analysis. Thus, granularity is a very critical factor influencing the quality of design (and thus the quality of data / information) of a Decision Support System. From the above table, we see that one of

the DQ Criteria included in process criteria is "amount of data". In the context of this work, we would like to rephrase the same as granularity.

Relevance of dimensions: Design and presentation of appropriate and adequate dimensions, with relevant hierarchies and dimensional attributes is essential for successful implementation of any Decision Support System. For effective decision making, the users of a DSS slice and dice the business facts (e.g. sales or number of resources recruited etc.) by different dimensions and at different levels (or hierarchies) within the selected dimension. E.g. sales figures for a selected quarter is analyzed by dimension named geography and this is done by country / state / store location and/or other combinations. This analysis may be extended to include additional dimensions such as above sales turnover by quarter, product line etc. As such quality decisions from DSS is taken after thorough analysis of information based on different dimensions; therefore, it is imperative to study relevance of dimensions as a DQ factor for studying quality of outcomes from decision making process.

Relevance of measures: On the same lines as above, inclusion of adequate business measures (sometimes directly derived from source systems or derived / computed measures) also assumes significance for successful implementation of any DSS. For effective decision making, the users of a DSS compare various measures or business facts (e.g. sales or number of resources recruited etc.) for their decision making. E.g. analyze not just sales figures in isolation, but along with cost of sales, margin, inventory etc. Such an analysis is intended to provide insights to the users of a DSS and help them identify the root cause of problems that they like to fix. Therefore, it is imperative to study relevance of dimensions as a DQ factor for studying quality of outcomes from decision making process.

Based on the above discussions, the revised set (including additional DQ factors proposed to be introduced through this work) of DQ factors relevant for the purpose of this Study are listed below. New IQ criteria introduced via the above analysis are shown in ***Bold Italics***.

TABLE XII. DQ FACTORS CONSIDERED FOR THE RESEARCH WORK

Assessment Class	***IQ Criteria***
Subject Criteria	Believability
	Concise representation
	Interpretability
	Reputation
	Understandability
	Value-added
	Granularity
Object Criteria	***Relevancy – Measures***
	Relevancy – Dimensions
	Aggregation
	Completeness
	Customer Support
	Documentation
	Objectivity
	Price
	Reliability
	Security
Process Criteria	Accuracy
	Availability
	Consistent representation
	Latency
	Response Time
	Timeliness
	Verifiability

Research Methodology

Evidence in literature establishing the relationship between management of information quality and organizational outcomes has to this point been limited and sparse, with much of that evidence being anecdotal. A research model was proposed for investigating this relationship. In the next few paragraphs, the detailed step-by-step approach of the current study has been explained, starting with introduction of the framework / methodology that served as the approach for this Study.

This new research model is inspired from the QAFD methodology, Amicis and Batini, 2004, discussed in detail in Chapter 2. The detailed steps proposed in QAFD methodology and the manner in which they have been adopted in this research work to evolve the framework to measure the relationship between DQ factors and business outcome is outline below:

TABLE XIII. QAFD STEPS AND THEIR ADOPTION FOR THIS WORK

Step as described in QAFD	***Adoption in this research work***
First, the methodology selects the most relevant DQ variables. Selection is usually based on knowledge from previous assessments, according to their practical effectiveness. Variables are grouped in categories of "related issues" that are characterized by the same risk, business, and descriptive factors.	DQ factors relevant for the study were selected from existing research materials (Naumann and Rolker, 2000). In parallel, business decisions were logically grouped based on "related issues" into "decision categories"
The second phase aims at discovering the main causes of errors. The most relevant data quality dimensions are identified in this phase and data quality rules are produced. Data quality rules represent the dynamic semantic properties of variables that cannot be measured along quality dimensions.	Additional variables appropriate for the subject of study were added based on recommendations from existing literature Batini et. al., 2009 and English, 1999.

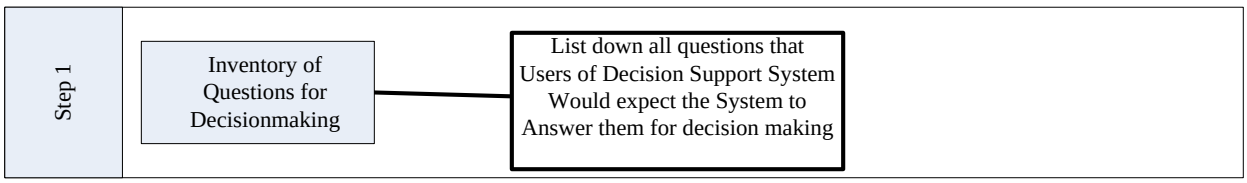
Step as described in QAFD	*Adoption in this research work*
In the third phase, the objective assessment is performed based on quantitative indexes.	Step introduced for calibration, capturing "weightage" for the selected DQ factors, serving as a variable in computation of DQ score.
The subjective assessment is performed in the fourth phase from three different perspectives; business experts, customers, and data quality experts. Each interviewee has to assess the quality level along each quality dimension. An overall assessment is obtained as the mean value of the subjective assessment of each class of experts.	Step introduced to capture "confidence factor" experts in DSS, serving as 2nd critical variable in computation of DQ score.
Finally, objective and subjective assessments are compared in the fifth phase.	DQ score is computed based on both "weightage" and "confidence factor" and is compared with independent assessment of DQ from users of DSS.

Research approach – DQ Assessment Framework

This section covers in detail the individual steps that were involved in evolution of the new framework proposed through this work.

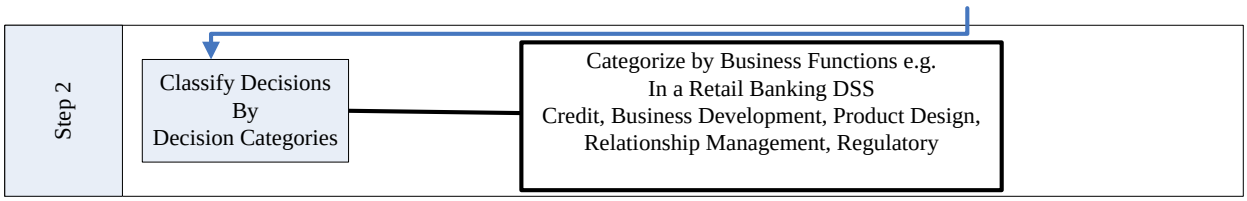
Step 1 – Refer DIAGRAM VIII. : Conduct research within the business function for which the framework is to be implemented and identify list of possible answers that users of DSS expect to get from the System. This step is critical as it sets the foundation for multiple business decisions that the users of DSS are likely to arrive at using the data that resides in the system. Though not directly, this work seeks to assess the impact of DQ on the quality of these decisions. As we all are aware, decisions do not exist without underlying questions and thus building this framework starts with arriving at the master list of all potential questions.

DIAGRAM VIII. STEP 1: IDENTIFY QUESTIONS



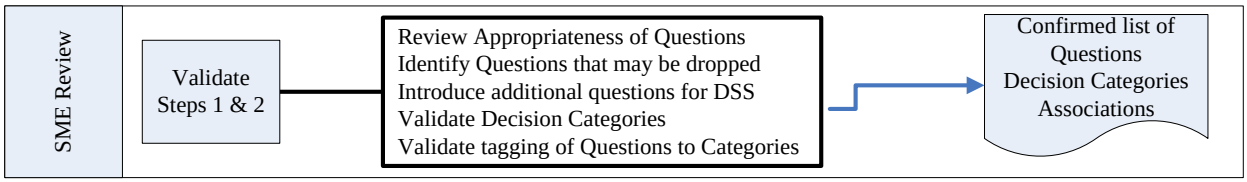
Step 2 - DIAGRAM IX. : Group these decisions logically into sets of decision categories. As part of evolution of the framework, concept of 'decision category' was introduced, serving as the driver for context-sensitive analysis of DQ. This concept of decision category is the central theme of the framework, values for which needs to be derived separately for selected industry while implementing this framework. This logical grouping is based on business context and "related issues" were grouped into "decision categories". This grouping was again an essential aspect of this research work, as the process of logical grouping provides the context sensitive measurement of DQ. As summarized at the end of Chapter 2, a major gap that exists in the current literature pertains to lack of context for use. Past research has failed to capture the context (i.e. how the data is used and its related outcome) while studying DQ factor. Data is intended to be used for different business decisions and as such any measure of DQ should be sensitive to this context of business decisions for which the data is being used. As such, this step introduces the ability to approach DQ factors based on categories of business decisions and measure DQ factors differently for different decision categories.

DIAGRAM IX. STEP 2: IDENTIFY DECISION CATEGORIES



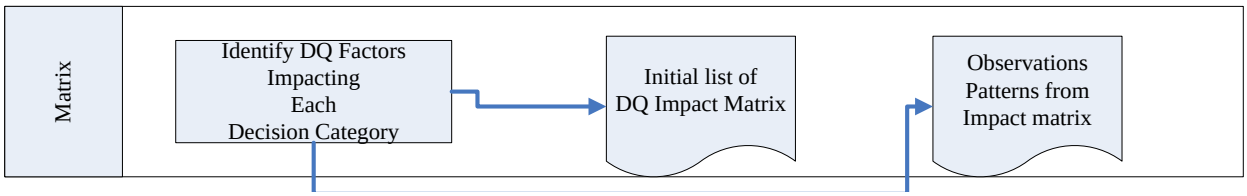
Step 3 – Refer DIAGRAM X. : Validate the list and decision categories with subject matter experts from the selected business function. This step involves proofing the base identified so far, by seeking expert inputs from industry specialists in the selected function for which the framework is proposed to be implemented. The objective is to validate the completeness of the questions and more importantly validate correctness of their logical grouping into decision categories.

DIAGRAM X. STEP 3: VALIDATION BY EXPERTS



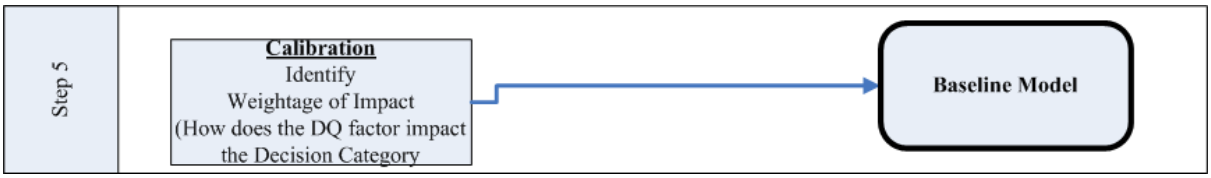
Step 4: Refer DIAGRAM XI. Build the decision category-DQ factor matrix to assess the applicability or otherwise of each DQ factor to selected decision category and vice versa. Before this, the DQ factors applicable for DSS were identified, as explained in TABLE XII. and the paragraphs preceding the same. Creating this matrix is again significant as it adds further context to DQ measurement. In addition to the context (of decision categories) introduced in the previous step, this matrix provides further context to follow different set of DQ factors for different decision categories, based on their specific applicability.

DIAGRAM XI. STEP 4: DQ FACTOR-DECISION CATEGORY MATRIX



Step 5: Refer **Error! Reference source not found.**Calibrate the above association by industry experts by capturing value for weightage component. This step involves walking through the model as has been evolved so far, with Industry experts to refine the initial settings based on the experience of Industry Experts. The objective of this step is to get the model calibrated with appropriate weightage assignments for selected decision category-DQ factor combination. This calibrated weightage factor serves as the baseline for subsequent steps.

DIAGRAM XII. STEP 5: CALIBRATION



Step 6: Refer DIAGRAM XIII. Measure DQ for each decision category, using the baseline model and capturing the confidence of the users in the quality of data, capturing value for confidence level component within the boundary of above calibration.

DIAGRAM XIII. STEP 6: DQ MEASUREMENT

To further add context to DQ, 2 sets of variables were introduced i.e. weight and confidence level. Weight represents how strongly a specific DQ factor influences a decision category of interest, and confidence level represents confidence of DSS users in quality of data existing in "their" organization for each DQ factor. Values for "weight" were obtained by calibrating the framework with the help of industry experts, who provided judgment on whether a DQ factor affected the selected decision category or not, and if yes, how much does it impact on a scale of "High / Medium / Low". Values for "confidence level" were obtained from a set of actual users of the DSS under study to capture their experience in the quality of data that exists in the System under study. This is again captured on a scale of "High / Medium / Low". At the end of these steps, the mathematical model (refer DQ expression following DIAGRAM XV. , page no. 84) is used to compute the DQ score for each decision category.

Measuring DQ through a composite score (such as the one proposed through this research work) can be tricky, since such a measure is expected to meet both precision and practicality expectations. These 2 are conflicting goals and achieving a balance remains a significant aspect of the framework. A DQ score should reflect reality as precisely as possible; yet, the assessment method adopted for such a measure should be as practical as possible. Any assessment method should be understood by the user and should be easy to adapt. Despite the sizeable body of literature available on DQ, relatively few researchers have tackled the difficult task of quantifying some of the conceptual definitions DQ. In fact, a general criticism within the DQ research field is that most approaches lack methods or even suggestions on how to assess quality scores, Naumann and Rolker, 2000.

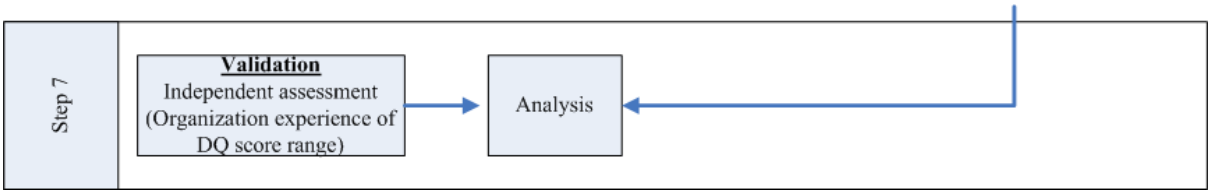
The DQ score computed by the framework introduced through this work represents an objective view of the quality of data that exists in the Decision Support System under study,

for each of the decision categories and thus serves as the basis to link DQ to the quality of decisions derived from DSS. As explained in this section, this work introduces a framework that is comprehensive, based on existing DQ assessment methodology, context sensitive and addresses the gaps in current literature summarized at the end of Chapter 2. This approach also addresses the key concern raised by Kaiser, 2010 i.e. "What is still lacking is an approach towards a unified measure of DQ based on metrics for different dimensions." Emphasizing the need for research in DQ assessments and frameworks, Sadiq et. al., 2011 have concluded that even in those themes and topics where there have been the most significant contributions from (past) research, i.e. DQ assessment and DQ frameworks there remains major concern in industry (with research to research advances).

Step 7: Refer DIAGRAM XIV. Validate the scores. In this step, the DQ score computed through the previous steps is subjected to independent validation. A different set of users are asked to provide their overall assessment of the DQ as it exists in the DSS under study. This independent measure is based on the quality of business outcomes they have experienced from decisions arrived using the DSS under study. The objective of this independent assessment is to compare the DQ scores computed through the model with these objective assessments and carry out statistical tests, explained in the next Chapter. The responses from this second group of users served as the gold standard, and outputs produced by the framework were compared to the gold standard to determine the validity of the framework.

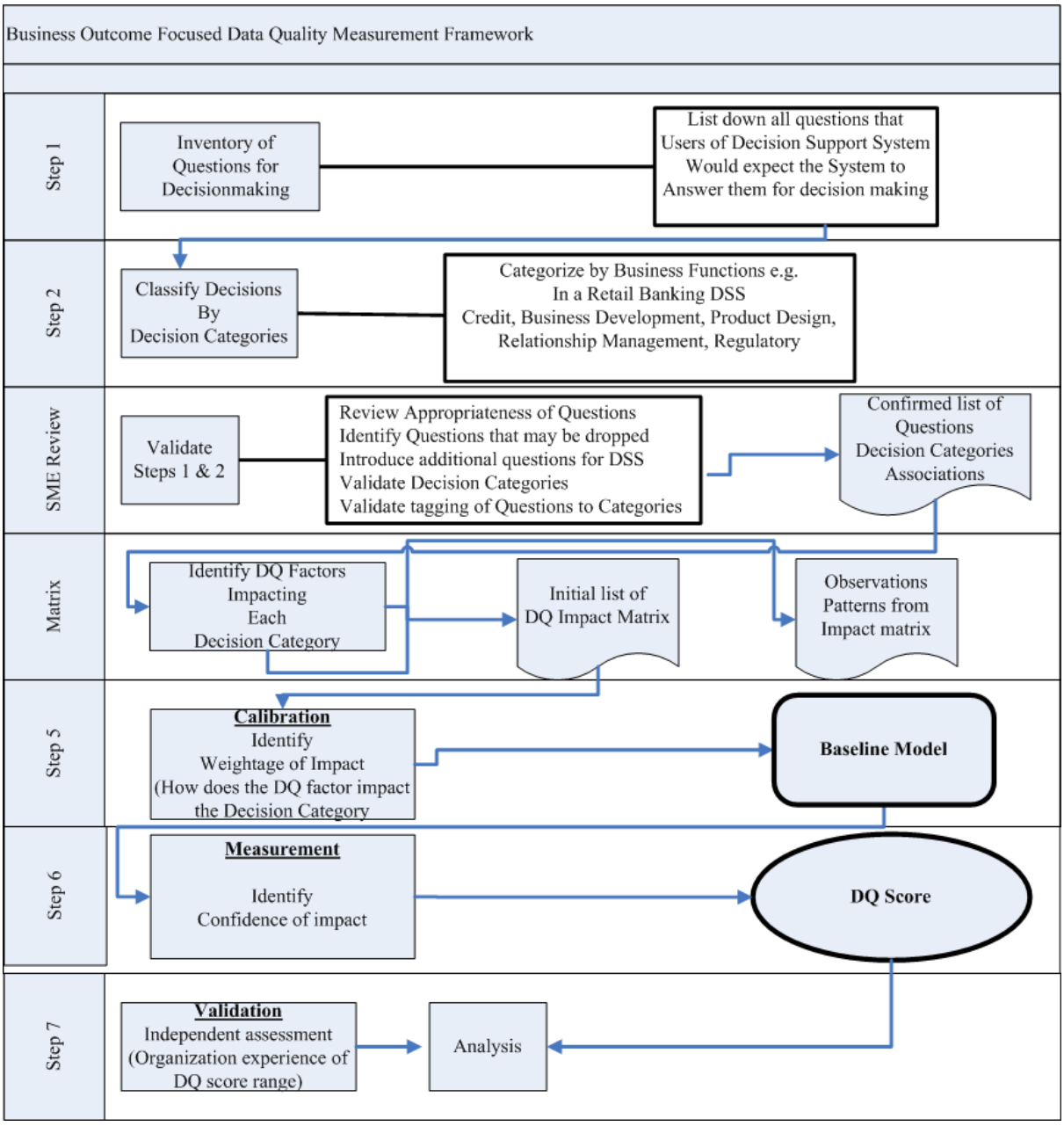
If the DQ score and the assessment values are in alignment (based on results from the statistical tests), it can be confirmed that the framework introduced through this work has been appropriately implemented for the selected industry / organization. If not, steps 1 to 6 will need to repeated either with a larger subject population or the basis of context i.e. decision categories and/or DQ factor-Decision category mapping be revisited with more focus, to refine the context.

DIAGRAM XIV. STEP 7: VALIDATE DQ SCORE



The proposed framework can be implemented in a selected industry by following the steps described above. The proposed framework is context sensitive in that it differentiates different business domains and within a selected domain it recognizes the specific context associated with different functions and their decision making requirements. It is expected that this framework may be administered separately for different business functions to arrive at relative weighted scores. For example, the weighted scores for a Datawarehouse System that is used largely for Strategic decision is expected to be quite different from those that are designed for tactical decision support. However, within the scope of a business function and the DSS that support decision making for the said function, this framework is expected to give pointers towards the relative ranking of DQ factors that needs to be focused for DQ improvement and improve quality of decision making. The following Diagram depicts a comprehensive view of the framework for measurement of DQ that is proposed through this work:

DIAGRAM XV. DQ ASSESSMENT FRAMEWORK



As mentioned earlier, the above framework has been inspired from QAFD Methodology (page 76). The below diagram depicts a mapping of QAFD Methodology to the above framework.

Research problem and hypothesis

Hypotheses based on the above model are discussed in the following section.

Null Hypothesis: A direct association does not exist between business outcomes and Information Quality factors in Banking domain.

Alternate Hypothesis: A direct association exists between business outcomes and Information Quality factors in banking domain.

This association is further influenced by the relative importance of the DQ factors and confidence of users of Information Systems on the quality of data contained in the underlying decision support system. This direct association can be identified by measuring the relative importance (weightage) of the DQ factors and confidence in the data; moreover the strength of the association can be measured through these influencing components.

Considering the possibility that the framework can be tested and analyzed under different decision categories, a set of decision-support specific hypotheses were identified which are detailed in Chapter 4.

The above association can be measured and expressed through the following equation:

$$\text{DQ Score (d)} = \frac{\sum_{ieDQ}^{n} w(i)c(i)a(i,d)}{\sum_{ieDQ}^{n} a(i,d)}$$

where

d = Decision Category (i.e. Credit, Tactical, etc.)

DQ = DQ Criteria

w(*i*) = Weight (and hence impact) of DQ factor *i* on decision category *d*.

c(*i*) = Confidence in the quality of DQ factor *i*

a(i,d) = Whether DQ factor *i* applies to decision category *d* – 0 or 1

such that $\sum_{ieDQ}^{n} a(i,d) <> 0$

The above hypothesis seeks to establish that primarily a relationship exists between the quality of business outcomes that an Organization derives based on decisions obtained from

an underlying Decision Support System. The hypothesis further seeks to prove that the above relationship is systematically measurable. The underlying philosophies of this work (and the above hypothesis) are as below:

The hypothesis seeks to prove the impact of 2 sets of influencing factors i.e. Relative Importance factors (weightage associated with each DQ factor) and Confidence Level of the users of Decision Support Systems on the quality of data existing in "their" organization, based on "their" experience of data as exists in "their" organization.

Chapter 4 – Empirical Study Design

Introduction

This chapter presents the details of the study designed and conducted, construction of survey instruments and design for data analysis. The first section describes the study that was designed to test the hypothesis underlying the research. The next section describes the setup activities that were performed and the outputs of such exercise. The next section presents the survey instruments that were designed for the study. The last section describes the statistical methods that were considered at different stages of the study, including techniques that were adopted for data analysis and study of results from the study.

Study Design

This section describes the way the study under this work was designed to test the hypotheses listed in the previous section.

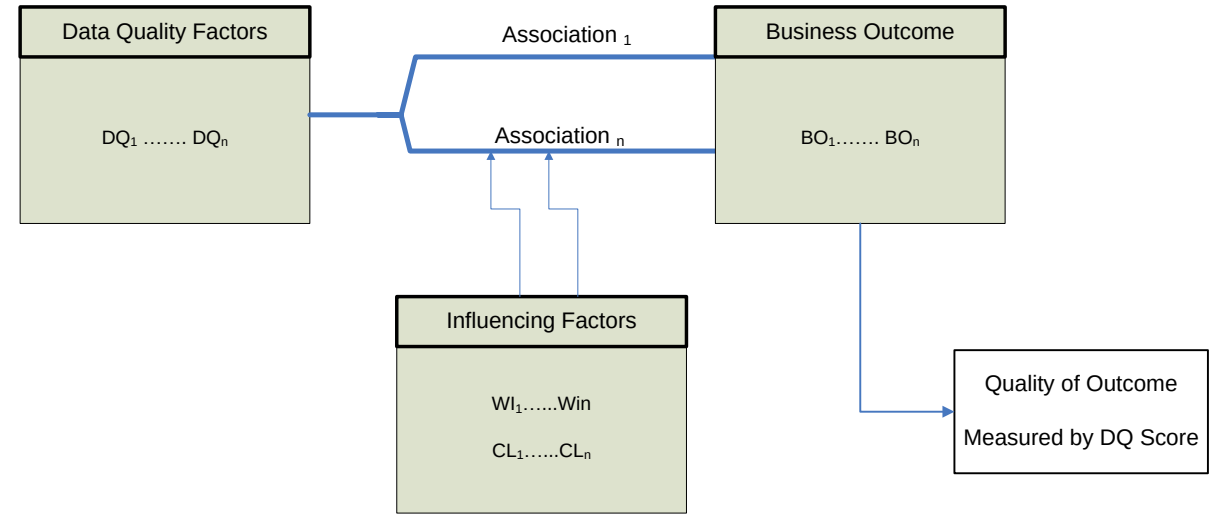
The research hypothesis (refer Chapter 3, page no. 85) seeks to establish that primarily a relationship exists between the quality of business outcomes/decisions and the quality of the underlying data within Decision Support Systems from which these decisions are based. The hypothesis further seeks to prove that the above relationship is systematically measurable.

The underlying philosophies of this work (and the above hypothesis) are as below:

The DQ measurement framework introduced in the previous chapter differs from the existing framework as it distinguishes itself from the "one size fits all" or "one definition of DQ captures it all" syndrome; it is context sensitive in that it distinguishes between different business domains (Banking, Telecom, Retail etc.) and within a selected domain it recognizes the specific sensitivities associated with different functions and their decision making requirements.

A diagrammatic representation of the hypothesis of the research is as below:

DIAGRAM XVI. RESEARCH PROBLEM AND HYPOTHESIS



Where, BO = Business Outcome, DQ = Data Quality Factor, WI = Weightage, CL = Confidence Level

Hypothesis:

A direct association exists between business outcomes and Information Quality factors in Banking domain.

- This association is further influenced by the relative importance of the DQ factors and confidence of users of Information Systems on the quality of data contained in the underlying decision support system.
- This direct association can be identified by measuring the weightage of the DQ factors and confidence in the data.

An empirical study was designed and carried out to test the hypothesis. The overall approach of the study was to involve the actual business users and experts from the Industry with the following key objectives in mind:

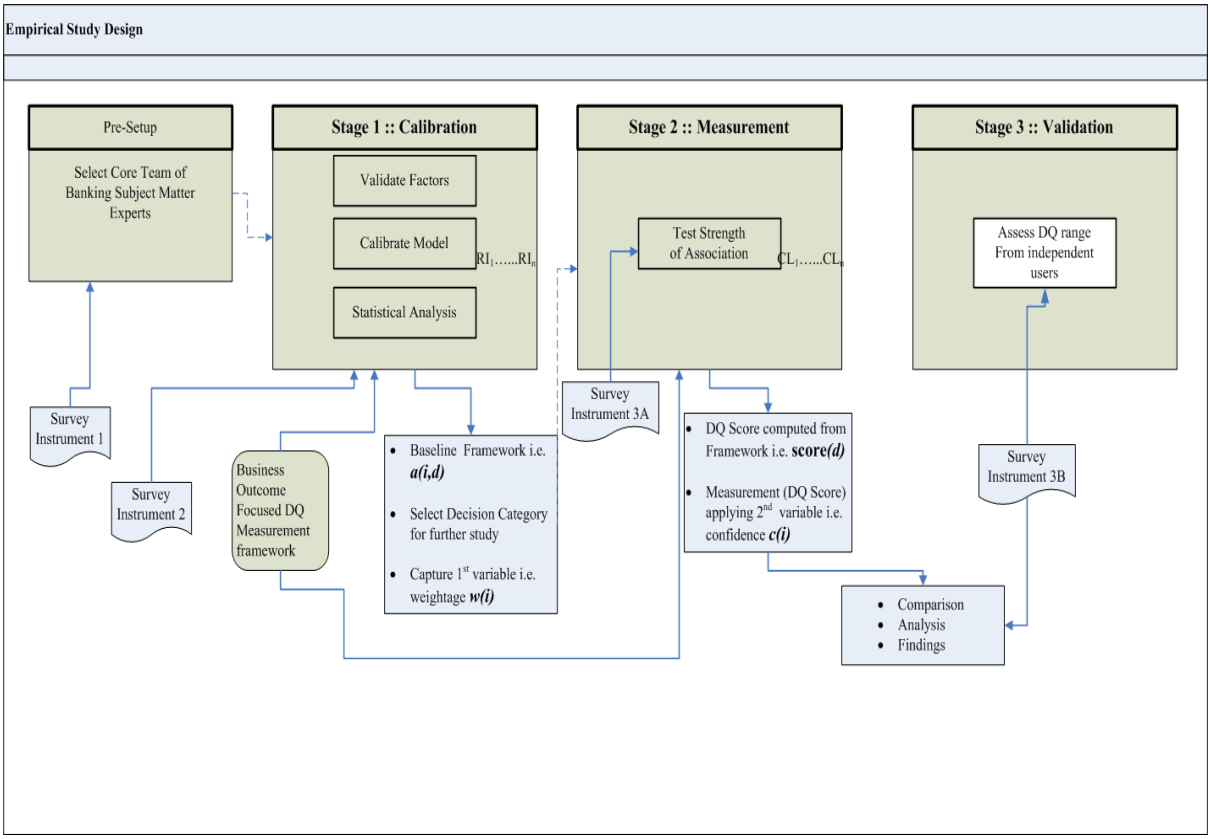
It was essential to keep the subjects of the population of the study / data collection to be homogeneous. To meet this objective, involving randomly selected individuals and/or organizations was not considered ideal; the reason being that one of the key elements in the model being tested was ***confidence factors*** and this confidence factor had to be based on identified target set of subjects to bring the actual experience in the environment that is being studied, as the underlying philosophy of the model is being context sensitive.

Often times, data collected through public surveys, lack the important element of capturing the respondent's specific experience (whereas such data collection mechanisms are modeled to capture respondents' opinion on the selected subject); this limitation is likely to lead to incorrect survey data which may adversely impact the objective of the study; in order to overcome this limitation, it was felt that users' commitment to the study and providing a "take-away" to the respondent, would bring in more seriousness to the Survey response and thus the data capturing exercise. To meet this objective and to obtain good level of involvement of the subjects in the study, a toolkit on the study was prepared. This short toolkit contained information on the background of the study, briefly introduce the framework and explain the nature of survey / survey instruments that were intended to be sent out to the respondents and the expected time to be spent by them. A briefing session was conducted with target respondents to walk through the contents of the toolkit and that the target subjects were being reached out because of their length of experience in the industry and their expertise on the subject (in this case retail banking). This recognition of expertise served as a motivation factor for the experts to involve in the study.

The 3rd objective considered was to group the influencing factors and seek inputs for such groups from different sets of target audience. Accordingly, the influencing factors were categorized as those that can be identified by experts (relation importance or weightage) and the rest as that can be identified by practical users of Decision Support Systems (confidence factors). Accordingly, different stages were designed for the study and appropriate survey instruments were deployed to capture such specific data.

The overview of the design of the study is depicted in the following diagram and a high level flow of the study is described in the paragraphs following the diagram.

DIAGRAM XVII. DESIGN OF EMPIRICAL STUDY



The empirical study in this research work was structured in 3 phases or stages and this structuring was required considering the multiple steps involved in implementing the framework referred in "Research Methodology" section of Chapter 3 (page no. 76). Pre-setup stage involved steps 1 to 4 of the framework.

- Stage 1 of the study covers step 5 of the framework. The focus of stage 1 of the study is to share the suggested model with the industry subject experts and subject the framework to calibration exercise. This stage also involves conducting statistical tests (explained in subsequent paragraphs) to ensure consistency of calibration outcomes. At the end of this stage of the study, the DQ scoring model is established with appropriate weightage factors assigned for each decision category that serves as the basis for stage 2 of the study.

- Stage 2 of the study maps to step 6 in implementing the framework. This stage involves administering Survey Instrument 3A. This survey obtains confidence factors in DQ, which when applied to the DQ scoring model base lined in the previous stage generates a set of DQ scores i.e. DQ score for each decision category.
- Stage 3 of the study maps to step 7 in implementing the framework. This stage involves administering Survey Instrument 3B. This survey obtains independent assessment inputs that serve as a validation of the DQ scoring model. Based on the DQ scores generated and the independent assessment inputs, statistical tests are conducted to test the Hypothesis set out for research.

Sampling methodology

This paragraph explains the approach towards sample size and sampling methodology for the empirical study. As explained in a previous paragraph, it is critical to conduct the study in selected organization units in order to get inputs from a set of homogeneous subjects. There were 2 levels of sample sizes considered – i.e. organization to be targeted for the study and the individual participants within the selected organizations.

As a first step, the list of public sector and private sector banks in India was identified (source: Indian Banks' Association @ www.iba.org.in/). This initial list contained a total of 27 Banks (20 from public sector and 7 from private sector). The objective was to cover maximum number of Banks from this list as part of the study. It was not feasible to include many of them due to practical reasons, primarily due to the current state of availability (or non-availability) of appropriate technical environment i.e. effective DSS backed up by enterprise DW Systems in these Banks. In its recent report, the Reserve Bank of India, 2011 had called out this state of technology adoption by Banks in India "*adoption of Core Banking Solutions has emerged as a single most significant innovation which has transformed the way banks have managed their businesses. However, these systems have not been fully exploited for information management and decision support.*" The said report further lists the below areas as major gaps that needs to be addressed during the ensuing years.

- issues in integration of information
- focused approach in usage of data for MIS and DSS
- inadequacies in information needed to take vital decisions
- disparate IT systems at different levels of maturity
- adoption of data mining and business analytics for information refinement

The said report further states that over the last decade and a half, attention has been focused particularly on operationalizing and maintaining the payment systems in a safe, secure and efficient manner. Consequently, the implementation of distinct information systems to support decision making activities could not be accorded equal priority. These observations were with respect to the current state of technology adoption by Banking sector in India.

However, there exists a few Banks which had made good progress in technology adoption and had embarked on their technical journey to implement DSS / DW systems for decision making process. As a second step, the initial list of banks was refined to capture the existence of DSS/DW for decision making and this information was captured in 2 forms i.e. banks that have implemented DW and those that were in the process of implementing DW in phases. This resulted in 7 Banks that were already using DW for decision making and the other 7 that were in the process of implementing DW, reducing the scope of study from the initial list of 27 Banks to 14 (refer TABLE XIV.). Sample size calculators were used (for this target population of 14 Banks with a confidence interval of 10) that pointed to a sample required as 12 (confidence level 95 %) or 13 (confidence level 99 %). The author approached these banks with the toolkit discussed earlier. These sample subjects can be categorized into 3 sets:

- Banks familiar to the author through professional association
- Banks in which the author has strong and longstanding relationship
- Banks which are the Author's employer's client organization (author has familiarity with these banks in the capacity of serving their Information Technology requirements)

TABLE XIV. BANKS COSNIDERED FOR DATA COLLECTION

S No	*Name of the Bank*	*Status of DW System*	*Scope of DW System*
Public Sector Banks			
1	Allahabad Bank	Not present	
2	Andhra Bank	Not present	
3	Bank of Baroda	Implementation underway	Customer data analysis
4	Bank of India	Implementation underway	Consolidate customer data
5	Bank of Maharashtra	Not present	
6	Canara Bank	Implementation underway in phases	Enterprise wide
7	Central bank of India	Implementation underway in phases	Enterprise wide
8	Corporation Bank	Not present	
9	Dena Bank	Not present	
10	Indian Bank	Not present	
11	Indian Overseas Bank	Operational	Enterprsise wide
12	Oriental Bank of Commerce	Not present	
13	Punjab & Sind Bank	Not present	
14	Punjab National Bank	Operational	Enterprise wide
15	Syndicate Bank	Implementation underway	Enterprise wide
16	UCO Bank	Operational	Customer data
17	Union Bank of India	Implementation underway	Enterprise wide
18	United Bank of India	Not present	
19	Vijaya Bank	Not present	
20	State Bank of India	Implementation underway in phases	Enetperise wide
New Private Sector Banks			
1	Axis Bank Ltd.	Not present	
2	Development Credit Bank Ltd.	Not present	

S No	*Name of the Bank*	*Status of DW System*	*Scope of DW System*
3	HDFC Bank Ltd.	Operational	Enterprise
4	ICICI Bank Ltd.	Operational	Enterprise
5	Indusind Bank Ltd.	Not present	
6	Kotak Mahindra Bank Ltd.	Operational	Credit Card data consolidation
7	YES Bank	Operational	Enterprise

Across the above categories, the author approached 12 Banks for their consent to participate in the study. Sampling methodology involved applying parameters such as easy access to subject matter experts, existence of DSS, turnover exceeding $ 500 MN, 1000+ branches etc. Many of the target subjects had stringent confidentiality requirements either in sharing profile information of their experts or details of their DSS, primarily due to factors such as to sensitivity of data, data security requirements and highest level of confidentiality that Banks needed to maintain for their competitive advantage. Based on these responses, the author signed up to work with the shortlisted Banks (that had implemented an enterprise wide DW System for its decision making) with significant retail banking business (above INR 5000 Crores), larger national and international network. This involved signing up a non-disclosure agreement as it pertains to the name of the Bank or its officials involved in the empirical study or details of its DW System / DSS.

Having selected the organization for the study, the next step involved defining an approach to select the individuals and the sample size of such target population. The next step in the data collection exercise was related to computing sample size for DQ scores that need to be validated to address the research question. The total possible DQ scores were 72 (12 eligible banks and 6 decision categories each). Applying the same sample size calculator pointed to target a minimum of 50 DQ score comparisons. As an accepted practice in the discipline of sampling, the first step in sample design is to ensure that the specification of the target population is as clear and complete as possible to ensure that all important characteristics within the population are represented.

The topic being researched is a specialized subject (DQ in DSS in retail banking) and requires inputs from experts with experience in the selected industry. This prompted the use of purposive sampling (a non-probability sampling technique) to identify the selected individuals within the Banks to provide necessary inputs required for rest of the study. Tansey, 2007 dealt with the topic of how researchers should approach sampling their subjects, while, in particular arguing for the use of non-probability sampling approaches to elite interviewing. This cited work further suggests that in order to pursue non-probability sampling, researchers need to consciously consider the criteria they will use to select their respondents. Purposive sampling is a selection method where the study's purpose and the researcher's knowledge of the population guide the process of selection. The basic assumption behind purposive sampling is that with good judgment and an appropriate strategy, researchers can select the cases to be included and thus develop samples that suit their needs.

Following these principles, a process was developed that involved selection of individuals based on the below parameters, from their initial response to the demographics survey.

- The respondent has experience in retail banking function
- The respondent has a minimum of 6 years experience in the selected field
- The respondent has been using DSS / DW System for a period between 1 to 2 years, at the minimum. This criteria was necessary, because, only based on this usage the respondents are expected to develop an understanding of the DSS, DQ existing within the said DSS and importance of DQ factors (listed in TABLE XII.) for the decision making process.
- The respondent belonged to any of the functions involving retail banking decision making e.g. Management or Executive or finance / accounts or administration

Note: In a related work Rudra et. al., 1999 suggest that the respondents could be classified into two categories according to the length of time they have been employed in their respective organizations and that that the respondents who have had 6 years or greater

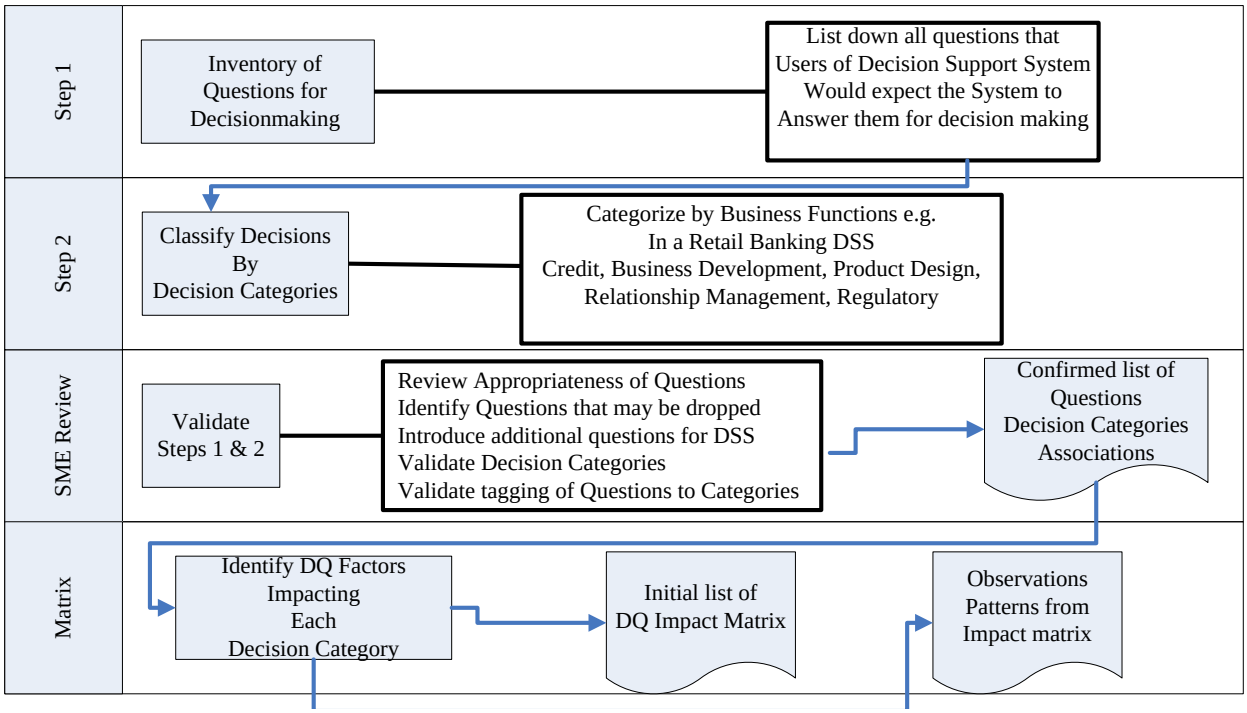
working experience in the organization had a more realistic opinion of the purpose of the warehouse.

This step led to identification of homogeneous groups that works in same business function, uses same DSS etc. Survey Instrument 1 captured the demographic information of the target population (individuals within the Bank). This information was used to identify such experts with length of experience in Banking Industry and thus the target subjects for calibration (subject matter experts were identified). Through this process top ranking officials of the Bank were identified to provide calibration inputs. Similar exercise was performed to use demographics information to identify users of DSS in the middle management function of the Banks. This methodology of sample size and sampling resulted in 300 data points of comparison (for DQ score) to validate the framework. This is considered to be a good number of data points for the intended empirical study; for example, Shankaranarayanan et. al., 2012, conducted a similar exercise to study DQ metadata and decision making, however with students from business schools as target population, generating about 60 responses. The referred work calls out sample size as a limitation suggesting extension of the work to "larger sample of decision makers in a real-world setting". The number of data points generated for study in the present research work is significant both in terms of numbers (nearly 5 times the population size of the above referred work) and source i.e. real-life industry practitioners.

Study set-up activities

As part of the research, the author carried out the initial steps that served as predecessor to the detailed study to test the Hypothesis listed in the previous chapter. These steps were part of the research framework (DIAGRAM XV.) referred in the previous chapter. The first 4 steps of this framework (highlighted in DIAGRAM XVIII.) formed the study setup activities.

DIAGRAM XVIII. PRE STUDY SET-UP STEPS



Step 1: The objective of the framework introduced through this research work is to approach the problem of DQ not from the perspective of measuring extent of quality itself, but, by measuring it through a factor of its impact on business decisions. Accordingly, the framework identifies a list of DQ criteria (refer TABLE XII. , Chapter 3, page no. 75) and potential decision categories, maps them on a many-to-many basis. Once the mappings are performed, an additional assessment of the extent of impact of a DQ factor on a decision category and the confidence level of each DQ factor is conducted to arrive at a weighted average DQ score that represents the quality of decisions from Decision Support Systems.

Identify list of potential decisions from DSS: In order to relate the DQ Factors to outcomes, it is imperative that we identify the key decisions that are typically taken by users of DSS. Of course, these decisions vary by Industry, type or size of the Organization, Function within the Company, role and level of the decision maker in the organization etc. As such, a comprehensive list of decisions applicable to all the above scenarios would become too farfetched for the purpose of this study. At the same time, it will be essential that the framework that we develop distinguish the variables depending on the above factors and the constants, irrespective of the above scenarios and thus yield itself to a maximum degree of reusability at all times. Therefore, it was considered that the framework be subjected to deep analysis for a specific function within a selected Industry and analyze applicability of DQ factors (listed in TABLE XII.) and their impact on quality of outcomes, relevant to this function / Industry. With these objectives in mind, this Study will focus on the Retail Banking Industry.

It is expected that this model will be administered separately for different purposes to arrive at relative weighted scores. For example, the weighted scores for a DSS that is used largely for Strategic decision is expected to be quite different from those that are designed for tactical decision support.

Step 2: Identify decision categories: Since the Hypothesis and the underlying model is organization and business function specific, it is essential to group the decisions into logical groups, named in this model as "decision categories". Information for decision making in a business can broadly be broken down into three types:

- That required for regulatory, financial and tax reporting, without which the business could not legally function and which must be produced quickly, accurately and efficiently for the purpose of external reporting (operational)
- That required to manage day-to-day situations and decisions (tactical)
- That needed to support longer-term decision making and the development of strategy. (strategic)

Step 3: SME Review: The model developed so far is based on requirements and experience from past Decision Support projects; it is critical to validate these experiences and their summarization in the form of model in this work. This step is proposed to engage Banking experts, leverage their experience, to validate the findings so far with the following activities:

- Review appropriateness of the questions that have been identified in Step 1
- Identify questions that may not be appropriate and have to be dropped
- Introduce additional questions for DSS
- Validate decision categories based on above

Step 4: Evolve DQ Matrix: In this step it is proposed to get a view of the likely impact of the DQ Factors (the number of factors impacting the decision category on the quality of outcomes). This examination is a key component of the framework, since, this is likely to identify the DQ factors that have greater influence on outcomes from DSS or those that are unlikely to have any impact on business outcomes and thus help filter out DQ factors (from previous sections) that are not relevant for rest of the study. At this stage, only existence or otherwise of impact is identified at this stage and not "how" a DQ criteria impacts decisions or "how much" a DQ criteria impacts such decisions, which will be subject of rest of the framework.

In the remainder of this section, the author describes how the first 4 steps of the Setup activities are implemented. To implement Step 1, the author referred to the Functional Requirements documents of several DW Projects in the retail banking industry; based on the requirements stated in these documents, the author identified several business decisions that users of DSS typically expect in retail banking. Listed below is a sample of the business decisions that were identified as related to retail banking:

- Do I need to create a new customer code or is there existing code?
- Do I need to associate a new customer code with any class of customers?
- Are there any specific restrictions or approvals that need to be verified for this customer?
- Whether or not to extend a specific status (e.g. preferred customer or classic customer) to a selected customer

- What products do we offer to this customer?
- How much credit limit to allow for this customer?
- What is the risk profile that needs to be assigned to this customer?
- Should this customer be included in a specific promotion scheme or not?
- What should be rate to be quoted for a loan product to this customer?
- Which segment of customers should the Bank choose to sell a specific loan or deposit product?
- Which class of customers should we follow up more vigorously to improve the loan recovery status?
- What is the lowest (or highest) interest rate that can be quoted to this customer for a new loan product, based on the risk profile?
- When should the Bank initiate request for additional security from a particular customer?
- When can the Bank make the next cold call to this client for upselling?
- Should the customer be tagged for additional credit checks or background verification?
- Is the customer eligible for specific incentives e.g. additional interest for Senior Citizens
- Can the Bank approach the customer to seek referral customers?
- Can the customer's information be shared to database companies?
- Can a customer's request for a specific fund transfer be approved (i.e. do we need joint approval or single or surviving)
- Can a discretionary authority be used to benefit a customer (e.g. honor a check in overdraft mode)
- Can request for incremental loan be from a specific customer be approved?
- Is the aging bucket for loans collections and follow up appropriate?
- Whether a product (loan or deposit) needs to be discontinued in specific market segments?
- Which specific channel needs to be explored for promoting a specific loan or deposit product?

- Which specific segment of customers should be targeted for new product offering? (multiple factors of decision are involved such as choice of location or age group or class of customers etc.)
- Can the Bank waive off specific charges that are otherwise applicable, for a selected customer?
- Whether to include a specific account in the physical account statement distribution list or not?
- Whether a specific customer needs to be called for personal greetings (e.g. Birthday wishes)
- Do the customers' transactions warrant any specialized reporting e.g. Anti Money Laundering or Fraud Investigation etc.
- Fee based services:- based on fees income from the customer in this month, is the customer eligible for a discount in fees for a selected transaction?
- Fee based services:- Is the current request from a customer within the limits applicable for services that can be availed in a specified time period – e.g. $ 10 MN bills discounting in a calendar month
- Delinquency – does the loan account quality for write off based on age of delinquency?
- Delinquency – are there any charges that need to be applied to the Customer for delinquency status?
- Delinquency – When is the next legal follow up / attorney letter to be sent to the customer?

To group and validate these decisions (and hence implement Steps 2 and 3), the author engaged with a team of Banking Domain Experts from his organization. These experts helped in further refining this list of decisions, either by rephrasing some of these questions or deleting a couple of them as not relevant or by adding new business questions that were not envisaged by the Author. Again, with the help of these Banking experts, the decisions were logically grouped into decision categories listed in TABLE XV.

TABLE XV. DECISION CATEGORIES IDENTIFIED

S No	*Decision Category*
1	Credit Decisions
2	Business Promotion Decisions
3	Product Decisions
4	Tactical Decisions
5	Relationship Decisions
6	Regulatory Decisions

Extending the alternate hypothesis from Chapter 3, considering the possibility that the framework can be tested and analyzed under the above listed decision categories, a set of additional hypothesis were developed and subjected to study and tests elaborated in subsequent paragraphs of this Chapter.

TABLE XVI. HYPOTHESIS OF THE RESEARCH WORK

Hypothesis Number	*Decision Category*	*Null Hypothesis*	*Alternate Hypothesis*
H1	Main research problem	There is no relationship between DQ factors and Business Outcomes in Banking domain	A direct association exists between business outcomes and DQ factors in Banking domain
H2	Business Promotion	For business promotion decisions, there is no relationship	For business promotion decisions, there exists a direct relationship between DQ and Business

Hypothesis Number	***Decision Category***	***Null Hypothesis***	***Alternate Hypothesis***
		between DQ and Business Outcomes	Outcomes
H3	Credit Decisions	For Credit decisions, there is no relationship between DQ and Business Outcomes	For Credit decisions, there exists a direct relationship between DQ and Business Outcomes
H4	Product Design	For product decisions, there is no relationship between DQ and Business Outcomes	For product decisions, there exists a direct relationship between DQ and Business Outcomes
H5	Regulatory	For regulatory decisions, there is no relationship between DQ and Business Outcomes	For regulatory decisions, there exists a direct relationship between DQ and Business Outcomes
H6	Relationship	For relationship decisions, there is no relationship between DQ and Business Outcomes	For relationship decisions, there exists a direct relationship between DQ and Business Outcomes
H7	Tactical	For tactical	For tactical decisions, there

Hypothesis Number	***Decision Category***	***Null Hypothesis***	***Alternate Hypothesis***
		decisions, there is no relationship between DQ and Business Outcomes	exists a direct relationship between DQ and Business Outcomes

To implement Step 4, an objective mapping of each of the DQ factors to the applicable decision categories was conducted based on the above definitions and the decision categories listed above. This initial mapping is done based on experience shared by DW projects referred in the previous paragraph. Once again validation inputs from industry were obtained. The Banking Domain Experts validated the matrix by revisiting the appropriateness of the decision categories and mapping of DQ factors to the decision categories; recommendations to revise the mappings initially established by the Author were implemented. With this step, the working model was established that carried decision categories as applicable to Banking together with a matrix of DQ factors as applicable to those decision categories. The following matrix emerged from this exercise, providing an association of the DQ factors listed earlier to the decision categories listed above.

TABLE XVII. DECISION CATEGORIES – DQ FACTORS MATRIX

DQ Factor / Decision Category	*Credit*	*Business*	*Product*	*Tactical*	*Relationship*	*Regulatory*
Believability	Yes	No	Yes	Yes	No	Yes
Concise representation	No	No	No	Yes	Yes	No
Interpretability	Yes	Yes	Yes	Yes	No	No
Reputation	Yes	No	Yes	No	No	Yes
Understandability	Yes	No	Yes	Yes	No	Yes
Value-added	Yes	Yes	Yes	No	No	Yes
Granularity	Yes	Yes	Yes	No	No	Yes
Relevancy – Measures	Yes	Yes	Yes	No	No	Yes
Relevancy – Dimensions	Yes	Yes	Yes	No	No	Yes
Aggregation	Yes	Yes	Yes	Yes	No	No
Completeness	Yes	Yes	Yes	Yes	Yes	Yes
Customer Support	Yes	No	No	Yes	Yes	Yes
Documentation	Yes	Yes	Yes	No	No	Yes
Objectivity	Yes	No	Yes	Yes	Yes	Yes
Price	No	No	No	No	No	No
Reliability	Yes	Yes	Yes	Yes	Yes	Yes
Security	Yes	Yes	Yes	Yes	Yes	Yes
Accuracy	Yes	No	Yes	Yes	Yes	Yes
Availability	Yes	No	No	No	Yes	Yes
Consistency	Yes	No	No	Yes	No	Yes
Latency	No	No	No	No	No	No
Response Time	No	No	No	Yes	No	No
Timeliness	Yes	No	No	Yes	No	Yes
Verifiability	Yes	No	Yes	Yes	No	Yes

The following observations and decisions were made based on the above mapping and this resulted in setting up the baseline framework for further experiment and research.

1. DQ factors that do not seem to impact any of the Decision Categories may be dropped from further analysis i.e. Price and Latency

2. For the purpose of materiality, DQ factors that do map to less than 3 Decision Categories viz., Response Time (1) and Concise Representation (2) are not considered for further analysis.
3. The table below provides a view of the likely impact of the DQ Factors based on the number of factors impacting the decision category on the quality of outcomes. The potential impact on quality of business outcomes are indicative and are based on the banking experts engaged in previous steps. However, these observations do not have any impact on the rest of the study.

TABLE XVIII. OBSERVATIONS FROM MAPPING EXERCISE

Decision Category	***No. of DQ Factors***	***# of decisions***	***Potential impact on quality of business outcomes***
Credit Decisions	20	8	Credit decisions vary from smaller $ value of transactions to potentially huge values; besides, the credit decisions are taken not just in relation to a single customer or transaction, rather, these are taken in relation to a group of customers; e.g. which class of customers to follow up for loan recovery? Hence, the impact of outcome is very high from a direct $ value associated with the decisions.
Business Promotion Decisions	10	6	Business promotion decisions are taken with a view to guide the decision maker on a select customer or a class of customers that can be approached for cross selling. As such there is no direct $ value involved; even if involved, the

Decision Category	*No. of DQ Factors*	*# of decisions*	*Potential impact on quality of business outcomes*
			probability of the promotion resulting in business is not often predictable. The only direct $ value associated with the decision is the cost of carrying out business promotion /campaign. In view of the uncertainty involved in the outcomes, even if the decision maker were provided with information of very high quality, the impact is marked as low. More often, the impact of incorrect decision in this category implies more of a notional cost e.g. potential loss of business due to targeting an incorrect customer class.
Product decisions	16	4	There are both direct $ value and indirect $ value associated with these decisions. E.g. decision related to continuation or discontinuation of a product in a market segment implies an direct (or notional) cost as above; whereas, rate to be fixed for a selected product lines has a direct $ value associated, as any incorrect decision in this regard may result in loss of revenue.
Tactical decisions	15	8	Tactical decisions need to be valued based on the $ value associated with

Decision Category	*No. of DQ Factors*	*# of decisions*	*Potential impact on quality of business outcomes*
			the underlying transaction; e.g. incorrect approval of a fund transfer may result in low or heavy damages based on the value of the funds being transferred on a case to case basis. Unlike credit decisions, tactical decisions are taken just in relation to a single customer or transaction. Another cost that can be potentially incurred based on tactical decisions is the cost of damages and law suits. E.g. incorrect approval of a credit request may lead to the customer claiming heavy damages due to mental pressure etc. Hence, the impact of outcomes are general high from a direct $ value associated with the decisions and also potential indirect $ value
Relationship Decisions	8	4	These decisions appear to be more of "good to have" and do not have greater influence on the $ value of associated outcomes.
Regulatory decisions	18	4	These decisions are mostly post-event reporting in nature; as such the quality of the decision do not alter the direct $ value associated with the decisions; However, any incorrect decision is likely to result in non-

Decision Category	*No. of DQ Factors*	*# of decisions*	*Potential impact on quality of business outcomes*
			compliance and may attract penalties or fines from the Law Enforcing Agencies, based on the nature of the compliance requirements. Since, these cannot be quantified and are also only potential in nature, these factors leave a medium impact on the decisions.

Survey Instruments Design

Before describing Stages 1 and 2 of the study, this section describes the design of various survey instruments used in these stages.

The survey items were based on existing items from validated instruments found in the research literature (Kahn et. al., 2001). Many of the survey items had been widely validated in a variety of populations and organizational settings, while others had been validated in more limited contexts. Lee et. al., 2002 observe that "despite a decade of research and practice, only piece-meal, ad hoc techniques are available for measuring, analyzing, and improving IQ in organizations". In response to this situation they developed a measurement instrument, known as the Information Quality Assessment (IQA), which measures stakeholder perceptions of each dimension. This instrument, which employs 69 items to measure the various information quality dimensions, has been used as the basis of several studies requiring information quality measurement e.g. Slone, 2006; Kahn et. al., 2002 and Pipino et. al., 2005 as well as for studies that extend this measurement concept further. As such, the survey instruments used in a recent related work Slone, 2006 was used as the basis and further modified as appropriate for study conducted in this research work.

Empirical Study Design: Setup Questionnaire

In general, data collection in the past works was targeted at anonymous target subjects. For example, Slone, 2006 used the contacts database of a vendor- and technology-neutral industry consortium; attendees at the consortium's conference were selected for administering the surveys for data collection. However, since the focus of this study is to address one of the major gaps i.e. lack of context sensitive approach to measurement of DQ, the author decided to approach specific target users and the reasons are elaborated below. As a first step in this direction, Survey Instrument 1 was administered to a wider set of Banking Professionals. This instrument primarily captured demographics data and professional background of the potential respondents, so that the target audience for rest of the data collection exercise could be identified based on their background, experience, role and other factors. The overall approach of the study (data collection process) was to involve the actual business users and experts from the Banking Industry with the following key objectives in mind:

It was essential to keep the subjects of the population of the study / data collection to be homogeneous. To meet this objective, involving randomly selected individuals and/or organizations was not considered ideal; the reason being that one of the key elements in the model being tested was ***confidence factors*** and this confidence factor had to be based on specific set of subjects, as the underlying philosophy of the model is being context sensitive.

Often times, data collected through public surveys, lack the important element of capturing the respondent's specific experience (whereas such data collection mechanisms are modeled to capture respondents' opinion on the selected subject); this limitation is likely to lead to incorrect survey data which may adversely impact the objective of the study. In order to overcome this limitation, it was felt that users' involvement in the study would bring in better quality responses for the Survey response.

The 3rd objective considered was to target different target groups for different variables associated in the model, based on appropriate profile of the target groups and their background to provide appropriate values for the variables under study. To elaborate, it is appropriate to collect calibration data (weightage) from more experienced professionals in the

Industry as opposed to others and this requires targeted administration of the survey. Accordingly, the variables were categorized as those that can be identified by experts (weightage) and the rest as that can be identified by actual users of DSS (confidence factors). Accordingly, different study stages were designed and appropriate survey instruments were deployed to capture such specific data.

Based on the above rationale, the 3 different survey instruments were created for use in different stages of the study. TABLE XIX. provides a detailed description of these survey instruments.

TABLE XIX. STAGES OF DATA COLLECTION

Study Design Stage	***Purpose***	***Instrument Reference***	***Target Audience***	***Target preferred Background***
1	Calibration of framework	Survey 2	Senior Bankers	Experience in Policy decision making Representation in high level decision making process (e.g. credit policy decisions for the Bank) Preferably exposure to Bank's IT Systems
2	Data Quality Confidence – Bankers	Survey 3 A	Branch Managers or Department Managers in HO (e.g. Credit Processing	Experience in understanding the rationale behind Policies and implementation of those policies Representation in operational decision making process (e.g. credit decisions) Preferably exposure to Bank's IT Systems

Study Design Stage	*Purpose*	*Instrument Reference*	*Target Audience*	*Target preferred Background*
			Managers)	Use of IT Systems for data analysis and decision support.
3	Data Quality assessment – Senior Bankers	Survey 3 B	Senior Bankers	Same profiles as in 2 and 3, but, different set of respondents

Empirical Study Design: Stage 1 – Model Calibration

After the set-up stage, it is important to capture the extent to which selected DQ criteria impacts a selected decision category; this impact captures industry level impact by identifying relative importance of factors (weightage associated with each DQ factor); this step is intended to add further context to DQ assessment, by associating extent of impact for each DQ-Decision Category combination.

To implement this stage and collect the necessary data for calibration, the author prepared a toolkit introducing the problem, the framework and data required for the study. A request was sent out to few Banks seeking their support in this study; along with the request, this toolkit was also sent out to set context of the study and the request. These Banks included medium to large sized Banks (which were either clients of the Author's employer or whose senior employees were professionally known to the Author). However, there were 2 roadblocks in getting acceptance from Banks – security considerations and time constraints. After repeated attempts, the author got concurrence from select set of banks as discussed in page no. 93(name and profile not to be disclosed due to non-disclosure requirements) that offered access to its officials in its Retail Banking Division to support data collection for the intended study.

A set of Banking subject matter experts that have experience in handling Retail Banking portfolio were identified; the identified subjects were administered a Survey Instrument (Survey Instrument 1) that focused on capturing the demographics of the respondents; these demographics contained data elements that helped determine the relative experience of the respondents in the Banking Industry, not just as users of decision support, but as business owners that influenced the decision making process. Once a set of senior Banking experts were identified, a briefing session was conducted to explain the research problem and introduce the framework proposed through this research work (Chapter 3, page no. 61).

The next step involved calibration of the framework. As part of this step the team of Banking experts was administered Survey Instrument 2 to capture 2 data points for each identified decision category i.e. whether a DQ factor impacted the business outcome for the selected decision category and if yes, to what extent does the DQ factor impact quality of business decision for the selected decision category. The impact was assessed using a scale of 3 points viz., High / Medium / Low impact. The objective of this step is to identify that part of the problem as it relates to "***how much*** do these factors influence the business outcome". Inputs from this survey exercise served as value for the weightage variable for each DQ for each Decision category - (*w(i)* referred in Chapter3, page no. 85).

At this stage, a critical question that the author had was whether or not to include the weightage factors as determined by the author (set-up activities), in the Survey Instrument 2. The advantage of including the factors was that this would provide some guidance (or baseline for the Banking experts to start response to the survey); whereas, the disadvantage was that it could carry potential risk of biasing the response of the experts. After several rounds of deliberations, it was decided to use 2 variations of the instruments (Survey Instrument 2A with the set-up values and Survey Instrument 2B without the set-up values) and administer them to 2 different sets of experts. It was also decided to conduct kappa statistical analysis (Carletta, 1996) to determine the degree of variance between the 2 approaches and introduce any corrections to the study design, if warranted based on findings from kappa statistical analysis.

Empirical Study Design: Stage 2 – Measure DQ

Measure DQ Score / Assess confidence level: At this stage, it is important to capture the extent to which, data quality exists in the organization under study for selected DQ criteria in the context of the selected decision category; this step captures organization level impact by capturing selected experience of users of DSS to identify extent of impact of each DQ factor for the type of decisions associated; this step is extremely critical as it provides values for key influencing factor in the process of DQ assessment.

As a next step, the team of users from the same Bank (identified using the same methodology described in the previous section) were administered Survey Instrument 3A to capture their confidence in DQ as it exists in their Organization; the objective of this step is to identify values for the variable *c(i)* of the research problem (referred in Chapter3, page no. 85).

Empirical Study Design: Stage 3 - Validation

Validation of the framework: To validate the results from the study (i.e. DQ scores computed by the model) an independent assessment was carried out to compare the DQ scores computed through the model with these objective assessments and carry out statistical tests as part of the validation process. As the final step, another team of users from the same Bank (again identified using the same methodology described in the previous section) were administered an independent survey to score the impact of DQ on each decision category as it exists in their Organization. The objective of this step is to validate the decision category score (i.e. impact of DQ on a decision category) calculated by the model with the feedback from users within the organization. The identified subjects were administered a Survey Instrument [Survey Instrument 3 B] for this purpose.

Statistical methods and instruments used in the study

Different statistical techniques were applied at different phases of the study either to validate reliability of survey responses or to analyze the study results. Usage of such techniques (kappa statistics, chi-squared test and Friedman test) are described in this section. The actual

results obtained from applying these statistical tests as part of the study described above and the inferences from them are discussed in detail in Chapter 5.

Kappa Statistics

As referred in the previous paragraph, it was important to compare data obtained for calibration of the framework, through different versions of Survey Instrument 2 – one that displays weightage for DQ factors as arrived at in the set-up activities and the other without revealing these values.

Researchers require evidence that people besides the authors themselves can understand and make the judgments underlying the research reliably. This is a reasonable requirement because if researchers can't demonstrate that different people can agree about the judgments on which their research is based, then there is no chance of replicating the research results. Published literature (Siegel et. al., 1988) shows strong argument for using kappa coefficient of agreement as a measure of reliability.

In a recent work, Sim and Wright, 2005 deal at length the applicability of kappa statistics for studies that are similar to the current research work. This cited work studied the appropriateness of kappa statistic to test the reliability of clinicians' ratings (which is an important consideration in areas such as diagnosis and the interpretation of examination findings). It was observed that these ratings lie on a nominal or an ordinal scale and that for such data kappa coefficient is an appropriate measure of reliability. In clinical practice and research, there is frequently a need to determine the reliability of measurements made by clinicians—reliability here being the extent to which clinicians agree in their ratings, not merely the extent to which their ratings are associated or correlated. Thus, 2 types of reliability exist: (1) agreement between ratings made by 2 or more clinicians (inter-rater reliability) and (2) agreement between ratings made by the same clinician on 2 or more occasions (intra-rater reliability). Sim and Wright, 2005 further elaborate these findings with examples from practice when they observe that in some cases, the ratings in question are on a continuous scale. In other instances, however, clinicians' judgments are in relation to discrete categories, which may be either nominal (eg., “present,” “absent”) or ordinal (eg., “mild,” “moderate,” “severe”); in each case, the categories are mutually exclusive and collectively

exhaustive, so that each case falls into one, and only one, category. These data require specific statistical methods to assess reliability, and the kappa statistic is commonly used for this purpose. Other examples of applicability of kappa statistics in practice include clinical diagnoses (Petersen et al., 2004) or classifications or assessment findings (Kilpikoski et al., 2002).

In clinical practice, a common situation in which a researcher may want to assess agreement on a nominal or ordinal scale is to determine the presence or absence of some disease or condition. This agreement could be determined in situations in which 2 researchers or clinicians have used the same examination tool or different tools to determine the diagnosis. Simple mechanisms that exist to gauge the agreement between 2 clinicians (i.e. overall percentage of agreement or effective percentage of agreement) do not take into account the agreement that would be expected purely by chance. If clinicians agree purely by chance, they are not really "agreeing" at all; only agreement beyond that expected by chance can be considered "true" agreement. Kappa is a measure of "true" agreement that indicates the proportion of agreement beyond that expected by chance, that is, the achieved beyond-chance agreement as a proportion of the possible beyond-chance agreement. Based on the above principle, Sim and Wright, 2005 describe a situation that is appropriate to use kappa statistics in practice, when they state that the simplest use of kappa is for the situation in which **2 clinicians each provide a single rating of the same patient**, or where a clinician provides 2 ratings of the same patient, representing inter-rater and intra-rater reliability, respectively. In the cited work, Sim and Wright studied the results of a hypothetical reliability study of assessments of movement-related pain. The assessment categories were on an ordinal scale with values "no pain," "mild pain," "moderate pain," and "severe pain."

A nominal level of measurement organizes data by name and thus refers to quality more than quantity. An example of a nominal scale is blood group, with the choices labeled A+ or A- or AB+ etc. An ordinal scale indicates direction, **in addition to providing nominal information**. An example of an ordinal scale is a measure of extent of pain, with choices labeled as low, medium, high or intense. In the current research work, as stated in the previous section (page no. 112) the study involved obtaining calibration inputs on an ordinal

scale to assess the extent of impact (High / Medium / Low impact of DQ factors on the decision and business outcome). As stated earlier (Sim and Wright, 2005) Kappa statistic is appropriate in measurement of reliability in ratings with nominal or ordinal scale and in a case where 2 clinicians each provide a single rating of the same patient. Applying these principles to the topic of current research, use of kappa statistics to measure degree of agreement between the calibrators was considered appropriate.

Kappa statistics could be effectively deployed to study degree of variation between 2 similar datasets. In scenarios where two individuals measure the same thing, Cohen's Kappa (referred as Kappa) is a statistical test to measure degree of agreement between the two individuals. In this exercise, the observed level of agreement with the value expected if the raters were totally independent. Kappa is always less than or equal to 1. A value of 1 implies perfect agreement and values less than 1 imply less than perfect agreement. Statisticians suggest that the below interpretations of Kappa.

- Poor agreement = Less than 0.20
- Fair agreement = 0.20 to 0.40
- Moderate agreement = 0.40 to 0.60
- Good agreement = 0.60 to 0.80
- Very good agreement = 0.80 to 1.00

The kappa coefficient (K) measures pair-wise agreement among a set of respondents making category judgments, correcting for expected chance agreement and is computed as below:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

Where $P(A)$ is the proportion of times that the coders agree and

$P(E)$ is the proportion of times that we would expect them to agree by chance

When there is no agreement other than that which would be expected by chance K is zero. When there is total agreement, K is one. An excel spreadsheet based calculation was built to

compute kappa statistics to compare responses from different versions of Survey Instrument 2; this sheet contained a matrix with 2 versions on columns and DQ factors listed in rows and the corresponding cells containing values of responses. This calculation sheet was replicated for each decision category.

An excel spreadsheet was set up to plot the calibration inputs from 2 sets of inputs i.e.

- Calibration inputs obtained by survey instrument that carried the initial weightage from the pre-study set up activities
- Calibration inputs obtained by survey instrument that DID NOT carry the initial weightage from the pre-study set up activities

These 2 pairs were compared for different decision categories to identify number of agreements and no agreements. This comparison (observed level of agreement) was used to compute Kappa statistic. An example of such comparison, for "Product Decision category" is given in the below table:

TABLE XX. KAPPA STATISTICS STUDY SET UP

Calibration Inputs (Set 1 above)	***Calibration Inputs (Set 2 above)***	***Observed Agreement***
H	H	Y
M	M	Y
M	H	N
H	M	N
M	M	Y

Kappa statistics revealed high degree of agreement between the different calibration inputs. For example, in the above case of products decision category, Kappa statistic produced a result of 0.8889, which maps to "Very good agreement" as listed in the previous paragraph. It is observed that there is no difference between the 2 versions of Survey 2 and thus calibration inputs received from both sets of survey administration approaches were considered valid for use in rest of the stages of the study.

Chi-Squared Test

To validate the results of the study (DQ score computed by the framework) and the levels assessed by independent users, Pearson's chi-squared test was designed as part of the data analysis phase of the study. This computation involved comparison of results from Survey 3 B (refer TABLE XIX.) that served as the gold standard and DQ score to determine the validity of the framework. In cases where the DQ score from the framework was within the independent assessment values, the counter for "Match" was incremented by 1 and in other cases, the counter for "Mismatch" was incremented by 1. Thus, the total of matches and mismatches were arrived at by each decision category. This aggregate matches and mismatches were used to conduct the Chi-square test which set up to compare a random baseline with these actual result aggregates from study at degree of confidence, df=1. Chi-square test is an important statistical test that allows us to test for deviations of observed frequencies from expected frequencies. As such, this test can be effectively conducted as part of this research work to compare results of DQ score as computed by the framework (based on weightage and confidence factor) and as assessed by independent users, the former being the observed values and the later being expected values. An excel spreadsheet based instrument was developed to capture the results from the study and compute chi-square test.

DIAGRAM XIX. STUDY RESULTS DATA ANALYSIS - CHI-SQUARE TEST

Enter your Null Hypothesis Below
There is no relationship between DQ and Business Outcomes
Enter your Alternative Hypothesis Below
There is a relationship between DQ and Business Outcomes
Result
There is a relationship between DQ and Business Outcomes

Yates Correction	0.5

#Rows	2
#Cols	2
df	1

Test Statistic

χ^2	42.62224692
p-value	0.0000000001
α	0.05
Result	Reject Null

Data Table- :: Overall results

	Results (Actual)	Baseline (Expected)	
Match	136	60	
Mismatch	164	240	

Friedman Tests

It was critical to detect any systemic bias in the responses received from the subject matter experts, either with respect to weightage or confidence levels. A series of Friedman tests were conducted to determine whether any bias existed in the responses from the chosen subjects. The Friedman test is a test for comparing three or more related samples and which makes no assumptions about the underlying distribution of the data. The data is set out in a table comprising *n* rows by *k* columns. The data is then ranked across the rows and the mean rank for each column is compared.

This test involves the computation given below.

$$M = \frac{12}{nk(k+1)} \sum_{Rj}^{2} - 3n(k+1)$$

Where:

k = number of columns (often called "treatments")

n = number of rows (often called "blocks")

Rj = sum of the ranks in column *j*.

Friedman test is defined as a non-parametric test (distribution-free) used to compare observations repeated on the same subjects. It was initially used to detect differences in treatments across multiple test attempts. The Friedman test is a rank-based, nonparametric test for several related samples where "related samples" may arise from a variety of research settings, such as the homogenous subjects that were part of the study in the current research work. Friedman's test is a test for treatment differences for a randomized complete block (RCB) design, where RCB design uses blocks of participants **who are matched closely on some relevant characteristic** (Salkind, 2007). - In the current research work the relevant characteristics are use of DSS for decision making, similar experience in retail banking industry, assessment involving same decision categories and use of same set of questions for survey in the study.

Friedman test allows for the analysis of repeated-measures data if participants are assessed on two or more occasions or conditions or to matched-subjects data if participants are matched in pairs, triplets, or in some greater number. Friedman test is applicable when the following conditions / assumptions are met:

- Each subject does all of the experimental conditions with more than two related samples on ordinal data
- All observations are mutually independent
- The rows are mutually independent. The results in one block (row) do not affect the results within other blocks.

- Data can be meaningfully ranked

In a recently published work, Sheldon et. al., 2006 reviewed the use and interpretation of the Friedman two-way analysis of variance by ranks test for ordinal-level data in repeated measurement designs and concluded that when the measurements are ordinal-scaled, such as some ratings of functional status and muscle strength, statistical significance may be determined by the Friedman test. In the cited work, Sheldon et. al., illustrated the use of the Friedman test with data from 27 subjects whose performance on a lifting task was rated by use of an ordinal scale.

In the current research work, the study involved obtaining independent validation of DQ scores from a different set of users. This set of independent users was given the same questions using the same scale. Moreover, the subjects chosen shared similar demographics and profile characteristics (e.g. experience, subject of specialization), besides the area of assessment being common i.e. use of DSS for decision making and same set of decision categories. Thus all the conditions and assumptions applicable for the test listed above are met in the current study. Therefore, based on the above discussions on the concepts of Friedman test and the principles referred in the work of Sheldon et. al., 2006, application of Friedman statistics to measure existence of any bias in the confidence level of DQ as perceived by them, based on their experience in using the DSS is considered appropriate. The above test was carried out to derive the measure described in the previous paragraph (computation involved in Friedman test) with DQ factors as rows and individual respondents in columns and the respective response in the cells; these responses were ranked to compute the above measure. This test was repeated separately for each of the decision categories. This test was performed separately for responses received for confidence factors (Survey 3A) and DQ Assessment (Survey 3B).

Chapter 5 – Results and analysis

Introduction

This chapter presents the details of the survey results, introduces DQ score computed using the framework for different decision categories, analysis of survey results and suggested areas for future research work based on this work. The first section presents results observed during the course of setting up of the study i.e. identification of DQ factors, decision matrix, and calibration results. The next section provides insights to the survey results i.e. calibration findings, DQ score for each decision category, confidence levels observed by different respondents. The next two sections describe overall analysis of results and recommendations for further work.

Results of empirical study

Many interesting results were observed in the course of the work.

1. This work resulted in finalization of 23 DQ factors as relevant for study of DQ and its relationship with business outcomes in Decision Support Systems. These factors, along with their definitions, are listed in the table below.

TABLE XXI. DQ FACTORS FOR DSS AND THEIR DEFINITIONS

S No	***Information Quality Attribute***	***Definition***
1	Believability	The extent to which data are accepted or regarded as true, real and credible
2	Concise representation	The extent to which data are compactly represented without being overwhelming (i.e. brief in presentation, yet complete and to the point)
3	Interpretability	The extent to which data are in appropriate language and units and data definitions are clear

S No	*Information Quality Attribute*	*Definition*
4	Reputation	Knowledge and awareness about the sources from which data is gathered and perception of overall trustworthiness of information available in the Decision Support System.
5	Understandability	The extent to which data are clear without ambiguity and easily comprehended
6	Value-added	The extent to which data are beneficial and provide advantages from their use
7	Granularity	Level of detail (fineness) of data provided for decision making process. Greater the granularity, deeper the level of detail (fineness of data)
8	Relevancy – Measures	Measurements or metrics or facts associated with a business function e.g. # of loans, outstanding amount, interest income etc.
9	Relevancy – Dimensions	Context or different perspectives for understanding the above facts i.e. characteristics such as who, what, where, when, how of a measure (subject). E.g. Housing Loan Amount (measure) by region, branch, agent, loan slab, borrower age, borrower occupation etc. (dimensions)
10	Aggregation	Summary or pre-computed measures that are used to enhance query performance
11	Completeness	The extent to which data are of sufficient breadth, depth and scope for the task at hand
12	Customer Support	Amount and usefulness of support
13	Documentation	Information on data attributes, their meaning and tips on using information
14	Objectivity	Extent to which data is free from bias or manipulation to direct analysis and decision making to pre-concluded results.
15	Reliability	The extent to which data that are being provided by the

S No	*Information Quality Attribute*	*Definition*
		Decision Support System, considered as trustworthy for decision making
16	Security	Extent to which access to information is restricted appropriately to maintain its security
17	Accuracy	The extent to which data are correct, reliable and free of error
18	Availability	Extent to which information is physically accessible
19	Consistency	Extent to which data (source data definition and data capture process) is uniform across time periods, person seeking the information.
20	Latency	The time between initiating a request in the computer System and receiving the information.
21	Response Time	Time until complete response reaches the user (turnaround time)
22	Timeliness	The extent to which the age of the data is appropriate for the task at hand
23	Verifiability	Degree and ease with which the information can be checked for correctness

2. The exercise of logically grouping the decisions likely to be taken from DSS in retail banking resulted in finalization of 6 decision categories viz., Credit Decisions, Business Promotion Decisions, Product Decisions, Tactical Decisions, Relationship Decisions and Regulatory Decisions.

3. Data Quality factors that do not map to any of the Decision Categories may be dropped from further analysis i.e. Price and Latency

4. DQ score computed using the framework (weightage determined initially by the author) are listed in the below table:

TABLE XXII. SETUP DQ SCORES

S No	*Decision Category*	*DQ Score*
1	Credit Decisions	73 %
2	Business Promotion Decisions	67 %
3	Product Decisions	54 %
4	Tactical Decisions	82 %
5	Relationship Decisions	58 %
6	Regulatory Decisions	84 %

5. DQ measurement introduced through this work meets the normative requirements (for DQ metrics) from published literature (Heinrich et. al., 2009). A list of DQ metrics requirements and how the DQ measurement introduced through this work meets those requirements are given in TABLE XXIII.

TABLE XXIII. DQ MEASUREMENT REQUIREMENTS

S No	*Requirements*	*How met by the proposed DQ measurement*
R 1	Normalization: An adequate normalization is necessary to assure that the values of the metric are comparable	This framework can be implmented to measure DQ in an organization at different time periods and thus the values of the measurement are comparable
R 2	Interval Scale: To support both the monitoring of the DQ level over time and the economic evaluation of measures the metrics should be interval scaled.	The framework doesn't cast any constraints related to data collection on a continuous basis; the implementation of the framework can be planned with better time intervals (during which DQ improvement measures can be implmeneted)
R 3	Interpretability: The metrics should be "easy to interpret by business users" and the values of the DQ metrics have to be comprehensible	DQ measurement proposed is a simple % and thus us easy to imterpret

S No	*Requirements*	*How met by the proposed DQ measurement*
R 4	Aggregation: the metrics must allow aggregation of the quantified values on a given level to the next higher level	This requirement remains to be addressed
R 5	Adaptivity: To quantify DQ in a goal-oriented way, the metrics need to be adaptable to the context of a particular application	The entire framework is context sensitive and provides for selection of DQ factor appropriate for the specific business / function for which the DQ measurement is taken up
R 6	Feasibility: To ensure practicality, the metrics should be based on input parameters that are determinable. When defining metrics, methods to determine the input parameters shall be defined	Inpuits for the DQ measurement are practically simple to capture (weigtage, users confidence and assessment) and can be obtained within the organization

Survey Results

This section covers the results from data gathered through administration of the surveys to validate the framework. Results presented in this section include DQ score computed based on calibration and confidence level inputs from the survey respondents, Chi Square test results and detailed comparison of individual DQ scores with assessment from users of DSS in the Bank.

DQ scores generated using the framework (identified as DQ1 to DQ 60) yielded the results presented in the tables in this Section. These tables present the survey results in the following structure:

- As discussed in Chapter 4 the calibration exercise was conducted through 2 sets of banking experts, each of them providing different weightage values for different decision category-DQ factor combinations.

- This resulted in two different versions of the calibrated model. The DQ score expression (refer chapter 3) was computed for both of these resulting calibrated models.
- Column 2 (Calibration) refers the DQ score using inputs from calibration set 1 or calibration set 2.
- Column 3 (Survey 3A respondent), as the title of the column indicates, the target subject that provided the confidence level for each the DQ factors for selected decision category (refer chapter 4)
- Column 4 (DQ score) refers to the DQ score generated from the framework, for the combination of inputs from calibration and confidence levels.

TABLE XXIV. DQ SCORE COMPUTED USING THE FRAMEWORK FOR CREDIT DECISION CATEGORY

S No	***Calibration***	***DQ Score ID***	***DQ Score***
1	Expert 1	DQ 1	60 %
2	Expert 1	DQ 2	55 %
3	Expert 1	DQ 3	66 %
4	Expert 1	DQ 4	62 %
5	Expert 1	DQ 5	75 %
6	Expert 2	DQ 6	59 %
7	Expert 2	DQ 7	55 %
8	Expert 2	DQ 8	65 %
9	Expert 2	DQ 9	61 %
10	Expert 2	DQ 10	74 %

TABLE XXV. DQ SCORE COMPUTED USING THE FRAMEWORK FOR BUSINESS PROMOTION DECISION CATEGORY

S No	***Calibration***	***DQ Score ID***	***DQ Score***
1	Expert 1	DQ 11	63 %
2	Expert 1	DQ 12	58 %

S No	*Calibration*	*DQ Score ID*	*DQ Score*
3	Expert 1	DQ 13	66 %
4	Expert 1	DQ 14	64 %
5	Expert 1	DQ 15	73 %
6	Expert 2	DQ 16	53 %
7	Expert 2	DQ 17	49 %
8	Expert 2	DQ 18	56 %
9	Expert 2	DQ 19	54 %
10	Expert 2	DQ 20	62 %

TABLE XXVI. DQ SCORE COMPUTED USING THE FRAMEWORK FOR PRODUCT DECISION CATEGORY

S No	*Calibration*	*DQ Score ID*	*DQ Score*
1	Expert 1	DQ 21	62 %
2	Expert 1	DQ 22	58 %
3	Expert 1	DQ 23	67 %
4	Expert 1	DQ 24	64 %
5	Expert 1	DQ 25	74 %
6	Expert 2	DQ 26	62 %
7	Expert 2	DQ 27	57 %
8	Expert 2	DQ 28	66 %
9	Expert 2	DQ 29	64 %
10	Expert 2	DQ 30	76 %

TABLE XXVII. DQ SCORE COMPUTED USING THE FRAMEWORK FOR TACTICAL DECISION CATEGORY

S No	*Calibration*	*DQ Score ID*	*DQ Score*
1	Expert 1	DQ 31	68 %
2	Expert 1	DQ 32	65 %

S No	*Calibration*	*DQ Score ID*	*DQ Score*
3	Expert 1	DQ 33	77 %
4	Expert 1	DQ 34	71 %
5	Expert 1	DQ 35	85 %
6	Expert 2	DQ 36	68 %
7	Expert 2	DQ 37	65 %
8	Expert 2	DQ 38	77 %
9	Expert 2	DQ 39	70 %
10	Expert 2	DQ 40	89 %

TABLE XXVIII. DQ SCORE COMPUTED USING THE FRAMEWORK FOR RELATIONSHIP DECISION CATEGORY

S No	*Calibration*	*DQ Score ID*	*DQ Score*
1	Expert 1	DQ 41	60 %
2	Expert 1	DQ 42	57 %
3	Expert 1	DQ 43	67 %
4	Expert 1	DQ 44	62 %
5	Expert 1	DQ 45	77 %
6	Expert 2	DQ 46	58 %
7	Expert 2	DQ 47	56 %
8	Expert 2	DQ 48	64 %
9	Expert 2	DQ 49	61 %
10	Expert 2	DQ 50	76 %

TABLE XXIX. DQ SCORE COMPUTED USING THE FRAMEWORK FOR REGULATORY DECISION CATEGORY

S No	*Calibration*	*DQ Score ID*	*DQ Score*
1	Expert 1	DQ 51	69 %
2	Expert 1	DQ 52	66 %

S No	*Calibration*	*DQ Score ID*	*DQ Score*
3	Expert 1	DQ 53	77 %
4	Expert 1	DQ 54	72 %
5	Expert 1	DQ 55	85 %
6	Expert 2	DQ 56	68 %
7	Expert 2	DQ 57	66 %
8	Expert 2	DQ 58	77 %
9	Expert 2	DQ 59	71 %
10	Expert 2	DQ 60	87 %

A comparison of DQ score as per the framework and as assessed by the users in the Bank was done; in cases where the DQ score was within the range as assessed by the user the count of "Correct Assignment" was incremented; else, the count of "Incorrect Assignment" was incremented. A summary of results from this exercise is as below:

TABLE XXX. SUMMARY OF SURVEY RESULTS

S No	*Decision Category*	*Correct Assignment*	*Incorrect Assignment*
1	Business Promotion	14	36
2	Credit Decisions	15	35
3	Product Design	27	23
4	Regulatory	16	34
5	Relationship	24	26
6	Tactical	40	10
	Overall	136	164

A series of chi-square tests were conducted to compare a random baseline with the actual results from the study. These tests were carried out at a degree of confidence, *df*=1 and the results of this test are presented in this Section.

TABLE XXXI. CHI-SQUARE TEST RESULTS

Hypothesis Number	***Decision Category***	***p-value***	***Result***	***Descriptive results***
H1	Main research problem	0.0000000001	Reject Null Hypothesis	There exists a direct relationship between DQ and Business Outcome
H2	Business Promotion	0.4824	Accept Null Hypothesis	For business promotion decisions, there is no relationship between DQ and Business Outcomes
H3	Credit Decisions	0.35561	Accept Null Hypothesis	For Credit decisions, there is no relationship between DQ and Business Outcomes
H4	Product Design	0.0009	Reject Null Hypothesis	For product decisions, there is a relationship between DQ and Business Outcomes
H5	Regulatory	0.2543	Accept Null Hypothesis	For regulatory decisions, there is no relationship between DQ and Business Outcomes

Hypothesis Number	***Decision Category***	***p-value***	***Result***	***Descriptive results***
H6	Relationship	0.0061	Reject Null Hypothesis	For relationship decisions, there exists a direct relationship between DQ and Business Outcomes
H7	Tactical	0.00000001	Reject Null Hypothesis	For tactical decisions, there exists a direct relationship between DQ and Business Outcomes

Analysis of results

Key observations of results from data points generated (DQ scores computed using the model described in Chapter 3) from the above study are as below:

1. Chi-square test conducted on the entire set of DQ scores generated from the model using the framework introduced in this research yielded a p-value of 0.0000000001. This result is statistically significant to support the hypothesis associated with the main research problem. The results establish that a relationship exists between DQ and quality of business outcome.
2. The results also confirm that DQ can be measured based on factors associated with business outcome.
3. Results of Kappa statistics test (TABLE XXXII.) revealed high degree of agreement between the calibration inputs and were thus valid for use in rest of the stages of the study. Results from the Kappa statistical analysis is presented below.

TABLE XXXII. KAPPA STATISTICAL ANALYSIS

S No	***Decision Category***	***Kappa Statistics Value***
1	Business Promotion	0.6111
2	Credit Decisions	0.7222
3	Product Design	0.8889
4	Regulatory	0.8329
5	Relationship	0.8323
6	Tactical	0.8333

- Results from Chi Square test conducted (to compare a random baseline with the actual results from study) **yielded a *p* value of 0.0000000001** (at degree of confidence, *df*=1), which is **statistically very significant** confirming points 1 and 2 listed above.
- Almost half of the results show that scores measured by the framework match with the assessment level expressed by the data users (45 %); this metric was higher at 49 %, considering the instances where the result from the framework and assessment level expressed by users varied marginally +/- 2 %.
- In selected decision categories, this agreement level was very high (e.g. in tactical decisions the match was 80 % or in product design decision category the same was 54 %). Above Chi-Square test in these decision categories showed a *p* value of 0.00000001 and 0.0009 respectively.
- Results of Friedman tests (referred in Chapter 4, page no. 120), revealed that there was no bias in responses received either for confidence factors (Survey 3A) or for DQ assessment (Survey 3B). The value of *M* as per this test for Survey 3B yielded a value of 3.266666667 which is much lower than the critical value (at $\alpha = 5\%$) 9.067. Similarly the value of *M* as per this test for Survey 3A also yielded much lower values as compared to the critical value (at $\alpha = 5\%$) for each of the decision categories. (e.g. Business Decisions category: -111.825 against 9.400 or Product Decisions category -124.022 as against 9.422 and so on).

In addition to the chi-square tests discussed above, 2 additional tests were carried out. These additional tests also involved chi-square; for the purpose of these tests, the study results were organized into 2 sets i.e. DQ Scores generated by the model as computed using the calibration inputs provided by each calibration expert. These actual results were compared with a random baseline. These tests were carried out at a degree of confidence, *df*=1 and the results of this test are presented below. The objective of this additional test is to verify if the results from the study using different calibration inputs yield different results.

TABLE XXXIII. ADDITIONAL TEST RESULTS

Hypothesis Number	***Data Set***	***p-value***	***Result***	***Descriptive results***
H8	DQ Score generated by the model based on callibration inputs by Expert 1	0.0000010244	Reject Null Hypothesis	As per the scoring model set up in Calibration Cycle 1 , there exists a relationship between DQ and Business Outcomes
H9	DQ Score generated by the model based on callibration inputs by Expert 2	0.0000244425	Reject Null Hypothesis	As per the scoring model set up in Calibration Cycle 2 , there exists a relationship between DQ and Business Outcomes

It can be seen from the above Table that results from these additional tests yielded a *p* value of 0.0000010244 and 0.0000244425 (at degree of confidence, *df*=1), which is **statistically very significant supporting research hypotheses**.

Chapter 6 – Findings and recommendations

Introduction

This chapter discusses the findings from the study, interpretation and their relevance to DQ research. This chapter also deals with uniqueness of this research work. The last section of this chapter covers the recommended use and application of this research work.

Research findings

This study has investigated the relationship between the DQ and business outcomes, in DSS. A literature review revealed the need to explore this area of study and establish a model to support measurement of this relationship. Many open research challenges from published literature have been addressed through the current research work. For example, Keeton et. al., 2010 had outlined a research challenge related to DQ metrics when they state that ***the first research challenge is in providing lightweight, scalable mechanisms for determining DQ metrics,*** without which users of Systems cannot make decisions with high confidence. The context based DQ measurement framework introduced and experimented in this current work addresses the above challenge.

Summary of results from this research and findings there from are as below:

- The *p* value for the chi-squared test conducted is **significantly low** at **0.0000000001**. Hence, the Null Hypothesis referred in Chapter 3 (page no.85) may be rejected and the alternate Hypothesis presented in Chapter 3 (page no. 85) may be accepted.
- The results establish that a relationship exists between DQ factors and quality of business outcome in Banking domain.
- The results also confirm that DQ can be measured based on factors associated with business outcome.

The research for this study has demonstrated that the relationship between the DQ and business outcomes is systematically measurable, through the framework introduced by this

research work. Empirical study conducted and analysis of results of the study establish that this work marks a significant extension to existing works found in literature in the study of DQ assessment / measurement, yet, different in many ways explained later in this chapter. Additionally, this research has set forth a framework that may be considered useful in positioning and describing this and other research on the topic of study of DQ in Decision Support Systems with focus on business outcome, within the broader context of the body of literature.

Uniqueness of this work

This work and the DQ measurement framework introduced through this work are different from existing work in the area of DQ research in the following ways (refer DIAGRAM XX.):

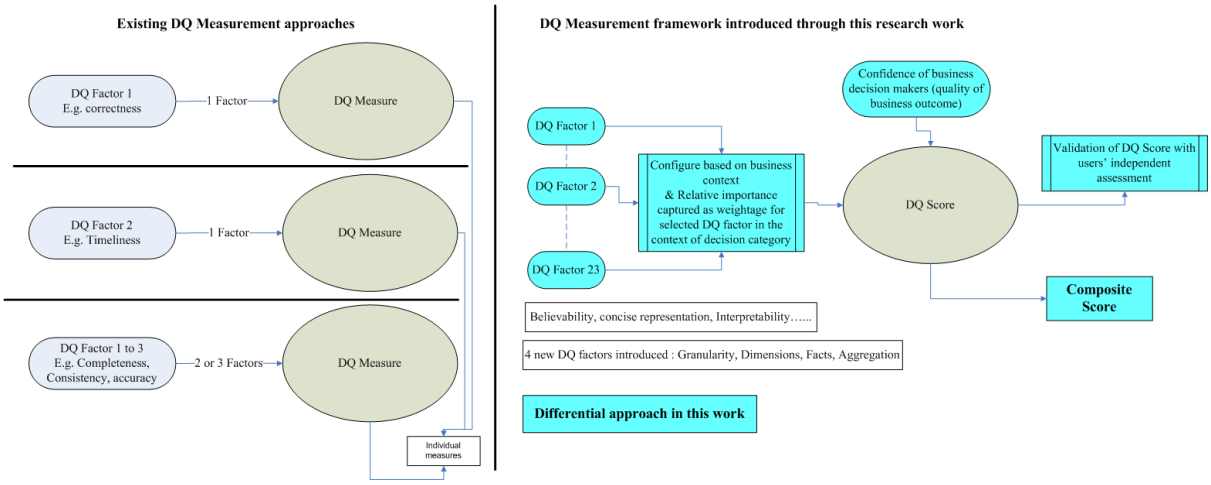
TABLE XXXIV. UNIQUENESS OF THIS WORK

S No	***Area***	***Existing works***	***Current research work***
1	DQ factors considered	Measure either one or few DQ factors in isolation (refer TABLE VI. , Chapter 2, page no. 56)	Considers a set 23 DQ factors (or a sub set from them) for measurement of DQ Score
2	DQ factors for decision support	Not covered	Included 4 new DQ factors relevant to DSS (Batini et. al., 2009) (refer discussions following TABLE XI. , Chapter 3, page no. 72)
3	Business context in DQ measurement	Not considered	Steps recommended for implementing this new measurement framework (refer DIAGRAM XV. , Chapter 3, page no. 84) allows the DQ measurement exercise to be context sensitive to the specific industry (e.g. Banking) and further for specific function

S No	*Area*	*Existing works*	*Current research work*
			within the selected industry (e.g. retail banking). Moreover, the framework provides for finer context based on the type of decision to be derived from underlying data, by introducing the concept of decision categories. This has been addressed through an association variable *a(i,d)* in DQ computation.
4	Relative importance of DQ factors	In the past works where multiple DQ factors are used (Lima et. a., 2006) all DQ factors are treated alike without consideration for the context	This work recognizes that based on specific types of decisions (decision categories), selected DQ factor impacts DQ differently. Therefore, this work considers weightage of DQ factor as an important aspect in DQ measurement. Accordingly, a new variable *w(i)* has been introduced in the mathematical model for computation of DQ score.
5	Link to business outcome	Do not consider	This work recognizes the need to link the quality of business

S No	*Area*	*Existing works*	*Current research work*
		quality of business decisions / outcomes in the DQ assessment / measurement process	outcomes / business decisions derived from use of data and hence confidence level of the users in the quality of data is included in the mathematical model, by introducing a new variable *c(i)*
6	Validation of DQ scores	Do not provide mechanism for validating the scores	This work recognizes the need for closed loop approach (Heinrich et. al., 2007) and accordingly introduces a step within the framework to seek independent assessment of DQ from consumers of data for decision making. The responses from this independent assessment phase serves as the gold standard, and DQ score generated by the framework are compared to the gold standard to determine the validity of the framework

DIAGRAM XX. UNIQUENESS OF THIS RESEARCH WORK



Recommended use of findings

This research work and its findings are expected to be useful in different ways for researchers, industry experts and practitioners of IT Systems. For researchers, this work serves as a reference in different areas – grouping of literature survey as detailed in Chapter 2, unified framework for DQ measurement (presented in chapter 3), practical approaches to experiments for data collection from the Industry (detailed in chapter 4) and finally explore future areas of research recommended at the end of this chapter. This research work is expected to add special attention to the domain/ sub-domain specific issues in dealing with quality of data for decision making. Industry experts may explore implementing this framework by selecting appropriate DQ factors as it applies to their entity (Industry / Firm / Division / Function / Sub-function etc.), identify and apply relevant weightage for the above selection and arrive at baseline DQ scores.

Rodríguez et. al., 2010 observe that assessing the quality of an IT System has been one of the major challenges for researchers as well as for the practitioners (organizations) and that there is the need for a multidimensional instrument capable of measuring the quality of an information system. Such an instrument is expected to provide valid information that helps

an organization to take decisions in order to improve and assure the quality of its information systems. Findings from the current research work are expected to be useful in addressing such expectations. IT Practitioners (DW Applications / DSS) can look forward to using an objective measure of DQ and use the framework introduced and validated through this research work for measuring DQ over a period of time as data is incrementally loaded in these systems. For example, DQ scores can be measured every quarter to monitor the movement of DQ over time (based on data from different source systems that get integrated into the System and/or the extent of use of the system for decision making for business purposes). This in turn can help IT practitioners to plan appropriate DQ improvement plans in their organizations.

In their recently published work Shankaranarayanan et. al., 2012, while examining the DQ (metadata) in DSS, have inferred that DQ positively impacts decision performance. The said study was carried out through a study covering ~60 students from a business school. In this study, the authors adopted a method of simple additive weighting or the weighted linear combination (which is often used in spatial multi-attribute decision making and in quality function deployment that considers multiple product criteria in operations management). One of the key limitations of the said study was the population size and the authors have recommended that the study be extended to cover larger sample size in a real-world setting. This research work has addressed this limitation by carrying out detailed study with a larger population, using the model in real-world setting with larger retail banks, to generate DQ score observations from the industry practitioners (and users of DQ/DSS).

Chapter 7 – Discussions

Introduction

This chapter revisits the importance of DQ in decision making process using DSS, in light of more recent published work. The objective of this discussion is to set the tone for future research work in the context of more recent industry developments. These discussions cover DQ in industry in general, but, with special reference to Banking industry, since the empirical study as part of this research work was carried out in the Banking industry.

DQ – an area of enduring research

The importance of DQ for better decision making and this better business outcomes can never be over emphasized. Pfeffer et. al., 2006 observe that a company's success hinges on the quality of the decisions its executive teams make and that towards making these decisions in any function a Company needs data of good quality to select the best course of action. Pfeffer et. al., 2006 further observe that lack of good quality data and not using such data in the decision making process leads to poor-quality decisions that waste time and money (at best) and risk a company's future (at worst). English, 2011 in a recently published book quantifies the cost of poor DQ for organizations to be in the range of 20-35% of operating revenue.

There has been a growing interest among researchers and industry practitioners in DQ, DQ assessment and improvement. However, the basic problem of DQ assessment and improvement still remains largely open. For example, Massachusetts Institute of Technology (MIT) has instituted Total Data Quality Management (TDQM) research effort which is grown from industry needs for high quality data. The objective of this program is to establish a solid theoretical foundation and devise practical methods for business and industry to improve DQ. The research scope of this program covers 3 areas i.e. 1) definition of DQ that addresses issues of data quality definition, measurement, and derivations; 2) Analysis of DQ Impact on Business that addresses the value chain relationship between DQ and business outcome and 3) improvement of DQ that addresses various methods for improving DQ.

Thus research interest in DQ remains an enduring subject, which is reiterated even by very recent literature. For example, Fehrenbacher et al., 2012 state that the importance of DQ is ever increasing and that research in this field focuses mainly on **2 aspects i.e. criteria and assessment**. This recent work observes that while researchers have developed a number of frameworks, criteria lists and approaches for assessing and measuring DQ, still research in this discipline indicates that assessing DQ remains to be challenging. This work argues that although DQ is subjective, most of the existing frameworks and assessment methodologies do not often consider the context in which the assessment is performed. Through empirical data research this cited work suggests that the perceived importance of DQ criteria has changed over the last decade.

DQ in relation to Banking

Banks worldwide use DW / DSS solutions for a variety of scenarios in the decision making process e.g. performance measurement, profitability analysis, risk management, compliance requirements, regulatory reporting and customer relationship management. Deployment of a data warehouse and business intelligence capability is the next logical step for Indian banks, especially those in the public sector, in their strategic use of information technology (Indian Bankers' Association, 2007).

"*Bad data is like dirty air. People have to use it although they know it has negative side effects*". Gartner research has cited an example of DQ impacting business outcome is highlighted in the case where Central banks were slow to recognize the scope of the 2007 economic crisis because they could not obtain reliable data to calculate the total risk exposure present in the global financial system (Bugajski, 2010).

In a recent global survey conducted by McKinsey (Roberts, 2011), 28 % of the 787 respondents have cited poor DQ as a critical barrier to increasing the use of data and advanced analytics for decision making.

On the same lines, the Reserve Bank of India, 2011 summarizes the need to improve DQ practices in the Banking sector. Few of the pertinent recommendations are listed below: (since this is a report submitted by a high level committee of the country's Apex Bank, after

several rounds of deliberations, some of the observations are reproduced verbatim (in italics), so as not to dilute either the focus or significance of the recommendations.

- Timeliness, availability and completeness (coverage) of data continue to be critical DQ factors that need focus and attention.
- "*Data Quality: Presently, the data reported to Reserve Bank is not necessarily forwarded from the centralised system or the MIS servers of banks. Manual intervention often leads to incorrect reporting. This raises doubts on the genuineness of data management. Policy decisions based on such data may be erroneous.*" Therefore, DQ needs to be improved for use in DSS and use of DSS / DW for decision making needs to be encouraged (at present there does not exist a strong linkage in the use and sharing of information for decision making process).

The present research work is aligned with this state of the industry and research on the need for making advances in the study of DQ. To begin with, this research accomplished an exhaustive study of literature related to DQ published till 2011 and identified the gaps (page No. 54) considered crucial for further research and listed the objectives for such study.

Further, this research work helped in consolidating published literature related to existing DQ assessment frameworks (TABLE IX.). This work covered the 1st aspect of MIT's TDQM research agenda i.e. DQ definition, in so far as the work identified new DQ factors as applicable for DQ in DSS / DW (page no. 69). Existing literature has time and again emphasized the need for a comprehensive DQ assessment / measurement framework as an important direction in DQ research. In line with these recommendations, this work introduced a comprehensive DQ assessment framework (DIAGRAM XV.). As often emphasized by published literature (Fehrenbacher et al., 2012), this work has introduced appropriate factors for considering context in the DQ assessment process. The empirical study was designed to involve the users of DQ to capture their confidence levels and assessment inputs. In summary, contributions of this research (DQ definition and assessment) and its findings are expected to be of significant value for researchers, academicians and practitioners of DQ.

DQ related to unstructured data

Unstructured data may be referred as information contained in desperate systems, but, without specific semantics. Typical examples of unstructured data are free-format text captured by customer feedback systems (mostly though web interface) or notes / observations /annotations on business documents captured during the course of workflow in an approval or authorization process.

The field of integrating unstructured data into DSS is still evolving. Generally, three steps are involved for making such information available for decision making:

- Data needs to be analyzed, identified as useful and categorized
- Unstructured data needs to be "conformed" to a structure
- In the process of such transformation DQ processes needs to be applied before data can be made available for decision-making.

Consolidating the above steps, leads to an effective integrated environment i.e. a platform where structured and unstructured data is integrated and presented through an integrated framework for decision support (Baars and Kemper, 2008).

A large majority of research contributions so far has been focused on DQ with respect to structured data (Batini et. al., 2009). Echoing the same views, Wang et. al., 2011 highlight the need for evolution of research in dealing with DQ for unstructured data. While there is an increasing demand for incorporating unstructured information in DSS, most of the current research on DSS has mainly focused on dealing with structured data and are inadequate to dealing with unstructured information (Wang et. al., 2011). It is observed that different technologies and techniques e.g., knowledge-based system or computational linguistics or artificial intelligence may be used for processing unstructured data and inferring useful knowledge for decision support (Wang et. al., 2011). A related area of emerging research involves using text mining as a means for deriving knowledge to support decision making. As per a recent published work, this field of decision support further extends to study of

ontology-guided information retrieval or domain-specific taxonomies for extracting knowledge from unstructured data (Bratus et. al., 2011).

Many of the DQ factors discussed in this work may be relevant for unstructured data e.g. reliability or timeliness or understandability or objectivity etc. However, existing literature doesn't cover (adequately) DQ factors applicable with specific reference to unstructured data.

The discussions and literature references lend a few view points:

- DQ for unstructured data remains less explored
- Different technologies that are still evolving e.g. text mining are involved in decision support using unstructured data.
- Different processes (such as analysis or extraction or integration) are associated with providing structure to unstructured data before they are presented in DSS for decision support
- Study of DQ applicable for unstructured data or decision support using unstructured data sources may, by themselves, be separate and significant areas of study / research.

Decision support Vs. decision making

The scope and focus of this research work has been on DQ in DSS; thus, decision making or decision process which is a different function by itself is not intended to be covered by this work. However, it may be relevant to briefly discuss the subject of decision making, particularly, multi-criteria decision making as the concepts involved in multi-criteria decision making are similar to the problem of multiple DQ factors in DSS.

Multi-Attribute Decision Making (MADM) is a branch of decision making that deals with decision problems under the presence of a number of decision criteria and is considered useful in many situations – economics, managerial, construction etc. (Zavadskas et. al., 2009). Multiple attribute decision-making problems are encountered under various situations

where a number of alternatives or actions or decisions need to be chosen based on a set of attributes. The main steps involved in multiple attributes decision-making are as follows:

- Establish attributes that relate to objectives or goals
- Generate alternatives i.e. developing alternative systems for attaining achieving the objectives / goals
- Evaluate alternatives in terms of attributes; this involves analysis of objectives / goals by using attributes, which have different dimensions, different weight or different directions of optimization (Petkus et. al., 2008) and (Ustinovichius et. al., 2007).
- Apply a normative multiple attributes analysis method
- Identify "optimal" or "preferred" alternative
- Optimization: If the solution is not identified, gather new information and go into the next iteration of multiple attributes selection.

Most of the MADM methods require that the attributes be assigned weights of importance. One of the most crucial steps in many decision making methods is the need to elicit qualitative information from decision makers, which is very often cannot be known in terms of absolute values. Therefore, many decision making methods attempt to determine the relative importance, or weight, of the alternatives in terms of each criterion involved in a given decision making problem. Such relative importance typically expressed qualitatively is translated to a quantitative value using a scale, which is either a linear scale or an exponential scale. This subject of decision making, by itself, forms a separate topic of research with several advances and open research issues and as such is not intended to be covered in detail in this thesis.

Chapter 8 – Conclusions and future work

This chapter discusses the conclusions that can be drawn from this research, identifies the limitations of the current work and provides recommendations for future extensions to this work and/or further research in DQ definition or DQ assessment / measurement. The objective of this discussion is to ideate on directions for continuous improvement and advances in the study of DQ in the context of industry developments and expectations from DQ researchers and practitioners.

Conclusions

This research work was undertaken to meet the 5 objectives listed in Chapter 2 (page no. 54). These objectives were based on the need to advance the study of DQ assessment to address the gaps that were identified from an exhaustive review of literature in this field. The key objective was to develop a DQ assessment / measurement framework that deals with DQ comprehensively and based on the context of such assessment / measurement. An associated objective was to identify the list of DQ factors pertinent to DSS. Other objective included appropriate involvement of users of data in the DQ assessment / measurement process.

This research work followed existing methodology to develop a DQ assessment framework, discussed in detail in Chapter 3. Alongside this framework, new DQ factors that were critical for DQ in DSS were introduced and a comprehensive list of DQ factors that were relevant for the study was drawn up. This work also introduced the concept of decision categories (page no. 102) to align the study of DQ with the context of the assessment / measurement process. This led to a new introduction of a new expression for measurement of DQ. An empirical study was conducted to test the framework (introduced through this research) and statistical tests were performed on the study results, to validate the research hypotheses.

As discussed earlier, the study results support the research hypotheses and are found to validate the framework (page no. 84) introduced through this work for DQ assessment and the equation (page no.85) proposed for DQ measurement. For example, the overall results of

the study conducted have shown a significantly low *p* value (chi-squared test) at **0.0000000001**. (Note: The overall results here mean the results of chi-squared test performed to validate DQ scores computed by the framework with the levels assessed by independent users that served as the gold standard, as detailed in page no. 119). Further, the results of chi-squared test conducted for each calibration cycle have resulted in a *p* value of 0.0000010244 and 0.0000244425, which are again significantly low. This seeks to prove the hypotheses set out in Chapter 4. In summary, the study findings lead to the conclusion that a direct association exists between business outcomes and DQ factors and that this relationship is systematically measurable. These results also point to the conclusion that the above association is a function of relative importance of the DQ factors (in the context of select decision category) and confidence of users in the quality of data used for decision making.

The results of the present research work (presented and analyzed in Chapter 5 and summarized in the preceding paragraphs) lead to the conclusion that **DQ is systematically measurable based on context in a comprehensive manner** i.e. by decision category with a complete set of applicable DQ factors. This finding meets objectives 1 and 2 of the research objectives set forth in Chapter 2 (page no. 54). The framework introduced in Chapter 3 – Research Methodology, page no. 84 and the DQ measurement expression introduced in page no.85, together with the results of the study (refer Analysis of results in page no. 133) support this conclusion, besides meeting the research objective 3. This research was focused on study of DQ in DSS. The methodology adopted (page no. 76) and approach followed (page no. 69) to achieve completeness with respect to DQ factors dimension, has resulted in expanding the DQ factors list by including 3 new DQ factors (viz., granularity, relevance of dimensions and relevance of measures) relevant to the study of DSS. This leads to the conclusion that **selection of DQ factors has to be considered based on the context and purpose of DQ measurement** (e.g. DQ in DSS for Banking). A critical step in the DQ assessment process introduced by this research work involves validation of the DQ scores independently by users of data (DIAGRAM XIV. Page no. 82). This validation process read with the study results help **conclude that users of data may be effectively involved in the DQ assessment / measurement process**. Further, research objectives 4 and 5 were met through this process of users' involvement in DQ assessment.

Limitations of current work

This study was based upon the assumption that the participant responses accurately reflected the knowledge and perceptions of those participants with respect to DQ and business and that measurement of those perceptions provided a reasonable representation of reality. The results of this study may not necessarily reflect the generalized views of broader selections of organizations or of organizations outside the industries represented in the sample.

Survey research is limited by the extent to which the responses accurately reflect the perspectives of the participants, and the extent to which those perspectives reflect the real-world situation under investigation. These limitations can be mitigated through rigorous attention to the design of the survey instrument and the extent of the limitation can be assessed by analyzing the construct validity of the instrument (Cooper and Schindler, 2003). The instrument used for this study was developed using accepted practices and the majority of the items used in the instrument had been validated previously (Slone, 2006). Further tests (kappa statistics) were conducted to assess the validity of the survey instruments used in subsequent phases of the study.

The population for this study was defined rather broadly; however, it was still limited to persons working in organizations that have implemented DSS and who use information regularly. The ability to generalize the results is limited to that population, and is further limited by the characteristics of those who actually participated.

Recommendations for future work

This framework is different from past work (of similar DQ measurement approaches) as it advances the conventional philosophy of "one size fits all" or "one definition of DQ captures it all". Echoing these thoughts, Weber et. al., 2009 suggest that companies should structure their own data governance models to meet their DQ management needs. In the said work, the authors introduce a model for DQ, which has however not been empirically tested; the authors suggest that the companies can use the proposed model to design a data governance configuration that fits their specific requirements, maximizing the positive contribution of DQ to their business objectives.

The current research work established the need for **2 broad characteristics in the study of DQ for DSS i.e. context and comprehensiveness**. The proposed framework is context sensitive in that it differentiates different business domains and within a selected domain it recognizes the specific sensitivities associated with different functions and their decision making requirements. Since the framework is context sensitive, it facilitates different business units or functional entities (e.g. Credit Department or Legal Department or Customer Relationship functions within Retail Banking) to focus on different sets of DQ attributes that are appropriate in their decision making. The proposed framework is comprehensive in that it defines a universal set of DQ factors relevant for the study and provides ability to measure DQ considering this wide range of DQ factors (and not limited to selected few factors).

The Hypothesis and Framework was evaluated in Retail Banking Industry; Sundararaman, 2012 recommends that future work may be carried out on similar lines to validate the same in other industries. Within Banking, DQ requirements for other business lines (Corporate Banking or Investment Banking) may be quite different with different decision categories and associated DQ factors. Based on the impact of decisions on outcomes of these business lines, weightage of the DQ factors may be different. Therefore, the author suggests that future work may be conducted using the framework in other business lines of Banking.

Similarly, DQ requirements for other industries (e.g. Telecom or Automobiles or Chemicals) may be quite different with different decision categories and associated DQ factors. Based on the impact of decisions on outcomes of these industries, weightage of the DQ factors may be different. Therefore, the author suggests that future work may be conducted using the framework in other industries.

In the current work, the author arrived at the decision categories by engaging with industry experts in Banking. The scope of this initial phase of the study may be expanded. The author suggests that empirical study may be conducted to cover initial steps of the work – set-up activities (refer DIAGRAM IX. and DIAGRAM XI. in Chapter 4) i.e. arrive at decision categories and DQ mapping matrix.

The framework may be experimented on a continuous basis; i.e. assess a baseline DQ Score and based on the weightage and confidence, select specific DQ factors around which DQ improvement initiatives can be implemented; post this implementation, DQ score can be computed using the framework with new set of weightage and confidence levels, once the improvement initiatives are implemented. Existing research (Madnick et. al., 2009; Ballou and Tayi, 1999; and Lyn, 2009) suggests that DQ measurements can be done periodically or continuously.

The model introduced through this work may be extended / refined further in the study of DQ factors applicable for unstructured data.

The method applied in this research work was empirical study. Researchers are encouraged to explore case study method to extend this work and address some of the limitations listed earlier.

References

[1] Alexander, J. E. and Tate, M. A. (1999), "*Web wisdom: How to evaluate and create information quality on the web*", L. Erlbaum Associates.

[2] Alkharboush, N and Li, Y. (2010), "A decision rule method for data quality assessment", proceedings of the *15th International Conference on Information Quality*.

[3] Alkharboush, N. and Li, Y. (2010), "A Decision Rule Method for Assessing the Completeness and Consistency of a Data Warehouse", *IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pp. 265- 268.

[4] AlMabhouh, A and Ahmad, A. (2010), "Identifying quality factors within data warehouse", *Second International Conference on Computer Research and Development*, pp.65-72.

[5] Amicis, Fabrizio De and Barone, D. (2006), "An Analytical framework to analyze dependencies among data quality dimensions", proceedings of the *11th International Conference on Information Quality (ICIQ)*, pp.369–383.

[6] Amicis, De F. and Batini, C. (2004), "A methodology for data quality assessment on financial data", *Studies in Communication Sciences*, Vol 4, pp.115-136.

[7] Asproth, V. (2007), "Visualisation of data quality in decision-support systems", *International Journal of Applied Systemic Studies*, Vol. 1, No.3, pp.280 – 289.

[8] Baars, H. and Kemper, H. (2008), "Management Support with Structured and Unstructured Data—An Integrated Business Intelligence Framework." *Information Systems Management*, Vol. 25, No. 2, pp. 132-148.

[9] Ballou,D. and Tayi, Giri Kumar. (1999), "Enhancing data quality in datawarehouse environments", *Communications of the ACM*, Vol. 42, Issue 1, pp.73-78.

[10] Batini,C. and Scannapieco,M. (2006), "*Data Quality. Concepts, Methodologies and Techniques*", 1st Edition. Springer.

[11] Batini, C., Barone, D., Mastrella, M., Maurino, A. and Ruffini, C. (2007), "A Framework and a Methodology for Data Quality Assessment and Monitoring", proceedings of the *12th International Conference on Information Quality*, Cambridge, pp. 333-346.

[12] Batini,C., Cappiello,C., Francalanci,C. and Maurino,A. (2009), "Methodologies for Data Quality Assessment and Improvement", *ACM Computing Surveys*, Vol. 41, No. 3, Article 16.

[13] Batini, C., Cabitza,F., Cappiello,C. and Francalanci, C. (2008), "A comprehensive data quality methodology for Web and structured data", *International Journal of Innovative Computing, Information and Control*, Vol. 1, No. 3, pp.205–218.

[14] Bratus, S., Rumshisky, A., Khrabrov, A., Magar, R. and Thompson, P. (2011). "Domain-specific entity extraction from noisy, unstructured data using ontology-guided search", *International Journal of Document Analysis and Recognition*, Vol. 14, No. 2, pp. 201-211.

[15] Bugajski, J. (2010), "Field Research Summary: Trials, Tribulations, and Testaments Along the Journey to Information Quality", Gartner Technical Professional Research. ID: ID:G00207415.

[16] Caballero, I., Calero, C., Piattini, M. and Verbo, E. (2008), "MMPro: A Methodology based on ISO/IEC 15939 to Draw up Data Quality Measurement Processes", proceedings of the *13th International Conference on Information Quality*.

[17] Cappiello, C., Francalanci,C. and Pernici, B. (2004), "A rule-based methodology to support information quality assessment and improvement", *Studies in Communication Sciences*, Vol 4, No.2, pp.137-154.

[18] Carletta, J. (1996), "Assessing agreement on classification tasks : : the kappa statistic", *Association for Computational Linguistics*, Vol. 22, No 2.

[19] Chung, W., Craig,F. and Wang, R. Y. (2005), "Redefining the Scope and Focus of Information Quality Work", Information Quality. New York: M.E.Sharpe, pp. 230-248.

[20] Cooper, D. R. and Schindler, P. S. (2003), "*Business research methods*", McGraw-Hill Irwin.

[21] Dedeke, A. (2000), "A conceptual framework for developing quality measures for information systems", proceedings of *5th International Conference on Information Quality*, pp.126–128.

[22] Dewan,R and Storey, V. (2008), "Guidelines for setting Organizational Policies for Data Quality", proceedings of the *41st Annual Hawaii International Conference on System Sciences*, USA, pp. 387- 395.

[23] Eckerson, W. (2002), "Data warehouse Institute Survey on Data Quality", proceedings of the *Seventh International Conference on Information Quality*, USA, 2002, Nov 8 – 10, pp.1-2.

[24] European Commission. (2007), "*Handbook of Data Quality Assessment method and tools*", Manfred Ehling and Thomas Körner (eds).

[25] English, L. (1999), "*Improving Data Warehouse and Business Information Quality*", Wiley & Sons.

[26] English, L. (2011), "*Information Quality Applied: Best Practices for Improving Business Information*", Processes and Systems. Wiley Publishing.

[27] Eppler. M. (2002), "Measuring information quality in the Web context: A survey of state-of-the-art instruments and an application methodology", proceedings of the *7th International Conference on Information Systems (ICIQ).*

[28] Eppler, M. and Muenzenmayer, P. (2002), "Measuring information quality in the web context: A survey of state-of-the-art instruments and an application methodology", proceedings of *7th International Conference on Information Quality*, pp.187–196.

[29] Even, A. and Shankaranarayanan, G. (2007), "Utility-Driven assessment of data quality", *The DATA BASE for Advances in Information Systems*. Vol.38, No. 2, pp.75–93.

[30] Even, A. and Kaiser, M. (2009), "A Framework for economics-driven Assessment of Data Quality decisions", proceedings of the *14th International Conference on Information Quality*.

[31] Falorsi, P., Pallara, S., Pavone, A., Alessandroni, A., Massella, E., and Scannapieco, M. (2003), "Improving the quality of toponymic data in the italian public administration", proceedings of the *ICDT Workshop on Data Quality in Cooperative Information Systems (DQCIS).*

[32] Fehrenbacher, D. and Helfert, M. (2008), "An empirical research on the evaluation of data quality dimensions", proceedings of the *13th International Conference on Information Quality*.

[33] Fehrenbacher, D and Helfert, M. (2012), "Contextual Factors Influencing Perceived Importance and Trade-offs of Information Quality", *Communications of the Association for Information Systems*, Vol. 30, Article 8. pp. 111-126.

[34] Frank, A. U. (2008), "Analysis of Dependence of Decision Quality on Data Quality", *Journal of Geographical Systems* Vol. 10(1), pp.71 - 88.

[35] Cong, G., Fan, W., Geerts, F., Jia, X. and Ma S. (2007), "Improving Data Quality: Consistency and Accuracy", Proceedings of the *33rd international conference on very large data bases* , pp. 315-326.

[36] Ge, Mouzhi; Helfert, M. and Jannach, D. (2011), "Information quality assessment: validating measurement dimensions and processes", proceedings of the *19th European Conference on Information Systems*.

[37] Ge,Mouzhi and Helfert, M. (2006), "A framework to assess decision quality using information quality dimensions", proceedings of the *11th International Conference on Information Quality (ICIQ)*, pp 455-466.

[38] Gibson, N. (2010), "Improving information products for System 2 Design Support", Ph. D. Thesis, University of Arkansas.

[39] Greene, J. C., Kreider, H., and Mayer, E. (2005), "Combining qualitative and quantitative methods in social inquiry", B. Somekh & C. Lewin (Eds.), *Research methods in the social sciences*, pp. 274-281.

[40] Gu Lin, Gu Jing and Dong Fang-fang. (2011), "Evaluation method of enterprise information quality based on QFD", *International Conference on Consumer Electronics*, pp 325-328.

[41] Gustafsson, P., Lindström, Åsa ., Jägerlind,C. and Tsoi, J. (2006), "A Framework for Assessing Data Quality – from a Business Perspective", *Software Engineering Research and Practice*, pp.1009-1015.

[42] Harold, L and Thenmozhi, M. (2008), "Information Quality and Banking Success: Evidence from the Indian Banking Industry", proceedings of the *13th International Conference on Information Quality*, pp. 146-158.

[43] Hasan, S., Padman, R and Duncan, George T. (2009), "On Data Quality and Risk in Guideline Based Clinical Decision Support", *Information Technology and Systems eJournal*, http://ssrn.com/abstract=1501762.

[44] Hasan, S and Padman, R. (2006), "Analyzing the Effect of Data Quality on the Accuracy of Clinical Decision Support Systems: A Computer Simulation Approach", proceeding of *AMIA Annual Symposium*, pp. 324–328.

[45] Heinrich,B., Kaiser,M. and Klier,M. (2007), "How to measure Data Quality? – A metrics based approach", 28th *International Conference on Information Systems*.

[46] Heinrich, B., Kaiser, M. and Klier, M. (2008), "Does the EU insurance mediation directive help to improve data quality? – A metric-based analysis", proceedings of the *16th European Conference on Information Systems (ECIS)*.

[47] Heinrich, B., Kaiser,M. and Kier,M. (2009), "A Procedure to Develop Metrics for Currency and its Application in CRM", *ACM Journal of Data and Information Quality*, Volume 1, pp. 1-28.

[48] Heinrich, B. and Klier, M. (2009), "A novel data quality metric for timeliness Considering supplemental data", proceedings of the *17th European Conference on Information Systems (ECIS)*, pp 2701-2713.

[49] Helfert, M. and Foley, O. (2009), "A Context Aware Information Quality Framework", Fourth *International Conference on Cooperation and Promotion of Information Resources in Science and Technology*, pp. 187-193.

[50] Helfert, M., Foley, O., Ge, Mouzhi and Cappiello, C. (2009), "Limitations of Weighted Sum Measures for Information Quality", proceedings of *Americas Conference on Information Systems*, Paper 277.

[51] Idris, N. Ahmad, K. (2011), "Managing Data Source quality for data warehouse in manufacturing services", *International Conference on Electrical Engineering and Informatics*, pp 1-6.

[52] Indushobha, N., Ballou, D. P., and Pazer,H. (1999), "The impact of data quality information on decision making: an exploratory analysis", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 11, Issue. 6, pp 853-864.

[53] ISO/IEC 25012. (2008), "Software Engineering -Software Product Quality Requirements and Evaluation (SQuaRE) -Data quality model", International Organization for Standardization.

[54] Jeusfeld, M.,Quix, C., and Jarke,M. (1998), "Design and analysis of quality information for datawarehouses", proceedings of the *17th International Conference on Conceptual Modeling*.

[55] Jung, W. (2004), "A Review of Research: An Investigation of the Impact of Data Quality on Decision Performance", proceedings of the *international symposium on Information and communication technologies*, pp.166-171.

[56] Kahn, B. K., Strong, D. M., and Wang, R. Y. (2002), "Information quality benchmarks: Product and service performance", *Communications of the ACM*, Vol. 45, Issue 4, pp. 184-192.

[57] Kaiser, M. (2010), "A conceptional approach to unify completeness, Consistency, and accuracy as quality dimensions of data values", *European and Mediterranean Conference on Information Systems*.

[58] Katerattanakul, P. & Siau, K. (1999), "Measuring information quality of web sites: Development of an instrument", proceedings of the *20th international conference on Information Systems*, pp.279–285.

[59] Keeton, K., Mehra, P. and Wilkes, J. (2010), "Do you know your IQ? A research agenda for information quality in systems" *SIGMETRICS Performance Evaluation Review*, Vol 37, Issue 3, pp. 26-31.

[60] Kilpikoski, S., Airaksinen, O., Kankaanpää, M., Leminen, P., Videman, T. and Alen, M. (2002), "Interexaminer reliability of low back pain assessment using the McKenzie method", *Spine*, pp E207–E214.

[61] Klein B. D. (2002), "When do users detect information quality problems on the World Wide Web?", *American Conference in Information Systems*, pp1101.

[62] Knight,S. (2011), "The combined conceptual life-cycle model of information quality: part 1, an investigative framework", *International Journal of Information Quality*. Vol 2, pp 205-230.

[63] Knight,S and Burn,J. (2005), "Developing a framework for assessing Information Quality on the World Wide Web". *Informing Science Journal*., Vol8, pp.159-172.

[64] Lee, Y. W., Pipino, L. L., Funk, J. D., Wang, R. Y. (2006), "*Journey to Data Quality*", MIT Press.

[65] Lee, Y.W., Strong, D.M, Kahn, B.K, Wang, R.Y. (2002), "AIMQ: A methodology for information quality assessment", *Information and Management* Vol. 40, No. 2, pp. 133–146.

[66] Leung, H. K. N. (2001), "Quality metrics for intranet applications", *Information & Management* , Vol. 38, No.3, pp. 137-152.

[67] Lima, Luís Francisco Ramos. Maçada, Antonio Carlos Gastaud and Vargas, Lilia Maria. (2006), "Research into Information Quality – A study of the state-of-the art IQ and its consolidation", proceedings of the *International Conference on Information Quality*.

[68] Long. J and Seko,C. (2005), "A cyclic-hierarchical method for database data-quality evaluation and improvement", *Advances in Management Information Systems*, Vol. 1 .

[69] Loshin, D. (2004), "*Enterprise Knowledge Management - The Data Quality Approach*", The Morgan Kaufmann Series in Data Management Systems , chapter 4.

[70] Lyn, R. (2009), "IT Metrics: Measuring IT's Business Value", Gartner for Leader. ID:G00203680.

[71] Madnick, S.,Wang,R.Y. and Zhu,H. (2009), "Overview and Framework for Data and Information Quality Research", *ACM Journal of Data and Information Quality*, Vol 2, pp 1-22.

[72] Marshall, L., De La Harpe, R. (2010), "Decision making in the context of business intelligence and data quality", *SA Journal of Information Management*, Vol. 11, pp 1-15.

[73] Naumann,F. and Rolker, C. (2000), "Assessment methods for Information Quality criteria", proceedings of the *14th International Conference on Information Quality*. pp.148-162.

[74] Olson, J. (2003), "*Data Quality, The Accuracy Dimension*", Morgan Kaufmann Publishers. 2003.

[75] Otto, B., Huner,K. and Osterle,H. (2009), "Identification of Business Oriented Data Quality Metrics" proceedings of the *14th International Conference on Information Quality*.

[76] Petersen, T., Olsen, S., Laslett, M. and Thorsen, H. (2004), "Inter-tester reliability of a new diagnostic classification system for patients with non-specific low back pain", *Australian Journal of Physiotherapy*, Vol. 50 pp. 85-91.

[77] Petkus, T., and E. Filatovas (2008), "Decision making to solve multiple criteria optimization problems in computer networks", *Information Technology and Control*, Vol. 37, pp 63–68.

[78] Pfeffer, J., Sutton, R., Bazerman, Max H., Chugh, D.and Davenport, T. (2006), "To Make the Best Decisions, Demand the Best Data", Harvard Business Review, pp. 225-240.

[79] Pipino, L. L., Wang, R. Y., Kopsco, D., and Rybolt, W. (2005), "Developing measurement scales for data-quality dimensions", *Information quality*, pp. 37-51.

[80] Pipino.L, Lee.Y and Wang.R. Y. (2002), "Data quality assessment", *Communications of ACM*, Vol. 45, No.4, pp. 211-218.

[81] Portela,F., Vilas, Marta-boas and Filipe, Manuel Santos. (2010), "Improvements in data quality for decision support in Intensive Care", *3rd International ICST Conference on Electronic Healthcare for the 21st century*.

[82] Prat, N and Madnick, S. (2008), "Measuring Data Believability: a Provenance Approach", proceedings of the *41st Hawaii International Conference on System Sciences*.

[83] Price, R and Shanks, G. (2011), "The Impact of Data Quality Tags on Decision-Making Outcomes and Process", *Journal of the Association for Information Systems*, Vol. 12, Issue. 4, Article 1.

[84] Price,R and Shanks,G. (2010), "DQ Tags and Decision-Making", *43rd Hawaii International Conference on System Sciences (HICSS)*, pp 1-10.

[85] Price,R and Shanks,G. (2008), "Data Quality and Decision Making", *Handbook on Decision Support Systems*, pp 65-82.

[86] Roberts, R. (2011), "A rising role for IT: McKinsey Global Survey results", McKinsey Global Survey results.

[87] Rodríguez, N and Casanovas, J. (2010), "A structural model of information system quality: an empirical research", proceedings of the *Americas Conference on Information Systems*.

[88] Roger, B and Shankaranarayanan, G. (2010), "Framing Data Quality Research : A Semantic Analysis Approach", proceedings of the *15th International Conference on Information Quality*.

[89] Rudra, Amit and Yeo, Emilie. (1999), "Key Issues in Achieving Data Quality and Consistency in Data Warehousing among Large Organisations in Australia", *32nd Hawaii International Conference on System Sciences*.

[90] Ryan, Anne B. (2006), "*Post-Positivist Approaches to Research. In: Researching and Writing your thesis: a guide for postgraduate students*", MACE: Maynooth Adult and Community Education, pp. 12-26.

[91] Sadiq, S., Indulska,M. and Jayawardene,V. (2011), "Research and industry synergies in data quality management", proceedings of the *16th International Conference on Information Quality*. pp. 314-326.

[92] Sadiq, S., Indulska, M. and Khodabandehloo, N.Y. (2011), "20 years of data quality research: Themes, trends and synergies", proceedings of the *22nd Australasian Database Conference (ADC 2011),* pp. 1-10.

[93] Salkind, N. J. (2007), "*Encyclopedia of measurement and statistics*", SAGE Publications. pp E207–E214.

[94] Scannapieco, M., Pernici, B., and Pierce, E. (2005), "IP-UML: A methodology for quality improvement-based on IP-MAP and UML", *Advances in Management Information Systems*, Vol. 1.

[95] Shankar, G. and Watts, S. (2003), "A relevant, believable approach for data quality assessment", proceedings of *8th International Conference on Information Quality*, pp. 178–189.

[96] Shankaranarayanan,G and Cai,Y. (2006), "Supporting data quality management in decision-making", *Decision Support Systems*, Vol 42, pp. 302-317.

[97] Shankaranarayanan, G. Zhu,B. and Cai,Y. (2008), "Decision support with data quality metadata", proceedings of the *International Conference on Information Quality*.

[98] Shankaranarayanan,G. (2005), "Towards implementing total data quality management in a datawarehouse", *Journal of Information Technology Management*, Vol XVI, pp. 21-30.

[99] Shankaranarayanan,G and Bin, Zhu. (2012), "Data Quality Metadata and Decision Making", *45th Hawaii International Conference on System Sciences*, pp.1434-1443.

[100] Shanks, G. and Corbitt, B. (1999), "Understanding data quality: Social and cultural aspects", proceedings of the *10th Australasian Conference on Information Systems*.

[101] Sheldon, M. R., Fillyaw, M. J. and Thompson, W. D. (2006), "The use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures designs", *International Journal of Physiotherapy Research*, Vol. 1 pp. 221-228.

[102] Siegel,S. and Castellan, N. J. Jr. (1998), "*Nonparametric Statistics for the Behavioral Sciences*", McGraw-Hill, second Edition.

[103] Sim, J and Wright, C. (2005), "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements", *Journal of the American Physical Therapy Association*, vol. 85 No. 3 pp. 257-268.

[104] Singh,R and Singh, K.Dr. (2010), "A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing", *International Journal of Computer Science Issues*, Vol. 7, Issue 3, No 2, pp 41-50.

[105] Sintchenko,V., Magrabi, F. and Tipper, S. (2007), "Are we measuring the right end-points? Variables that affect the impact of computerised decision support on patient outcomes: A systematic review", *Medical Informatics and the Internet in Medicine*. pp. 225-240.

[106] Slone, John P. (2006), "Information Quality Strategy : An empirical investigation of the relationship between information quality improvements and organizational outcomes". Ph.D. Thesis, Capella University.

[107] Stvilia,B., Gasser, L.,Michael, B. T and Linda, C. (2007), "A framework for Information Quality Assessment", *Journal of the American Society for Information Science and Technology*, Vol 58, No. 12, pp. 1720-1733.

[108] Su, Y and Jin, Z. (2004), "A methodology for information quality assessment in the designing and manufacturing processes of mechanical products", proceedings of the *9th International Conference on Information Quality* (ICIQ), pp. 447–465.

[109] Sundararaman, Arun. (2012). "A Framework for Study of Data Quality and its impact on Quality of Medical Decisions in Clinical Decision Support Systems", *Indian Journal of Medical Informatics*, Vol. 6, No. 1.

[110] Tansey, O. (2007), "Process Tracing and Elite Interviewing: A Case for Non-probability Sampling", *Political Science & Politics*, Vol 40, pp 765-772 .

[111] The Indian Bankers' Association. (2007), *The Indian Banker*, pp 55-59.

[112] The Reserve Bank of India. (2011), "*Report of the high level committee for preparation of the information technology vision document 2011-17*".

[113] Ustinovichius, L., Zavadskas, E.K. and Podvezko, V. (2007), "Application of a quantitative multiple criteria decision making (MCDM-1) approach to the analysis of investments in construction", *Control and Cybernetics*, Vol 36, pp. 251–268.

[114] Wang, R. Y. (1998), "A product perspective on total data quality management", *Communications of ACM* , Vol.41, No,2, pp. 58-65.

[115] Wang, R. Y., Storey, V. and Firth, C.P. (1995), "A Framework for Analysis of Data Quality Research", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 7, No.4, pp. 623-640.

[116] Wang, R.Y., Reddy, M.P. and Kon, H.B. (1995), "Towards Quality Data - An attribute-based Approach", *Decision Support Systems*, Vol. 13, pp. 349-372.

[117] Wang, R.Y. & Strong, D.M. (1996), "Beyond accuracy: What data quality means to data consumers", *Journal of Management Information Systems*, Spring, pp. 5–33.

[118] Wang, W.M. & Cheung, C.F. (2011), "A narrative-based reasoning with applications in decision support for social service organizations", *Expert Systems with Applications*, Volume 38, Issue 4, pp. 3336-3345.

[119] Weber, K., Otto,B., and ¨Osterle,H. (2009), "One Size Does Not Fit All—A Contingency Approach to Data Governance", *Journal of Data and Information Quality*, Volume 1, Issue 1.

[120] Wende, K. (2007), "A Model for Data Governance – Organising Accountabilities for Data Quality Management", proceedings of *18th Australasian Conference on Information Systems*, pp. 417-425.

[121] Woodall,P and Parlikad, A. (2010), "A hybrid approach to assessing data quality", proceedings of the *15th International Conference on Information Quality*.

[122] Yaari, E., Baruchson, Shifra -Arbib and Judit Bar-Ilan. (2011), "Information quality assessment of community generated content: A user study of Wikipedia", *Journal of Information Science*, vol. 37, No. 5, pp. 487-498.

[123] Zeist, R.H.J. and Hendriks, P.R.H. (1996), "Specifying software quality with the extended ISO model", *Software Quality Journal*, Vol. 5, No. 4, pp. 273-284.

[124] Zhu, X. and Gauch, S. (2000), "Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web", proceedings of the

23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp.288–295.

[125] Zavadskas, E.K., Kaklauskas.A., Turskis.Z., and Tamošaitien˙e. (2009), "Multi-Attribute Decision-Making Model by Applying Grey Numbers", *Informatica*, pp. 305-320.

Appendices

Appendix 1 :: Preliminary Survey Instrument

SURVEY INSTRUMENT

Please read this sheet carefully before you proceed with responding to Survey…….

Structure of the Survey

This data gathering exercise comprises the following Survey steps that are administered in different phases:

- **Initial Survey to Banking Experts**
- Survey to validate the DQ Scoring Framework
- Survey to capture end users confidence factor

Objectives

1. The key objectives of these surveys is to gather data from experts in the field of Banking, based on their vast experience in the Industry, use such data to validate the model proposed by the Researcher and to analyze the data in testing the Hypothesis involved in the research work.
2. The objective of this initial Survey No. 1 is to capture basic demographics and other information related to the Banking Experts so that these additional data elements can be used appropriately in later phases of data analysis and hypothesis testing.

Instructions

1. Please choose the most appropriate response from the options provided.
2. If you choose to response "Others", please specify / expand what you mean by others.
3. In questions related to your organization, have the context of your enterprise that you work for (e.g. Global Bank Ltd.) and not merely the specific entity that you work for within the enterprise (e.g. Retail Banking – North America Organization).

Part I – General

The following items address basic information about the organization in which you work and the nature of your interaction with computer-based Decision Support / Data Warehousing Systems.

1. In my work for this organization, I regularly interact with computer-based information systems in the following ways (check all that apply):

a) Prepare analytical reports (e.g. ad hoc queries that slice and dice data or drill down for additional information etc.)
b) Publish reports for others use
c) Look up information
d) Update or modify data in central repository
e) Perform modeling simulation or analysis
f) Monitor key operational metrics (e.g., trend of Ageing, sales by region etc.)
g) Monitor key strategic metrics through Dashboards (e.g., Growth percentage, utilization % etc.)
h) Prepare and provide reports to Management Teams or Governing bodies to support their decision making process
i) Manage, operate, or administer Decision Support Systems

Part II – Classification Data

You are almost finished! Questions in the final section will be used for classification and analysis by subgroups only. Please provide the appropriate response to each item.

1. Which of the following best describes the type of organization you work for?

a. For-profit.

b. Non-profit.

c. Governmental agency.

d. Other.

2. Which of the following best describes the industry in which you work or are most closely associated?

a. Manufacturing

b. Engineering

c. Transportation

d. Hospitality

e. Health care

f. Education

g. Other

3. What is the primary business activity at your location?

a. Banking

b. Insurance

c. Research and development

d. Manufacturing

e. Transportation

f. Hospitality

g. Health care

h. Retail

i. Education

j. Other

4. Which primary business function have you experience in working over your career?

a. Retail Banking

b. Mortgage Banking

c. Credit Cards

d. Loans – Policy formulation

e. Statutory compliance

f. Central processing

g. Credit Processing

5. How many employees work at your location?

a. Under 100

b. 101 to 1,000

c. 1,001 to 10,000

d. Over 10,000

6. How many employees are there in your entire organization?

a. Under 100

b. 101 to 1,000

c. 1,001 to 10,000

d. Over 10,000

7. What are your organization's approximate annual revenues in U.S. dollars or equivalent (approximate budget if non-profit or governmental)?

a. Under $1 million

b. At least $1 million, less than $10 million

c. At least $10 million, less than $100 million

d. At least $100 million, less than $1 billion

e. Greater than $1 billion

8. How long have you been with this organization?

a. Less than 1 year

b. At least 1 year, less than 5 years

c. At least 5 years, less than 10 years

d. At least 10 years, less than 20 years

e. 20 years or more

9. How long have you been in this industry?

a. Less than 1 year

b. At least 1 year, less than 5 years

c. At least 5 years, less than 10 years

d. At least 10 years, less than 20 years

e. 20 years or more

10. How long have you been in the primary function in Q. No 4 above?

a. Less than 1 year

b. At least 1 year, less than 5 years

c. At least 5 years, less than 10 years

d. At least 10 years, less than 20 years

e. 20 years or more

11. How long have you been using the Decision Support / DW System that was the subject of this survey?

a. Less than 1 year

b. At least 1 year, less than 2 years

c. At least 2 years, less than 3 years

d. At least 3 years, less than 5 years

e. 5 years or more

12. Which of the following best describes your job title or function?

a. Executive

b. Management

c. Sales / Marketing

d. Finance / Accounts

e. Procurement

f. Inventory / Warehouse Management

g. Consultant

h. Engineer

i. Researcher

j. IT Professional

k. Professional (other than IT)

l. Administration

m. Other

Appendix 2 :: Survey Instrument 2

SURVEY INSTRUMENT :: DQ Score Model Validation & Calibration

Please read this sheet carefully before you proceed with responding to Survey…….

Structure of the Survey

This data gathering exercise comprises the following Survey steps that are administered in different phases:

- Initial Survey to Banking Experts
- **Survey to validate the DQ Scoring Framework**
- Survey to capture end users confidence factor

Objectives

1. The key objectives of these surveys is to gather data from experts in the field of Banking, based on their vast experience in the Industry, use such data to validate the model proposed by the Researcher and to analyze the data in testing the Hypothesis involved in the research work.
2. The objective of this Survey No. 2 is to get the researcher's proposed model (of association between IQ factors and Business Outcomes) validated.
3. The other objective of this Survey is to calibrate the above model, which can be the basis for subsequent phases of the experiment.

Part I – Validation of impact map & calibrating the model

Credit Decisions: According to the Researcher, with respect to credit decisions in Retail Banking Industry, the Information Quality factors with potential impact are listed below. Please complete the table with your responses (Columns C and D based on the Instructions as above.

IQ Criteria \ Decision Category	Weightage (as per Model)	Banking Expert suggested	Remarks

(A)	(B)	Weightage (H / M / L / No) (C)	(D)
Believability	H		
Concise representation	No		
Interpretability	H		
Reputation	M		
Understandability	H		
Value-added	H		
Granularity	H		
Relevancy – Measures	H		
Relevancy – Dimensions	H		
Aggregation	L		
Completeness	H		
Customer Support	L		
Documentation	L		
Objectivity	M		
Price	No		
Reliability	H		
Security	M		
Accuracy	H		
Availability	H		
Consistency	H		
Latency	No		
Response Time	No		
Timeliness	M		
Verifiability	M		

Business Promotion Decisions: According to the Researcher, with respect to decisions involving business promotion in Retail Banking Industry, the Information Quality factors with potential impact are listed below. Please complete the table with your responses (Columns C and D based on the Instructions as above.

IQ Criteria \ Decision Category (A)	Weightage (as per Model) (B)	Banking Expert suggested Weightage (H / M / L / No) (C)	Remarks (D)
Believability	No		
Concise representation	No		
Interpretability	M		
Reputation	No		
Understandability	No		
Value-added	H		
Granularity	H		
Relevancy – Measures	H		
Relevancy – Dimensions	H		
Aggregation	M		
Completeness	M		
Customer Support	No		
Documentation	M		
Objectivity	No		
Price	No		
Reliability	M		
Security	M		
Accuracy	No		

Availability	No		
Consistency	No		
Latency	No		
Response Time	No		
Timeliness	No		
Verifiability	No		

Product Decisions: According to the Researcher, with respect to decisions involving Products in Retail Banking Industry, the Information Quality factors with potential impact are listed below. Please complete the table with your responses (Columns C and D based on the Instructions as above.

IQ Criteria \ Decision Category (A)	Weightage (as per Model) (B)	Banking Expert suggested Weightage (H / M / L / No) (C)	Remarks (D)
Believability	H		
Concise representation	M		
Interpretability	M		
Reputation	H		
Understandability	H		
Value-added	M		
Granularity	H		
Relevancy – Measures	H		
Relevancy – Dimensions	H		
Aggregation	M		
Completeness	No		

Customer Support	M		
Documentation	M		
Objectivity	No		
Price	H		
Reliability	H		
Security	H		
Accuracy	No		
Availability	No		
Consistency	No		
Latency	No		
Response Time	No		
Timeliness	M		
Verifiability	No		

Tactical Decisions: According to the Researcher, with respect to decisions involving tactical / operational in Retail Banking Industry, the Information Quality factors with potential impact are listed below. Please complete the table with your responses (Columns C and D based on the Instructions as above.

IQ Criteria \ Decision Category (A)	Weightage (as per Model) (B)	Banking Expert suggested Weightage (H / M / L / No) (C)	Remarks (D)
Believability	H		
Concise representation	H		
Interpretability	H		
Reputation	H		

Understandability	H		
Value-added	M		
Granularity	H		
Relevancy – Measures	H		
Relevancy – Dimensions	L		
Aggregation	H		
Completeness	H		
Customer Support	No		
Documentation	H		
Objectivity	No		
Price	H		
Reliability	H		
Security	H		
Accuracy	No		
Availability	H		
Consistency	No		
Latency	H		
Response Time	H		
Timeliness	M		
Verifiability	H		

Customer Relationship Decisions: According to the Researcher, with respect to decisions involving Customer Relationship in Retail Banking Industry, the Information Quality factors with potential impact are listed below. Please complete the table with your responses (Columns C and D based on the Instructions as above.

IQ Criteria \ Decision Category	Weightage (as per Model)	Banking Expert suggested	Remarks

(A)	(B)	Weightage (H / M / L / No) (C)	(D)
Believability	No		
Concise representation	L		
Interpretability	H		
Reputation	H		
Understandability	H		
Value-added	M		
Granularity	H		
Relevancy – Measures	H		
Relevancy – Dimensions	L		
Aggregation	M		
Completeness	H		
Customer Support	No		
Documentation	M		
Objectivity	No		
Price	H		
Reliability	M		
Security	H		
Accuracy	H		
Availability	H		
Consistency	No		
Latency	H		
Response Time	H		
Timeliness	M		
Verifiability	No		

Regulatory Decisions: According to the Researcher, with respect to decisions involving Regulatory aspects in Retail Banking Industry, the Information Quality factors with potential

impact are listed below. Please complete the table with your responses (Columns C and D based on the Instructions as above.

IQ Criteria \ Decision Category (A)	Weightage (as per Model) (B)	Banking Expert suggested Weightage (H / M / L / No) (C)	Remarks (D)
Believability	H		
Concise representation	No		
Interpretability	No		
Reputation	H		
Understandability	H		
Value-added	M		
Granularity	H		
Relevancy – Measures	H		
Relevancy – Dimensions	M		
Aggregation	No		
Completeness	H		
Customer Support	L		
Documentation	H		
Objectivity	H		
Price	No		
Reliability	H		
Security	H		
Accuracy	H		
Availability	H		

Consistency	H		
Latency	No		
Response Time	No		
Timeliness	H		
Verifiability	H		

Note – Checklist for completion of survey

Before you close this Survey, please pay attention to tasks listed in the following checklist.

- You have verified the mapping details (mapping between Information Quality factors and decision criteria).
- You have completed in all respects, the weigtage associated with each of the IQ Factors for all selected decision criteria.
- You have objectively used the overall industry perspective in mapping the influencing IQ factors (as opposed to organization specific experience)
- You have evaluated the weightage information in broader industry perspective.
- Wherever, definitions were not clear or were closely related to other IQ factor, you refer to the Definitions Appendix and applied the context provided in the definitions.
- You have carefully re-examined all your responses before final submission.

Appendix 3 :: Survey Instrument 3A

SURVEY INSTRUMENT :: DQ Score – User Confidence Factors

Please read this sheet carefully before you proceed with responding to Survey…….

Structure of the Survey

This data gathering exercise comprises the following Survey steps that are administered in different phases:

- Initial Survey to Banking Experts
- Survey to validate the DQ Scoring Framework
- **Survey to capture end users confidence factor**

Objectives

The key objectives of this survey is to gather data from users of Data Warehouse System in the field of Banking i.e. gather their confidence factors in the Data Quality on how much each factor impacts the Quality of business outcome for which the business decisions are taken. Such data is intended to be used to validate the model proposed by the Researcher and to analyze the data in testing the Hypothesis involved in the research work.

Instructions

1. Familiarize yourselves with the definitions of the Information Quality factors, together with appropriate examples, given as Appendix to this Survey.
2. While updating your confidence level in each of the IQ factors, please examine the same in the context of how this is applicable based on your specific experience in the DW Application that you have been using as part of your business decision making process in your organization.
3. Please evaluate your confidence level on the IQ factors based on repeated / time tested instances and not just by exceptional occurrences.
4. This Survey is organized in 2 Sections i.e.
    a. Section I to capture general information related to your role in the Organization and the Decision Support System that is the basis for this Survey response and
    b. Your response on the confidence on the data quality related to the above System

Part I – General

The following items address basic information about the organization in which you work and the nature of your interaction with computer-based Decision Support / Data Warehousing Systems.

1. In my work for this organization, I regularly interact with computer-based information systems in the following ways (check all that apply):

a) Prepare analytical reports (e.g. ad hoc queries that slice and dice data or drill down for additional information etc.)
b) Publish reports for others use
c) Look up information
d) Update or modify data in central repository
e) Perform modeling simulation or analysis
f) Monitor key operational metrics (e.g., trend of Ageing, sales by region etc.)
g) Monitor key strategic metrics through Dashboards (e.g., Growth percentage, utilization % etc.)
h) Prepare and provide reports to Management Teams or Governing bodies to support their decision making process
i) Manage, operate, or administer Decision Support Systems

Instruction: For this section of the survey, please think of a particular information system that you currently interact with in the performance of your job. This system can be a report that you receive regularly, an interactive system that you update, a Dashboard application that you interact with, a system that you operate or are deploying, or something similar. Please select a single system and keep this system in mind as you respond to the items in this section.

3. Please indicate which of the following best describes the nature of your interaction with the system you have selected:

a) I receive reports from this system
b) I provide information to this system for others to use
c) I use this system to look up information
d) I update or modify the data in this system
e) I use this system to perform modeling simulation or analysis
f) I use this system to monitor status of something (e.g., shipping, manufacturing, inventory)

g) I am responsible for managing, operating, or administering this system

Following Questions in this section will be used for classification and analysis by subgroups only. Please provide the appropriate response to each item.

4. Which of the following best describes the type of organization you work for?
a. For-profit.
b. Non-profit.
c. Governmental agency.
d. Other.

5. Which of the following best describes the industry in which you work or are most closely associated?
a. Manufacturing
b. Engineering
c. Transportation
d. Hospitality
e. Health care
f. Education
g. Other

6. What is the primary business activity at your location?
a. Banking
b. Insurance
c. Research and development
d. Manufacturing
e. Transportation
f. Hospitality
g. Health care
h. Retail
i. Education

j. Other

7. How many employees work at your location?

a. Under 100

b. 101 to 1,000

c. 1,001 to 10,000

d. Over 10,000

8. How many employees are there in your entire organization?

a. Under 100

b. 101 to 1,000

c. 1,001 to 10,000

d. Over 10,000

9. What are your organization's approximate annual revenues in U.S. dollars or equivalent (approximate budget if non-profit or governmental)?

a. Under $1 million

b. At least $1 million, less than $10 million

c. At least $10 million, less than $100 million

d. At least $100 million, less than $1 billion

e. Greater than $1 billion

10. How long have you been with this organization?

a. Less than 1 year

b. At least 1 year, less than 5 years

c. At least 5 years, less than 10 years

d. At least 10 years, less than 20 years

e. 20 years or more

11. How long have you been in this industry?

a. Less than 1 year

b. At least 1 year, less than 5 years

c. At least 5 years, less than 10 years

d. At least 10 years, less than 20 years

e. 20 years or more

12. How long have you been using the Decision Support / DW System that was the subject of this survey?

a. Less than 1 year

b. At least 1 year, less than 2 years

c. At least 2 years, less than 3 years

d. At least 3 years, less than 5 years

e. 5 years or more

13. Which of the following best describes your job title or function?

a. Executive

b. Management

c. Sales / Marketing

d. Finance / Accounts

e. Procurement

f. Inventory / Warehouse Management

g. Consultant

h. Engineer

i. Researcher

j. IT Professional

k. Professional (other than IT)

l. Administration

m. Other

14. Which of the following best describes your highest level of education?

a. High school or equivalent

b. Technical school certification

c. Associate's degree

d. Bachelor's degree

e. Master's or Specialist's degree

f. Doctoral degree or beyond

Part II – Data Quality Model

Instructions

1. Familiarize yourselves with the definitions of the Information Quality factors, together with appropriate examples, given as Appendix to this Survey.
2. While responding with your confidence level in the data, please examine the same in the context of data quality as it exists in your Organization.
3. Please evaluate the confidence level based on repeated / time tested instances and not just by exceptional occurrences.
4. Please provide your confidence level in column B in the below table. Please indicate your confidence level as High or Medium or Low level or as No confidence / No basis.
5. Please include additional remarks, if any, on the DQ factors and associated confidence level.

The objective of the following section of the Survey is to capture your views on the extent to which Information Quality exists within your Organization. This section focuses on measuring your confidence in the quality of data perceived / experienced by you in your organization within the purview of the Application System listed in the earlier section of this Survey.

According to the Researcher the Information Quality factors with potential impact are listed below. Please complete the table with your responses (Columns B and C based on the Instructions as above.

IQ Criteria \ Decision Category	Your confidence level in	Remarks

(A)	data quality (H / M / L / No) (B)	(C)
Believability		
Concise representation		
Interpretability		
Reputation		
Understandability		
Value-added		
Granularity		
Relevancy – Measures		
Relevancy – Dimensions		
Aggregation		
Completeness		
Customer Support		
Documentation		
Objectivity		
Price		
Reliability		
Security		
Accuracy		
Availability		
Consistency		
Latency		
Response Time		
Timeliness		
Verifiability		

Appendix 4 :: Survey Instrument 3B

SURVEY INSTRUMENT :: DQ Score – Independence Assessment

Please read this sheet carefully before you proceed with responding to Survey…….

Structure of the Survey

This data gathering exercise comprises the following Survey steps that are administered in different phases:

- Initial Survey to Banking Experts
- Survey to validate the DQ Scoring Framework
- **Survey to capture end users independent assessment**

Objectives

The key objectives of this survey is to gather data from users of Data Warehouse System in the field of Banking i.e. gather their independent assessment on the quality of data in Data captured through their experience of overall Data Quality in the Decision Support System. This data is intended to be used to validate the model proposed by the Researcher and to analyze the data in testing the Hypothesis involved in the research work.

Instructions

1. While responding to questions related to Data Quality Score, please examine the same in the context of how this is applicable based on your specific experience in the DW Application that you have been using as part of your business decision making process in your organization.
2. Please evaluate the overall Data Quality Score based on repeated / time tested instances and not just by exceptional occurrences.
3. This Survey is organized in 2 Sections i.e.
    a. Section I to capture general information related to your role in the Organization and the Decision Support System that is the basis for this Survey response and
    b. Your response on the confidence in the quality of data related to the above System

Part I – General

The following items address basic information about the organization in which you work and the nature of your interaction with computer-based Decision Support / Data Warehousing Systems.

1. In my work for this organization, I regularly interact with computer-based information systems in the following ways (check all that apply):

a) Prepare analytical reports (e.g. ad hoc queries that slice and dice data or drill down for additional information etc.)
b) Publish reports for others use
c) Look up information
d) Update or modify data in central repository
e) Perform modeling simulation or analysis
f) Monitor key operational metrics (e.g., trend of Ageing, sales by region etc.)
g) Monitor key strategic metrics through Dashboards (e.g., Growth percentage, utilization % etc.)
h) Prepare and provide reports to Management Teams or Governing bodies to support their decision making process
i) Manage, operate, or administer Decision Support Systems

2. In our Organization,

a) Our product or service operation involves substantial information processing.
b) We have many product or service varieties within a line of products or services.
c) Information is used to a great extent in our production or service operations.
d) Our product or service mainly provides information.
e) Many steps in our production or service operations require the frequent use of information.
f) Customers need a lot of information related to our products or services before purchasing the product or service.
g) Cycle time from the initial order to the delivery of our product or service is long.
h) Information used in our production or service operations is usually accurate.
i) Our product or service is complex (i.e., is contains many parts that must work together).
j) Information used in our production or service operations is frequently updated.

Instruction: For this section of the survey, please think of a particular information system that you currently interact with in the performance of your job. This system can be a report that you receive regularly, an interactive system that you update, a Dashboard application that

you interact with, a system that you operate or are deploying, or something similar. Please select a single system and keep this system in mind as you respond to the items in this section.

3. Please indicate which of the following best describes the nature of your interaction with the system you have selected:

a) I receive reports from this system
b) I provide information to this system for others to use
c) I use this system to look up information
d) I update or modify the data in this system
e) I use this system to perform modeling simulation or analysis
f) I use this system to monitor status of something (e.g., shipping, manufacturing, inventory)
g) I am responsible for managing, operating, or administering this system

Instruction: Each item below addresses your understanding of the benefits your organization derives from the use of the information in this system. For each item, select a number from 1 to 7 that best completes the sentence:

Following Questions in this section will be used for classification and analysis by subgroups only. Please provide the appropriate response to each item.

4. Which of the following best describes the type of organization you work for?

a. For-profit.

b. Non-profit.

c. Governmental agency.

d. Other.

5. Which of the following best describes the industry in which you work or are most closely associated?

a. Manufacturing

b. Engineering

c. Transportation

d. Hospitality

e. Health care

f. Education

g. Other

6. What is the primary business activity at your location?

a. Banking

b. Insurance

c. Research and development

d. Manufacturing

e. Transportation

f. Hospitality

g. Health care

h. Retail

i. Education

j. Other

7. How many employees work at your location?

a. Under 100

b. 101 to 1,000

c. 1,001 to 10,000

d. Over 10,000

8. How many employees are there in your entire organization?

a. Under 100

b. 101 to 1,000

c. 1,001 to 10,000

d. Over 10,000

9. What are your organization’s approximate annual revenues in U.S. dollars or equivalent (approximate budget if non-profit or governmental)?

a. Under $1 million

b. At least $1 million, less than $10 million

c. At least $10 million, less than $100 million

d. At least $100 million, less than $1 billion

e. Greater than $1 billion

10. How long have you been with this organization?

a. Less than 1 year

b. At least 1 year, less than 5 years

c. At least 5 years, less than 10 years

d. At least 10 years, less than 20 years

e. 20 years or more

11. How long have you been in this industry?

a. Less than 1 year

b. At least 1 year, less than 5 years

c. At least 5 years, less than 10 years

d. At least 10 years, less than 20 years

e. 20 years or more

12. How long have you been using the Decision Support / DW System that was the subject of this survey?

a. Less than 1 year

b. At least 1 year, less than 2 years

c. At least 2 years, less than 3 years

d. At least 3 years, less than 5 years

e. 5 years or more

13. Which of the following best describes your job title or function?

a. Executive

b. Management

c. Sales / Marketing

d. Finance / Accounts

e. Procurement

f. Inventory / Warehouse Management

g. Consultant

h. Engineer

i. Researcher

j. IT Professional

k. Professional (other than IT)

l. Administration

m. Other

14. Which of the following best describes your highest level of education?

a. High school or equivalent

b. Technical school certification

c. Associate's degree

d. Bachelor's degree

e. Master's or Specialist's degree

f. Doctoral degree or beyond

Part II – Data Quality Model

Instructions

1. While responding with your confidence level in the data, please examine the same in the context of data quality as it exists in your Organization.
2. Please evaluate the DQ score assessment based on repeated / time tested instances and not just by exceptional occurrences.
3. Please include additional remarks, if any, on the Decision Categories and associated DQ assessment.

The objective of the following section of the Survey is to capture your views on the extent to which Information Quality exists within your Organization. This section focuses on measuring your confidence in the quality of data perceived / experienced by you in your organization within the purview of the Application System listed in the earlier section of this Survey.

Credit Decisions: The quality of data related to credit decisions (and their impact on the quality of business outcome from such decisions) is …. (complete by checking appropriate box below)

Data Quality score	Confidence level / Impact on business outcome	Your Response
0 to 20 %	Very low; adverse	
21 to 40 %	Low; negative	
41 to 60 %	Average; neutral impact	
61 to 80 %	Good; positive	
81 to 100 %	Very Good; highly positive	

Business Promotion Decisions: The Quality of data related to business promotion decisions (and their impact on the quality of business outcome from such decisions) is …. (complete by checking appropriate box below).

Data Quality score	Confidence level / Impact on business outcome	Your Response
0 to 20 %	Very low; adverse	
21 to 40 %	Low; negative	
41 to 60 %	Average; neutral impact	
61 to 80 %	Good; positive	
81 to 100 %	Very Good; highly positive	

Product Decisions: The Quality of data related to product decisions (and their impact on the quality of business outcome from such decisions) is …. (complete by checking appropriate box below)

Data Quality score	Confidence level / Impact on business outcome	Your Response
0 to 20 %	Very low; adverse	
21 to 40 %	Low; negative	
41 to 60 %	Average; neutral impact	
61 to 80 %	Good; positive	
81 to 100 %	Very Good; highly positive	

Tactical Decisions: The Quality of data related to tactical decisions (and their impact on the quality of business outcome from such decisions) is …. (complete by checking appropriate box below)

Data Quality score	Confidence level / Impact on business outcome	Your Response
0 to 20 %	Very low; adverse	
21 to 40 %	Low; negative	
41 to 60 %	Average; neutral impact	
61 to 80 %	Good; positive	
81 to 100 %	Very Good; highly positive	

Customer Relationship Decisions: The Quality of data related to customer relationship decisions (and their impact on the quality of business outcome from such decisions) is …. (complete by checking appropriate box below)

Data Quality	Confidence level / Impact on	Your

score	business outcome	Response
0 to 20 %	Very low; adverse	
21 to 40 %	Low; negative	
41 to 60 %	Average; neutral impact	
61 to 80 %	Good; positive	
81 to 100 %	Very Good; highly positive	

Regulatory Decisions: The Quality of data related to Regulatory decisions (and their impact on the quality of business outcome from such decisions) is …. (complete by checking appropriate box below)

Data Quality score	Confidence level / Impact on business outcome	Your Response
0 to 20 %	Very low; adverse	
21 to 40 %	Low; negative	
41 to 60 %	Average; neutral impact	
61 to 80 %	Good; positive	
81 to 100 %	Very Good; highly positive	

Appendix 5 :: Data Quality factors – Definitions

S No	*Data Quality factor*	*Definition*
1	Believability	The extent to which data are accepted or regarded as true, real and credible
2	Concise representation	The extent to which data are compactly represented without being overwhelming (i.e. brief in presentation, yet complete and to the point)
3	Interpretability	The extent to which data are in appropriate language and units and data definitions are clear
4	Reputation	Knowledge and awareness about the sources from which data is gathered and perception of overall trustworthiness of information available in the Decision Support System.
5	Understandability	The extent to which data are clear without ambiguity and easily comprehended
6	Value-added	The extent to which data are beneficial and provide advantages from their use
7	Granularity	Level of detail (fineness) of data provided for decision making process. Greater the granularity, deeper the level of detail (fineness of data)
8	Relevancy – Measures	Measurements or metrics or facts associated with a business function e.g. # of loans, outstanding amount, interest income etc.
9	Relevancy – Dimensions	Context or perspectives for understanding the above facts i.e. characteristics such as who, what, where, when, how of a measure (subject). E.g. Housing Loan Amount (measure) by region, branch, agent, loan slab, borrower age, borrower occupation etc. (dimensions)
10	Aggregation	Summary or pre-computed measures that are used to

		enhance query performance
11	Completeness	The extent to which data are of sufficient breadth, depth and scope for the task at hand
12	Customer Support	Additional help mechanism on accessing data, understanding data.
13	Documentation	Information on data attributes, their meaning and tips on using information
14	Objectivity	Extent to which data is free from bias or manipulation to direct analysis and decision making to pre-concluded results.
15	Reliability	The extent to which data that are being provided by the Decision Support System, considered as trustworthy for decision making
16	Security	Appropriate level and detail of information is made available to the appropriate authorities within the Organization for decision making.
17	Accuracy	The extent to which data are correct, reliable and free of error
18	Availability	Adequate information is obtainable and accessible
19	Consistency	Extent to which data (source data definition and data capture process) is uniform across time periods, person seeking the information.
20	Latency	The time between initiating a request in the computer System and receiving the information.
21	Response Time	Time between initiating a detailed analysis or report in the Computer System and receiving the same.
22	Timeliness	The extent to which the age of the data is appropriate for the task at hand
23	Verifiability	Extent to which correctness of data can be validated.

Appendix 6 :: List of publications and presentations

Event	**3rd International Conference on Trendz in Information Sciences**
Date & Venue	8th & 9th December, 2011; Chennai, India
Organizers	Sathyabama University, Chennai in association with IEEE, NASSCOM and others (http://www.tisc2011.com/conference-highlights.html)
Highlight	Paper archived in IEEE xplore http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6169110&contentType=Conference+Publications&pageNumber%3D2%26queryText%3DTISC+2011

3rd International Conference on

TRENDZ IN INFORMATION SCIENCES & COMPUTING

TISC-2011

8th - 9th December, 2011

Organized by

SATHYABAMA UNIVERSITY

(Established under section 3 of the UGC Act, 1956)

Jeppiaar Nagar, Rajiv Gandhi Road,

Chennai – 600 119.

in association with

NASSCOM IEEE Cognizant WIPRO BANCA SELLA

Certificate of Presentation

This is to certify that Dr./Mr./Ms. ARUN SUNDARARAMAN

of ACCENTURE, CHENNAI has presented a paper

entitled A FRAMEWORK FOR LINKING DATA QUALITY TO BUSINESS OBJECTIVES IN DECISION SUPPORT SYSTEMS

in the 3rd International Conference on "Trendz in Information Sciences & Computing"

organized by Sathyabama University, Chennai during 8th & 9th December, 2011.

Er. MARIE JOHNSON
DIRECTOR
Sathyabama University

Er. MARIAZEENA JOHNSON
DIRECTOR
Sathyabama University

Event	**5th India Software Engineering Conference**
Date & Venue	22nd to 25th February, 2012; Kanpur, India
Organizers	IIT, Kanpur in association with Special Interest Group on Software Engineering (SIGSE) (http://www.csi-sigse.org/isec2012/)
Highlight	Paper presented at Doctoral Symposium and the research work presented at the Symposium

E-mail : ska@iitk.ac.in Fax : +91-512-259 7586, 0725 Phone : +91-512-259 7614 (O), 8460 (R)
URL : www.cse.iitk.ac.in

भारतीय प्रौद्योगिकी संस्थान कानपुर
INDIAN INSTITUTE OF TECHNOLOGY KANPUR
संगणक विज्ञान एवं अभियांत्रिकी विभाग
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Dr. Sanjeev K. Aggarwal
Professor

पत्रालय–आई.आई.टी. कानपुर–208016 (भारत)
P.O.-I.I.T. Kanpur-208016 (India)

December 12, 2011

To Whom It May Concern:

This is to certify that Mr Arun Sundaraman's paper (titled: A Framework for Linking Data Quality to Business Objectives in Decision Support Systems) has been selected for presented at the ISEC 2012 Doctoral Symposium.

Sincerely,

Sanjeev K Aggarwal
General Chair
ISEC 2012

Event	**8th National Conference on Medical Informatics**
Date & Venue	3rd to 5th February, 2012; New Delhi, India
Organizers	All India Institute of Medical Sciences / Indian Association for Medical Informatics (http://www.aiims.edu/aiims/events/ncmi/home.html)
Highlights	• Abstract published in "Abstracts Book" and topic presented in separate session in the conference • Sundararaman, Arun. (2012). "A Framework for Study of Data Quality and its impact on Quality of Medical Decisions in Clinical Decision Support Systems", *Indian Journal of Medical Informatics*, Vol. 6, No. 1.

8th NATIONAL CONFERENCE ON MEDICAL INFORMATICS
3rd -5th Feb. AIIMS, NEW DELHI, INDIA
www.NCMI2012.org

Patrons
Dr. R.C.Deka
Dr.V.M.Kattoch

Co-Patrons
Prof. Rani Kumar
Prof. A.B.Dey

Chairperson
Prof. P.P.Kotwal

Advisory Committee
Prof. Khalid Moidu
Mr S.K.Dey Biswas
Mr. B.S.Bedi
Prof. A.B.Dey
Dr. Arindam Basu
Prof. Shashi Kant
Dr. Zulfiquar hossain Khan
Dr. S.B.Bhattacharyya
Dr. A.Thangaprabhu

Scientific Advisory Board
Dr. S.B.Gogia
Dr. Ashutosh Biswas
Dr. Ashish Suri
Mr. N.K.Jain
Mr Sameshwar Mishra
Prof. Indrajit Bhattacharya
Ms. A. Ramachandran
Ms.Neera Palnija
Dr. Vivak Sahi
Dr. Karanvir Singh
Prof. Deepak Agarwal
Dr. Sanjeev Sood
Dr. D.Lavanian

Organizing Secretary
Mr. Sushil Kumar Mehar

Jt. Organizing Secretary
Dr.(Major) Anil Kuthiala

Treasurer
Mr. Sanjay Gupta

Organizing Committee
Mr. S.N.Ragu Kumar
Mr. Satish prasad
Mr Vinay Pande
Mr. S.P.Singh
Mr. Hari Sankar
Mr. Manoj K. Singh
Ms. Tripta Sharma
Mr. Shyamal Barua
Mr. Pawan Sharma
Mr. Sanjeev Kumar
Ms. Ankita Kumari
Ms. Ritu Gupta
Ms. Archana Tyagi
Ms Neelam Gautam
Mr. A.K.Singh
Mr. Aman Gupta

TO WHOMSOVER IT MAY CONCERN

This is to certify that the abstract titled "**A Framework for linking Data Quality to Quality of Medical Decisions using Clinical Decision Support Systems**" submitted by Arun Sundararaman was included in the abstracts book of the "**National Conference on Medical Informatics, 2012**" held between 3rd February and 5th February at the All India Institute of Medical Sciences by Indian Association for Medical Informatics. Arun Sundararaman presented this topic in a session as part of the conference proceedings.

S.K.Meher,
Organizing Secretary, 8th NCMI 2012
Department of Computer Facility, AIIMS,
Ansari Nagar, New Delhi-110 029 INDIA
ncmi2012aiims@gmail.com
Ph: 91-11-26588332, 9868397023

Appendix 7 :: Profile of Supervisor / Research Guide

Dr. Peter Zei-Chan Yeh

Accenture Technology Labs

50 W. San Fernando, Suite 1200

San Jose, CA 95113, USA

E-mail: peter.z.yeh@accenture.com

Web: http://www.cs.utexas.edu/~pzyeh

Research Interests

My research interests are in large-scale knowledge-based systems; semantic techniques & technologies (e.g. ontology, inference engine, ontology alignment, semantic matching, etc.); data and web mining; and automated natural language understanding. I am also interested in the application of techniques and technologies from these areas to business problems such as competitive intelligence and data & information management.

Education

University of Texas at Austin **Austin, TX**

Ph.D. in Computer Science, 2006.

Thesis: Flexible Semantic Matching of Rich Knowledge Structures.

Advisor: Bruce Porter

University of Texas at Austin **Austin, TX**

M.S. in Computer Science, 2001 (4.0/4.0 GPA).

University of Texas at Austin **Austin, TX**

B.S. in Computer Science with Honors, 1999.

Professional Experience

Research Manager, Accenture Technology Labs, 2006 to present

As a research manager at Accenture Technology Labs I shape and lead research initiatives related to artificial intelligence, semantic technologies, and natural language understanding. My responsibilities include defining and scoping research projects; project implementation; managing programmers; presentation of projects at both business and academic forums; assessing and advising on emerging technologies and vendor products and solutions; and technology transfer of capabilities developed to down stream groups within Accenture.

In this role, I have successfully led (or co-led) the following R&D projects:

- **Enterprise Corporate Radar:** The goal of this project is to provide enterprises with competitive insights about their external ecosystem through a technology platform that can support a wide range of corporate radar applications. These applications will automatically and systematically generate these insights from external information sources such as the Web. To achieve this goal, this project investigated how semantic and natural language understanding technologies can be applied to detect relevant events (e.g. product launch, acquisition, etc.) from the Web and to interpret how these events impact an organization's business. The following proof-of-concept corporate radar applications have been built as part of this project to demonstrate the technical feasibility and business value of this technology platform.

- **Technology Lifecycle Tracker:** An application that automatically tracks the maturity of technologies of interest using the Web to inform technology investment decisions. This application has been successfully piloted with Accenture's Wireless Community of Practice. The Technology Lifecycle Tracker was able to produce maturity assessments for various wireless technologies that were comparable to those given by the human experts.

- **Business Event Advisor:** A research prototype that detects and interprets business threats and opportunities which are relevant to an organization from publicly available information on the Web. This prototype is intended to demonstrate the feasibility of using semantic technologies to support critical business and competitive intelligence functions.

- **Data Management R&D Initiative:** This R&D initiative investigates how semantic and analytic technologies can enable new classes of solutions and capabilities that allow enterprises to more effectively (and efficiently) discover, understand, integrate, and govern their data – especially in a Big Data context. Areas of focus include meta-data management and data semantics; integration of

data across heterogeneous sources within (and outside) the enterprise; data quality; and data lineage. The following solutions has been developed as part of this initiative:

- **Data Quality Rules Accelerator (DQRA):** A technology solution that automatically discovers actionable, client-centric data quality rules that can detect and cleanse data inconsistencies (and conflicts) on a wide range of enterprise initiatives ranging from master data management to business intelligence. DQRA has been successfully used at over a dozen Accenture client engagements, and shown to significantly reduce the amount of time on these engagements (up to 90% in some cases). Moreover, the discovered rules have been show to effectively address data quality problems directly linked to revenue loss and operational inefficiency. DQRA has been successfully transitioned to Accenture's Product & Offerings Development (P&OD) division, and integrated into P&OD's Data Quality Management Services offering.

- **Data Mapping Accelerator (DMA):** A technology solution that aims to significantly reduce the cost associated with discovering mappings across data sources on enterprise initiatives – such as legacy & data migration, data consolidation, data warehousing, and more – by automatically discovering these mappings. DMA is currently being piloted with client teams at Accenture.

- **Data Enrichment Framework (DEF):** A customizable technology framework that can automatically enrich a wide range of data objects – such as customers, products, etc. – using heterogeneous data sources (i.e. structured and unstructured, internal and external, etc). This framework can provide numerous business benefits such as improved data quality (by "filling" in incomplete/missing data) and 360 degree view of the customer (which can enable micro-customer segmentation). DEF is currently being piloted with clients in the Retail and Energy & Resource industry.

- **Data Lineage Tracker:** A technology solution that effectively captures, manages, tracks, and reports rich data lineage information across heterogeneous platforms and applications in service of enterprise activities such as decision making, compliance & audit, data loss protection, and analytics. This solution is currently in the prototype stage.

In addition to the above responsibilities, I also help build working relationships with academia. Through my efforts, Accenture Technology Labs has collaborated with the Multi-Functional Knowledge Base (MFKB) group at The University of Texas at Austin on the Digital Aristotle project funded by Vulcan Inc. and the Kno.e.sis Center at Wright State University.

Consultant, BAE Systems, 2005

I consulted for BAE Systems on several occasions to provide the following services:

- Teach a tutorial on knowledge representation.
- Build a situation awareness ontology to detect and reason about terrorist activities.

Graduate Research Assistant, University of Texas at Austin 1999-2006

As part of my graduate education, I worked as a graduate research assistant in the Multi-Functional Knowledge Base (MFKB) lab headed by Professor Bruce Porter. Under the guidance of Professor Porter, I identified a common requirement of knowledge based systems: determining whether (and how) two knowledge representations match each other. A key challenge of this matching problem is similar information can be expressed in many different ways. My dissertation work explored how to build a semantic matcher to resolve these differences in a robust manner.

Through this work, I successfully applied the resulting semantic matcher to several projects at the MFKB lab.

- In DARPA's Rapid Knowledge Formation project, this matcher was used to match encodings of desirable (and undesirable) military situations to military course of actions to assess their strengths and weaknesses.

- In DARPA's PAL project, this matcher was used to stitch together a user's utterances to build a coherent model of what was said.

- In collaboration with Boeing, this matcher was used to perform both sense disambiguation and semantic role labeling in Boeing's control language system called CPL.

- In the Digital Aristotle project (funded by Vulcan Inc.), this matcher was used in the question answering module of the AURA system to match representations of questions to a knowledge base to select relevant knowledge that answers these questions.

Undergraduate Research Assistant , Applied Research Laboratories 1997-1999

As an undergraduate research assistant, I worked on several projects at the Applied Research Laboratories, which is a part of the University of Texas at Austin. I was involved in reverse

engineering a communications network modeling tool for the air force. I also explored embedding AI technology in real time systems.

Teaching Experience

Assistant Instructor, University of Texas at Austin, Fall 2005, Spring 2006

CS105: Introduction to Perl. The goal of this course is to teach students how to program in Perl and to apply concepts learned in class to real world applications. As the assistant instructor, my duties included designing the course (syllabus, lectures, assignments, and exams), delivering the lectures, and conducting office hours outside of class to assist students.

Publications

Journal and Book Chapter

- **Peter Z. Yeh** and A. Kass (2010). *"A Technology Platform to Enable the Building of Corporate Radar Applications that Mine the Web for Business Insight"*. In C. Soares and R. Ghani (Ed.), Data Mining for Business Applications (pp. 149-163). IOS Press.
- **Peter Z. Yeh,** C. Puri, and A. Kass (2010). "*A Knowledge Based Approach for Capturing Rich Semantic Representations from Text for Intelligent Systems*". International Journal of Advanced Intelligence Paradigms 2(1), 33-48.
- N. Friedland, P. Allen, G. Matthews, M. Witbrock, D. Baxter, J. Curtis, B. Shepard, P. Miraglia, J. Angele, S. Staab, E. Moench, H. Oppermann, D. Wenke, D. Israel, V. Chaudhri, B. Porter, K. Barker, J. Fan, S. Chaw, **Peter Z. Yeh**, D. Tecuci, and P. Clark (2004). "*Project Halo: Towards a Digital Aristotle"*. AI Magazine 25(4), Winter 2004, 29-47.

Conference, Workshop, and Symposium

- K. Gomadam, **Peter Z. Yeh**, and K. Verma (2012). "*Data Enrichment using Web APIs"*. In AAAI Spring Symposium Series, Intelligent Web Services Meet Social Computing. Stanford University, Palo Alto, California.
- **Peter Z. Yeh**, C. Puri, M. Wagman, and A. Easo (2011). *"Accelerating the Discovery of Data Quality Rules: A Case Study"*. Proceedings of the Twenty-Third Innovative Applications of Artificial Intelligence Conference (IAAI 2011). San Francisco, California.
- P. Jain, **Peter Z. Yeh**, K. Verma, R. Vasquez, M. Damova, P. Hitzler, and A. Sheth (2011). *"Contextual Ontology Alignment of LOD with an Upper Ontology: A Case Study with Proton"*. Proceedings of the Eighth Extended Semantic Web Conference (ESWC 2011). Heraklion, Greece.

- P. Jain, P. Hitzler, A. Sheth, K. Verma, and **Peter Z. Yeh** (2010). *"Ontology Alignment of Linked Open Data"*. Proceedings of the Ninth International Semantic Web Conference (ISWC 2010). Shanghai, China.
- **Peter Z. Yeh** and C. Puri (2010). *"Discovering Conditional Functional Dependencies to Detect Data Inconsistencies"*. Proceedings of the Eighth International Workshop on Quality in Databases in conjunction with VLDB 2010 (QDB 2010). Singapore.
- **Peter Z. Yeh** and C. Puri (2010). *"An Efficient and Robust Approach for Discovering Data Quality Rules"*. Proceedings of the IEEE Twenty-Second International Conference on Tools with Artificial Intelligence (ICTAI 2010). Arras, France.
- P. Jain, P. Hitzler, **Peter Z. Yeh**, K. Verma, and A. Sheth (2010). "*Linked Data is More Data"*. In AAAI Spring Symposium Series, Linked Data Meets Artificial Intelligence. Stanford University, Palo Alto, California.
- P. Jain, **Peter Z. Yeh**, K. Verma, C. Henson, and A. Sheth (2009). *"SPARQL Query Re-writing for Spatial Datasets Using Partonomy Based Transformation Rules"*. Proceedings of the Third International Conference on GeoSpatial Semantics (GeoS 2009). Mexico City, Mexico.
- **Peter Z. Yeh**, C. Puri, and A. Kass (2009). *"Towards a Technology Platform for Building Corporate Radar Applications that Mine the Web for Business Insight"*. Proceedings of the IEEE Twenty-First International Conference on Tools with Artificial Intelligence (ICTAI 2009). Newark, New Jersey.
- S. Chaw, K. Barker, B. Porter, D. Tecuci, and **Peter Z. Yeh** (2009). *"A Scalable Problem-Solver for Large Knowledge-Bases"*. Proceedings of the IEEE Twenty-First International Conference on Tools with Artificial Intelligence (ICTAI 2009). Newark, New Jersey.
- **Peter Z. Yeh** and C. Puri (2009). *"A Tool for Measuring the Reality of Technology Trends of Interest"*. Proceedings of the Twenty-First Innovative Applications of Artificial Intelligence Conference (IAAI 2009). Pasadena, California.
- M. Ginsburg, A. Kass, and **Peter Z. Yeh** (2009). *"Exploring Two Enterprise Semantic Integration Systems"*. Proceedings of the Hawaii International Conference on System Sciences (HICCS-42). Big Island, Hawaii.
- **Peter Z. Yeh** and A. Kass (2008). *"Capturing the Semantics of Online News Sources for Business Intelligence Applications"*. Proceedings of the IEEE Twentieth International Conference on Tools with Artificial Intelligence (ICTAI 2008). Dayton, Ohio.
- P. Jain, **Peter Z. Yeh**, K. Verma, A. Kass, and A. Sheth (2008). *"Enhancing Process-Adaptation Capabilities with Web-Based Corporate Radar Technologies"*. Proceedings of the ISWC First International Workshop on Ontology-supported Business Intelligence (OBI 2008). Karlsruhe, Germany.
- **Peter Z. Yeh**, D. Farina, and A. Kass (2008). "*A Knowledge Based Approach for Capturing Rich Semantic Representations from Text*". Proceedings of the Twelfth International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2008). Zagreb, Croatia.
- **Peter Z. Yeh**, A. Kass, and D. Farina (2008). "*Technology Investment Radar: A Tool to Automatically Track Technology Maturation*". Proceedings of the IEEE

Second International Conference on Semantic Computing (ICSC 2008). Santa Clara, California.

- **Peter Z. Yeh**, D. Farina, and A. Kass (2007). "*Semantic Interpretation of the Web without the Semantic Web: Toward Business-Aware Web Processors*". Proceedings of the IEEE First International Conference on Semantic Computing (ICSC 2007). Irvine, California.
- S. Chaw, J. Fan, D. Tecuci, and **Peter Z. Yeh** (2007). *"Capturing a Taxonomy of Failures During Automatic Interpretation of Questions Posed in Natural Language"*. Proceedings of the Fourth International Conference on Knowledge Capture (K-CAP 2007). Whistler, British Columbia.
- P. Clark, S. Chaw, K. Barker, V. Chaudhri, J. Thompson, P. Harrison, B. John, B. Porter, A. Spaulding, and **Peter Z. Yeh** (2007). *"Capturing and Answering Questions Posed to a Knowledge-Based System"*. Proceedings of the Fourth International Conference on Knowledge Capture (K-CAP 2007). Whistler, British Columbia.
- K. Barker, B. Agashe, S. Chaw, J. Fan, M. Glass, J. Hobbs, E. Hovy, D. Israel, D. Kim, R. Mulkar, S. Patwardhan, B. Porter, D. Tecuci, and **Peter Z. Yeh** (2007). *"Learning by Reading: A Prototype System, Performance Baseline and Lessons Learned"*. Proceedings of the Twenty-Second National Conference on Artificial Intelligence (AAAI 2007). Vancouver, British Columbia.
- **Peter Z. Yeh,** B. Porter, and K. Barker (2006). "*Flexible Semantic Matching for Link Analysis: A Proposal"*. AAAI Fall Symposium Series, Capturing and Using Patterns for Evidence Detection. Washington, DC.
- **Peter Z. Yeh,** B. Porter, and K. Barker (2006). *"A Unified Knowledge Based Approach for Sense Disambiguation and Semantic Role Labeling"*. Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006). Boston, Massachusetts.
- **Peter Z. Yeh,** B. Porter, and K. Barker (2005). *"Matching Utterances to Rich Knowledge Structures to Acquire a Model of the Speaker's Goal"*. Proceedings of the Third International Conference on Knowledge Capture (K-CAP 2005). Banff, Canada.
- **Peter Z. Yeh**, B. Porter, and K. Barker (2004). "*Mining Transformation Rules for Semantic Matching"*. Proceedings of the ECML/PKDD 2nd International Workshop on Mining Graphs, Trees and Sequences (MGTS'04). Pisa, Italy.
- K. Barker, S. Chaw, J. Fan, B. Porter, D. Tecuci, **Peter Z. Yeh**, V. Chaudhri, D. Israel, S. Mishra, P. Romero, and P. Clark (2004). "*A Question-Answering System for AP Chemistry: Assessing KR&R Technologies"*. Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning (KR 2004). Whistler, British Columbia.
- **Peter Z. Yeh**, B. Porter, and K. Barker (2003). "*Using Transformations to Improve Semantic Matching"*. Proceedings of the Second International Conference on Knowledge Capture (K-CAP 2003). Sanibel Island, Florida.
- K. Barker, J. Blythe, G. Borchardt, V. Chaudhri, P. Clark, P. Cohen, J. Fitzgerald, K. Forbus, Y. Gil, B. Katz, J. Kim, G. King, S. Mishra, C. Morrison, K. Murray, C. Otstott, B. Porter, R. Schrag, T. Uribe, J. Usher, and **Peter Z. Yeh** (2003). "*A Knowledge Acquisition Tool for Course of Action Analysis"*. Proceedings of the

Innovative Applications of Artificial Intelligence Conference (IAAI-2003). Acapulco, Mexico.

- B. Porter, K. Barker, J. Fan, P. Navratil, D. Tecuci and **Peter Z. Yeh** (2002). "*Mining Answers from Texts and KBs: Our Position"*. In AAAI Spring Symposium Series, Mining Answers from Texts and Knowledge Bases. Stanford University, Palo Alto, California.

Thesis and Technical Report

- S. Chaw, K. Barker, B. Porter, and **Peter Z. Yeh** (2007). "*Towards an Ontology-Independent Problem-Solver"*. UT-AI-TR-07-349, Department of Computer Sciences, University of Texas at Austin.
- **Peter Z. Yeh** (2006). *"Flexible Semantic Matching of Rich Knowledge Structures"*. Ph.D. Dissertation, Department of Computer Sciences, University of Texas at Austin.
- **Peter Z. Yeh**, B. Porter, and K. Barker (2003). "*Transformation Rules for Knowledge-Based Pattern Matching"*. UT-AI-TR-03-299, Department of Computer Sciences, University of Texas at Austin.

Other

- K. Gomadam, K. Verma, **Peter Z. Yeh**, A. Sheth, and P. Jain (2011) *"Knowledge Cloud: Harnessing Knowledge on the Web"*. Tutorial at IEEE Fourth International Conference on Cloud Computing (CLOUD 2011). Washington, DC.
- P. Jain, P. Hitzler, K. Verma, **Peter Z. Yeh**, and A. Sheth (2010) *"How to Make Linked Data More Than Data"*. Proceedings of the 2010 Semantic Technology Conference, San Francisco, California.
- P. Jain, A. Sheth, K. Verma, and **Peter Z. Yeh** (2009). "*Extending SPARQL to Support Spatially and Temporally Related Information".* Proceedings of the 2009 Semantic Technology Conference. San Jose, California.
- A. Kass and **Peter Z. Yeh** (2007). "*Business Aware Web Clients".* Proceedings of the 2007 Semantic Technology Conference. San Jose, California.
- **Peter Z. Yeh** (2005). *"Semantic Matching and Its Applications for Calo"*. Poster Presentation at DARPA's Y3 PAL Kick Off Meeting. SRI International, Menlo Park, California.

Professional Services and Activities

Program Committee

- Twenty-Fourth Conference on Innovative Applications of Artificial Intelligence (IAAI 2012).
- AAAI Fall Symposium Series: Open Government Knowledge – AI Opportunities and Challenges (2011).

- International Workshop on Web-Scale Knowledge Extraction (co-located with ISWC 2011).
- Twenty-Third Conference on Innovative Applications of Artificial Intelligence (IAAI 2011).
- Ninth International Semantic Web Conference (ISWC 2010).
- Twenty-Second Conference on Innovative Applications of Artificial Intelligence (IAAI 2010).
- Nineteenth International WWW Conference (WWW 2010).
- Fifth International Conference on Knowledge Capture (K-CAP 2009).
- Third International Conference on Advances in Semantic Processing (SEMAPRO 2009).
- Seventh International Semantic Web Conference (ISWC 2008).
- Second International Conference on Advanced Engineering Computing and Applications in Sciences (ADVCOMP 2008).
- Twenty-Second National Conference on Artificial Intelligence (AAAI 2007).
- Fourth International Conference on Knowledge Capture (K-CAP 2007).
- International Conference on Advances in Semantic Processing (SEMAPRO 2007).
- AAAI Fall Symposium Series: Capturing and Using Patterns for Evidence Detection (2006).

Reviewer

- AI Magazine, Special Issue on Question Answering Systems (2010).
- ACM Transactions on the Web (2006).
- First International Conference on Knowledge Capture (K-Cap 2001).

Member

- Association for the Advancement of Artificial Intelligence.

Awards and Honors

- *"Extended Collaboration Event Monitoring System",* patent pending.
- *"Data Enrichment Using Heterogeneous Sources"*, patent pending.
- *"Data Mapping Acceleration"*, patent pending.
- *"Information Source Alignment",* patent pending.
- *"Data Quality Enhancement for Smart Grid Applications"*, patent pending.
- *"Method and System for Accelerated Data Quality Enhancement"*, patent pending.
- *"Technology Event Detection, Analysis, and Reporting System*", patent pending.
- University Honors 1998 and 1999.

Appendix 8 :: Profile of Author

Arun Sundararaman

Accenture Technology Services P Ltd

Tek Meadows

51, Old Mahabalipuram Road

Sholinganallur, Chennai. 600 119. India

E-mail: arun.sundararaman@accenture.com

Research Interests

My professional and research interest include Information Management, Data Quality in Decision Support Systems. Apart from the current research work, the areas where I am interested in further study / research include Data Quality definitions and implementation involving advanced analytics computing techniques (such as datamining algorithms), Data Quality and data governance approach for Clinical Informatics, Decisions Support in Bio-Informatics.

Education and Professional Membership

- CA Final from The Institute of Chartered Accountants of India 1987
- ACS from The Institute of Company Secretaries of India 1995
- N C Krishnan Medal Prize for All India Best Student in Corporate Management Course Examination from The Institute of Chartered Accountants of India.
- Ph.D. Qualifying Examination, Birla Institute of Technology & Science 2006
- Certification in Management Development from XLRI, Jamshedpur 2002

- Fellow Member of The Institute of Chartered Accountants of India
- Associate Member of The Institute of Company Secretaries of India

Professional Experience

- More than 12 years of specialization with Datawarehousing / Business Intelligence technologies.
- Manage a large delivery team (off shore / onsite) to ensure timely quality deliverables with better productivity / profitability.
- Experience in handling multiple Enterprise wide DW/BI projects for large Banks and Healthcare clients in Asia, US and UK.
- Active role in implementation and practice of Software Quality Management Systems (ISO 90001:2000) and CMM; was one of the Assessors for CMM Assessment (SEI-CMM CBA IPI).

Senior Manager, Accenture, India 2005 to present

- Lead – Healthcare Analytics
- Delivery of Enterprise wide Information Management projects (Datawarehouse / Business Intelligence technologies) for large Healthcare clients.
- Develop competencies in select DW/BI Tools and Technologies
- Provide Thought Leadership initiatives in areas of Data Migration, Data Modeling, ETL Testing.
- Course Design, preparation of materials and roll out of HC201 – Informatics (training for IT professionals for specialization in DW/BI for Healthcare Industry)

Project Manager, HCL Technologies Ltd, India 1998 to 2005

- Head BI function for BFS vertical – business planning, partnership, delivery management
- Team building: mentoring and guiding the team on technical / business areas, development of standards and ensure process compliance
- Manage a large delivery team (off shore / onsite) to ensure timely quality deliverables with better productivity / profitability
- Program Management involving multiple projects for different engagements - ensuring that the project teams meet critical success factors – schedule, profitability, quality metrics, highlights, business continuity, customer satisfaction
- Facilitate exploring business opportunities and drive sales initiatives proposal / bid response preparation, estimation, due diligence, process definition

Product Development, Finance / Accounts Management and MIS 1988 to 1998

- Technical Architecture and design of Online Oil Purchases and Sales System
- Design and Development of Finance & Budgeting Systems
- Development and implementation of Depot Sales System
- Consolidation of Accounts (800 branches)
- Centralized payroll processing
- Functional testing and implementation of new Financial Accounting System
- Functional design – Budgeting & Branch Margin System
- Preparation of Project Reports – for expansion and new business lines

Publications and professional development activities

- Published a whitepaper on Datawarehouse Testing in DMReview which is being widely ready and referred by emerging professionals in Datawarehouse Testing [Testing for DW/BI - Current State and a Peep into the Future]. [http://www.information-management.com/infodirect/20070622/1083670-1.html?pg=2]
- Sundararaman, Arun. (2012). "A Framework for Study of Data Quality and its impact on Quality of Medical Decisions in Clinical Decision Support Systems", *Indian Journal of Medical Informatics*, Vol. 6, No. 1.

- Sundararaman, Arun (2011). "A framework for linking Data Quality to business objectives in decision support systems", proceedings of *3rd international conference on trendz in information science and computing*. (IEEE Xplore: http://ieeexplore.ieee.org/search/searchresult.jsp?queryText%3DSundararaman&pageNumber=2)
- Accepted for publication in Journal: *Advanced Engineering Forum* ISSN: 2234-9898 – "Exploring insights through visualization of association rules from text mining statistics"
- Conceptualized a Framework for definition and measurement of Data Quality in Datawarehouse and other Decision Support Systems; developed a software application for implementing this framework which is being piloted with a few clients.
- Designed and conducted a "Workshop on Data Mining" for Hindustan University, Chennai, August 2012.
- Software Metrics – Faculty Development Program at Kovai Kalaimagal Engineering College in association with All India Council of Technical Education
- Datawarehousing and Professional Opportunities for Chartered Accountants – Erode Branch of The Institute of Chartered Accountants of India
- BS 7799 – Faculty Development Program at Kovai Kalaimagal Engineering College in association with All India Council of Technical Education
- Software Quality (Review and Inspection) – workshop conducted by Accenture for Faculty Development at Anna University, Chennai
- Data Mining – Features, Tools and Techniques at Coimbatore Institute of Technology
- Data Warehousing – A Corporate Banking Case Study – as part of Workshop on Banking Data Warehousing and Data Mining conducted by the Institute for Development and Research in Banking Technology, Hyderabad.
- Data Warehousing – Estimation and Capacity Planning – Accenture workshop
- Software Quality Metrics – Program conducted by Accenture for Faculty Development at NIT, Trichy.
- Session on Information Security Frameworks – Industry View / Applications as part of FDP conducted by Nandha Engineering College, Erode.