THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

A PATTERN BASED APPROACH FOR DATA QUALITY REQUIREMENTS MODELLING

WELGAMAGE VIMUKTHI KAUSHAL JAYAWARDENE

MBA BSc

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2016

The school of Information Technology and Electrical Engineering

Abstract

Data quality (DQ) is a critical issue in today's information systems. Both academic researchers and industry practitioners have contributed to addressing the problem of data quality through management strategies and technological advancements over the last three decades, yet data quality management remains a challenge in the organizational management portfolio. Requirement models have been used successfully to develop solutions in areas like software and database development. The current state of the art in DQ management methodologies developed by both academic researchers and industry practitioners have largely overlooked the area of DQ requirements modelling and analysis. However DQ requirements are fundamental to DQ management since the ultimate purpose of DQ management is to develop solutions to satisfy the DQ requirements. Thus, a clearly defined DQ requirement model is a necessary prelude to systematically develop solutions to organizational DQ problems. In this research, we have developed a repository of thirty-three DQ patterns to model DQ requirements. The patterns are rich in representing the real world DQ requirements while free from notational complexities, thereby allowing them to be used practically to support DQ management. We used design science as the guiding methodology for developing DQ patterns while maintaining a rationale for the rigor and the relevance of our artefacts through appropriate validations and verifications throughout the design process. One of the challenges faced in the conceptualization of DQ patterns was the lack of shared understanding among researchers about DQ dimensions, which is a key concept in representing a DQ requirement. Owing to the importance of shared understanding we systematically refactor the concept of DQ dimensions by consolidating different viewpoints from both academic and practitioner community. As a secondary aim of this study, we adapted a credible requirements engineering methodology from literature to analyse and elicit DQ requirements. We demonstrate through empirical studies that by using this methodology DQ patterns can be effectively used to elicit and model DQ requirements in organizations.

Declaration by author

This thesis is composed of my original work and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

Conference papers:

Zhang, Ruojing, Jayawardene, Vimukthi, Indulska, Marta, Sadiq, Shazia, and Zhou, and Xiaofang Zhou (2014). A data-driven approach for discovering data quality requirements. In: *Proceedings of the Thirty Fifth International Conference on Information Systems* (*ICIS 2014)*, Auckland, New Zealand, 14-17 December 2014.

Jayawardene, Vimukthi, Sadiq, Shazia and Indulska, Marta (2013). The curse of dimensionality in data quality. In: *Proceedings of the 24th Australasian Conference on Information Systems (ACIS 2013)*, Melbourne, VIC, Australia, (1-11). 4-6 December 2013.

Jayawardene, Vimukthi, Sadiq, Shazia and Indulska, Marta (2012). Practical significance of key data quality research areas. In: *Poceedings. Of the 16th Pacific Asia Conference on Information Systems (PACIS 2012)*, Ho Chi Minh City, Vietnam, 11-15 July 2012.

Sadiq, Shazia, Indulska, Marta and Jayawardene, Vimukthi (2011). Research and industry synergies in data quality management. In: *Proceedings of the 16th International Conference on Information Quality (ICIQ2011)*, Adelaide, Australia, (314-326). 18-20 November 2011.

Technical Paper:

Jayawardene, Vimuthki, Sadiq, Shazia and Indulska, Marta (2013) An analysis of data quality dimensions. *ITEE Technical Report 2013-01, School of Information Technology and Electrical Engineering*, The University of Queensland

Publications included in this thesis

Sadiq, Shazia, Indulska, Marta and Jayawardene, Vimukthi (2011). Research and industry synergies in data quality management. In: *Proceedings of the 16th International Conference on Information Quality. 16th International Conference on Information Quality (ICIQ2011)*, Adelaide, Australia, (314-326). 18-20 November 2011.– Partially included in Chapter 1 and 2.

Contributor	Statement of contribution
Jayawardene	Data analysis and interpretation (100%) Paper writing (80%)
Sadiq	Survey design and execution (80%) Paper writing (10%)
Indulska	Survey design and execution (20%) Paper writing (10%)

Jayawardene, Vimukthi, Sadiq, Shazia and Indulska, Marta (2012). Practical significance of key data quality research areas. In: *Proceedings of the 16th Pacific Asia Conference on Information Systems (PACIS 2012)*, Ho Chi Minh City, Vietnam, 11-15 July 2012.– Partially included in Chapter 1 and 2.

Contributor	Statement of contribution
Jayawardene	Data analysis and interpretation (80%) Paper writing (80%)
Sadiq	Data analysis and interpretation (10%) Paper writing (10%)
Indulska	Data analysis and interpretation (10%) Paper writing (10%)

Jayawardene, Vimukthi, Sadiq, Shazia and Indulska, Marta (2013). The curse of dimensionality in data quality. In: *Proceedings of the 24th Australasian Conference on Information Systems (ACIS 2013)*, Melbourne, VIC, Australia, (1-11). 4-6 December 2013. – Partially included in Chapter 4.

Contributor	Statement of contribution
Jayawardene	Concept design (80%) Data analysis and interpretation (80%) Paper writing (80%)
Sadiq	Concept design (10%) Data analysis and interpretation (10%) Paper writing (10%)
Indulska	Concept design (10%) Data analysis and interpretation (10%) Paper writing (10%)

Jayawardene, Vimuthki, Sadiq, Shazia and Indulska, Marta (2013) An analysis of data quality dimensions. *ITEE Technical Report 2013-01, School of Information Technology and Electrical Engineering*, The University of Queensland - Partially included in Chapter 4

Contributor	Statement of contribution
Jayawardene	Concept design (80%) Data analysis and interpretation (80%) Paper writing (80%)
Sadiq	Concept design (10%) Data analysis and interpretation (10%) Paper writing (10%)
Indulska	Concept design (10%) Data analysis and interpretation (10%) Paper writing (10%)

Zhang, Ruojing, Jayawardene, Vimukthi, Indulska, Marta, Sadiq, Shazia, and Zhou, and Xiaofang Zhou (2014). A data-driven approach for discovering data quality requirements. In:*Proceedings of the Thirty Fifth International Conference on Information Systems (ICIS 2014*), Auckland, New Zealand, 14-17 December 2014. – Partially included in Chapter 8.

Contributor	Statement of contribution
Jayawardene	Concept design (20%) Empirical evaluation (20%) Paper writing (20%)
Zhang	Concept design (60%) Empirical evaluation (60%) Paper writing (60%)
Sadiq	Concept design (10%) Empirical evaluation (10%) Paper writing (10%)
Indulska	Concept design (10%) Empirical evaluation (10%) Paper writing (10%)

Contributions by others to the thesis

My supervisors, Professor Shazia Sadiq, and Associate Professor Marta Indulska each contributed to conceptualization, reviewing and editing the Thesis prior to submission. Both of them were also part of the empirical study performed in this thesis.

Statement of parts of the thesis submitted to qualify for the award of another degree

None

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Professor Shazia Sadiq and Associate Professor Marta Indulska for their continuous guidance and support. This thesis would not have been possible without their guidance, support and patience.

Special thanks and acknowledgement go to the chair of my research committee Dr. Mohamed Sharaf and, Professor Andrew Burton-Johns of UQ Business School for their valuable support feedback at various stages of this research project.

I would like to thank all my fellow Ph.D. colleagues in DKE Group especially Thien Wan Au, Syed Norris and Andri Setiawan for their support and encouragement at the initial stage of my research project. Special thank goes to my friend Chinthaka de Silva for his friendship and moral support throughout the time.

Finally, to those most affected and most important. To my parents, thank you for your blessings and well-wishes throughout my life. To my wife, Vijini and my gorgeous two children Chanuga and Chenulee, thank you for everything. Words cannot express my appreciation for all that you have done and all that you have put up with throughout this journey.

Keywords

data quality, data quality patterns, data quality management, data quality dimension, Data quality characteristics, Data quality requirements, data quality requirements analysis, meta-modelling, goal oriented requirement analysis.

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 080609, Information Systems Management, 60%
ANZSRC code: 080603, Conceptual modelling, 20%
ANZSRC code: 060104, Information Systems Development Methodologies, 20%

Fields of Research (FoR) Classification

FoR code: 0806, Information systems 100%

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Background and motivation

Data quality (DQ) has been widely researched over the past two decades (Sadiq et al. 2011b) and has developed into a professional discipline (Yonke et al. 2011), with a prominent focus in organizational strategy (Friedman 2012). Contributions from researchers as well as practitioners have resulted in advancements in data quality management. As a result, the data quality body of knowledge consists of diverse perspectives ranging from advanced computational methods (Al-Hakim 2007; Dasu and Loh 2012; Kimball and Caserta 2004; Sadiq 2013; Talburt 2011) data governance and information management (Batini and Scannapieco 2006; Eppler 2006; Floridi 2011; Wang et al. 2000) and experience based studies from the practioner community (English 2009; Loshin 2011; McGilvray 2008; Redman 2008). Professional initiatives such as the International Association of Information and Data Quality[1], Data Quality Pro[2] and The Data Governance Institute[3] provide further evidence that both academic and industry practitioner community are working towards establishing knowledge sharing and dissemination processes in order to foster the DQ professional community. Industry standard ISO 8000 (ISO 2011) has also emerged for DQ, further strengthening its identity as a professional discipline. Due to the increasing recognition of data as an organizational asset and the risks associated with inadequate data management, a number of regulations are also emerging (Digital Accountability & Transparency Act 2014; Principles for effective risk data aggregation and risk reporting 2013). In spite of several decades of contributions to data quality research and practice, data quality continues to be a significant problem for organizations of all sizes and nature. The first US chief data scientist Dr. Dhanurjay Patil summarises today's data quality problem as "*You have to start with a very basic idea: Data is super messy, and data clean up will always be literally 80% of the work. In other words, data is the problem"* (Caudron and Peteghem 2015). Below we outline and present the underlying motivation for the research objectives of this thesis.

1.1.1 Data quality research

A number of research communities have contributed to the data quality body of knowledge including those pursuing statistics e.g.:-(Cheng et al. 2003; Dasu and Loh 2012; Winkler 2004), computer

[1] http://www.iaidq.org/
[2] http://dataqualitypro.com
[3] http://www.datagovernance.com

science e.g.:-(Cong et al. 2007; Koudas et al. 2006), information systems (Mettler et al. 2008; Otto et al. 2007) and management e.g.:- (Khatri and Brown 2010; Loshin 2004; Wende 2007). Thus the themes in the data quality body of knowledge are significantly diverse.

A number of recent studies outlined below have attempted to structure the body of knowledge in data quality research. Madnick et al. (2009) presented a framework that characterizes data quality research based on the two dimensions of topics and methods, thereby providing a means to classify various works. This research identified four categories of data quality research, each having several subcategories: (1) Data quality impacts: development of methods, designs and test mechanisms that maximize positive impacts of data quality in organisations while minimising negative impacts; (2) Database related technical solutions: development of database technologies for assessing, improving, and managing data quality, including development of techniques for reasoning about data quality and for designing systems that result in data of high quality; (3) Data quality in the context of computer science and IT: technologies and methods (except for the specific database-related techniques) to manage, ensure, and enhance data quality; and (4) Data quality in curation: selection, preservation, and management of digital information in ways that promote easy discovery and retrieval for both current and future uses of that information. Further, they identified fourteen high-level research methodologies used to investigate data quality issues, which provides an indication of the span of the studies so far.

Lima et al. (2006) classified the DQ literature as either theoretical (conceptual, applied, illustrative) or practical (qualitative, experimental, survey, simulation). Further, Neely and Cook (2008) analysed data and information quality literature over the period of 1996-2007 and revealed the over-researched areas namely dimensions, semantics, standards, policies, procedures, strategies of DQ improvement; and the under-researched areas namely research associated with costs of data, and legal implications of fitness for use characteristics of data.

Sadiq et al. (2011b) analysed data quality research literature over the last two decades using over 1400 publications, aiming to identify main themes and to develop a taxonomy of the data quality domain. Their study revealed that the bulk of data quality research spans three distinct communities: business analysts, solution architects, and database experts (Sadiq et al., 2011c). Business analysts focus on organizational solutions for quality data, where they develop data quality objectives for the organization, as well as strategies to establish roles, processes, policies, and standards required to ensure the achievement of data quality objectives. Solution architects are focused on architectural solutions, where they design and develop the required technical landscape for DQ management

processes, standards and policies. Database experts and statisticians are focussed on computational solutions including effective and efficient IT tools, and computational techniques required to meet data quality objectives. These computational techniques include record linkage, lineage, and provenance, data uncertainty, semantic integrity constraints, as well as information trust and source credibility.

The above studies provide various lenses through which the body of knowledge can be classified and thus provide a means to study the depth and breadth of the body of knowledge. While this body of knowledge continues to expand over the last two decades, how this knowledge is successfully used in practice to achieve quality data is a question of concern.

1.1.2 Data quality practice

English (2009) reveals that over the past two decades 122 organizations have lost nearly one and a quarter trillion dollars due to poor quality data and he further emphasizes the gravity of economic disadvantage that poor quality data brings into modern organizations. An industry research by Gartner (Friedman and Smith 2011) reveals that 40% of the anticipated value of all business initiatives is never achieved due to poor quality data used in both the planning and execution phases of these initiatives. Further an industry survey by Pierce et al. (2012) reveals that only 22% of the organizations claim that the effectiveness of their data quality initiatives is 'Good' and they have achieved most of the intended goals while the rest 78% claims they have either achieved some of the goals or just a few of them. Thus, based on the above studies it is apparent that organizations are still struggling with achieving good quality data.

To explore the effectiveness of using the DQ body of knowledge in industry practice, we conducted an empirical study (Jayawardene et al. 2012; Sadiq et al. 2011a). This study investigated seven factors extracted from Sadiq et al. (2011b), through a survey with 60 DQ professionals working in government and private organizations. The survey responses indicated that three of the factors viz. data quality assessment, data quality frameworks, and data constraints and rules contribute most significantly towards the achievement of good data quality within organizations (Jayawardene et al. 2012). Data quality assessment involves measuring data quality, a fundamental aspect of management, owing to the popular management axiom *"what gets measured gets managed*" (Willcocks and Lester 1996). A data quality framework explains the overall strategy of data quality management starting from data quality goals/objectives and outlining a systematic process to achieve them. Therefore it should be noted that the two factors, DQ assessment and DQ framework focus on the management perspective necessary to achieve the quality of data. Rules and constraints refer to

the operational aspect of ensuring data quality, which can vary from database constraints to business rules, including all forms of specifications that affect organizational data.

In general terms, these three factors specify *'what to manage'* and *'how to manage'* in achieving DQ and, in turn indicate the fundamental importance of having a specification of DQ requirements and how the requirements can be implemented. Such a specification of DQ requirements is vital in successfully developing solutions to solve the problem of data quality. Requirements analysis and modelling is a fundamental aspect of developing solutions in the information systems domain (Machado et al. 2005). Machado et al. (2005) emphasise that a requirement model is essential to explain and share the views in human mind about complex problems, in designing and implementing successful solutions. In IEEE standards, a requirement is defined as a condition or capability needed to solve a problem or achieve an objective (IEEE 1990). Therefore, in DQ management, a DQ requirement is an essential prelude in designing solutions to solve DQ problems and achieving organizational DQ objectives.

According to our study (Jayawardene et al. 2012), this fundamental aspect is satisfactorily performed by only 30% of the organizations while the remaining 70% are struggling to effectively manage their DQ requirements (Jayawardene et al. 2012). Therefore, we posit that it is essential to model the data quality requirements in such a way that it brings together the management and operational aspects of data quality.

1.1.3 Limitations in DQ requirements analysis and modelling

DQ is a well-explored area and, in the literature prominent methodologies exist to manage DQ in organizational context e.g.:- (English 1999; Jeusfeld et al. 1998; Lee et al. 2002; Loshin 2004; Su and Jin 2006; Wang 1998). After considering thirteen DQ methodologies we observed that only five of them perform DQ requirement analysis and identify the DQ requirements in designing their solutions (see chapter 2 for more details). In the field of information systems, requirement analysis is a well-established concept in developing software solutions to organizational problems and, Bourque et al. (1999) point out that a sound requirements analysis is critical to the success of a software solution. Thus we further emphasize that the lack of DQ requirements analysis affects the success of DQ solutions developed in organizational context.

While DQ requirements analysis is overlooked by many DQ methodologies, when it comes to DQ requirements modelling we observed that there are a few attempts in the literature to model DQ requirements e.g.:- (Becker et al. 2007; Pierce 2002; Scannapieco et al. 2002; Serrano et al. 2009; Shankaranarayanan and Wang 2007; Shankaranarayanan et al. 2000; Storey and Wang 2001; Tu and

Wang 1993; Wang and V. 1998). In these approaches to modelling DQ requirements, there is no prominent approach which has unified the domain in the same way that Unified Modelling Language (UML) (Booch et al. 2000) has unified the software engineering domain or Entity Relationship Modelling (ERM) (Chen 1976) has unified the conceptual data modelling domain. These existing approaches for DQ requirement modelling have limitations such as deficiencies in domain concept coverage and notational complexities that distant them from practical usage (see Chapter 2 for more details). Thus, a necessity exists for a DQ requirement modelling framework that can be practically used to support DQ management to achieve the quality of data.

Since data quality requirements continue to be dictated by the "fitness for use principle" (Juran 2011), they are highly dependent on the organizational context. Due to the contextual nature, a plethora of diversified DQ requirements exist, they are complicated to model and analyse, and their re-use is extremely difficult, resulting in a prohibitive capacity for knowledge sharing for data quality requirements management.Thus an authentic and carefully developed set of generic data quality requirements that represents the universe of contextual DQ requirements can provide a valuable starting point for organizations embarking upon data quality initiatives, and for broader knowledge sharing within the data quality research and practitioner community.

In requirements modelling, pattern-based approaches can be successfully used due to the modelling power the patterns provides through design options in complex and overwhelming scenarios (Sutcliffe et al. 1998). Hoffmann et al. (2012) developed twenty software requirement patterns and the pre-defined requirement templates that can be used to specify trust requirements in recommender system development. In general, pattern-based approaches have already established well within the field of information systems to facilitate complex designing processes. For instance, Van Der Aalst et al. (2003) have developed workflow patterns to facilitate business process design. Further, Gamma et al. (1995) have defined design patterns to facilitate software design. Data quality requirements are complex by nature due to the cross-disciplinary nature of the domain (Sadiq et al. 2011c). Due to the complexity of DQ requirements, we posit that a requirements modelling approach that uses DQ requirement pattern will best suit the purpose of modelling DQ requirements

1.2 Aims and objectives of the research

The aim of this study is to develop a repository of data quality patterns to model data quality requirements in an organization which can be used to drive the data quality management function systematically. Thus, the research has the following five objectives:

1) Investigate and identify the requisite concepts to develop an abstract model of a DQ requirement
2) Develop a repository of reusable DQ patterns to model the DQ requirements in an organization.
3) Evaluate the DQ patterns for clarity and coverage to ensure that the patterns are sufficiently differentiated and provide sufficient coverage to represent real world data quality requirements.
4) Develop requirements modelling methodology to facilitate the use of the patterns to model data quality requirements in an organization.
5) Evaluate the applicability of the pattern-based approach in real world organizations by using the methodology developed in objective 4.

1.3 Thesis structure

This rest of this thesis is organised into nine chapters, as follows:

Chapter 2 presents a review of the literature and related work relevant to these research objectives. In particular, we discuss leading data quality management frameworks, conceptual modelling, and requirements engineering, and conclude the chapter with a discussion on the current research gap found in the literature.

Chapter 3 describes the methodology that governs the development of a modelling framework for DQ requirements. This discussion includes details of the approach that has been utilised to create and evaluate the artefacts produced in this research. In particular, we discuss design science as the principal methodology for this research.

Chapter 4 discusses the concepts behind DQ requirements and systematically defines a meta-meta-model for a DQ requirement (Objective 1).

Chapter 5 presents the concept of data quality dimensions, fundamental in defining a data quality requirement, and discusses the multiple interpretations of them available in the literature. Then we refactor this concept by assimilating it with other related concepts in the domain so that it can be used to represent a comprehensive data quality requirement. Finally, we produce a consistent set of eight data quality dimensions and thirty-three data quality characteristics related to these dimensions (Objective 1 and 2).

Chapter 6 documents the repository of data quality patterns by instantiating the meta-model developed in Chapter 5 using the meta-meta-model introduced in Chapter 4. (Objectives 2).

Chapter 7 presents a descriptive evaluation of the data quality patterns. The evaluation is performed using eight credible data quality problem classifications and three real world data quality rule repositories. We further identify and investigate relationships between patterns (Objective 3).

Chapter 8 presents a methodology for data quality requirement engineering, i.e. how the patterns can be used to model data quality requirements of an organization (Objective 4).

Chapter 9 evaluates the applicability of the pattern-based approach for modelling data quality requirements in organizations. We conduct empirical evaluations in two organizations to identify their data quality requirements using the pattern-based approach developed in this thesis (Objective 5).

Chapter 10 presents the conclusion and future work relating to this research.

Chapter 2

LITERATURE REVIEW

2.1 Overview

To achieve an in-depth understanding of data quality management principles and concepts, a review of the data quality body of knowledge was conducted. In this chapter, first, we present the fundamental concepts of data quality, the evolution of DQ management and the methodologies developed over the years to manage DQ. Second, we present fundamental concepts of requirements modelling, the composition of a meta-model and the notion of the pattern-based approach used in modelling. Finally, we point out a research gap in the body of knowledge with regards to analysis and modelling of DQ requirements in managing data quality.

2.2 Data Quality

What is data and what is data quality? Liebenau and Backhouse (1990) used modern semiotic theory principles developed by Morris (1938) to explain data. In their work, they define data as "…language, mathematical or other symbolic surrogates which are generally agreed upon to represent people, objects, events, and concepts". As per this definition, in its simplest form, data is a representation of objects or phenomena in the real world. Thus, when it comes to the discussion of the quality of data, we can say good quality data is a result of a good representation of the real world. In the context of information systems, this representation of a real world is further moderated by the needs of the system users, hence the reference framework to evaluate the representation is the set of user needs – i.e. the same object in the real world may have different representations in an information system depending on the need of the users. This semiotic perspective of data has been adopted by DQ researchers as well. For example, Price and Shanks (2004) define three quality levels for data, i.e. syntactic quality, semantic quality and pragmatic quality.

The application of semiotics can be considered as one of the philosophical approaches towards data and its quality. To date, however, the semiotic perspective of data has not become popular among researchers or practitioners. When it comes to supporting processes for managing data quality, a prominent approach, proposed by Wang (1998), uses a product perspective of data. Owing to the historical definition that 'information is processed data' Wang and Strong (1996) argue that information is analogous to products and data is analogous to raw materials in a typical product manufacturing process. Based on this argument, Wang (1998) considers information as a product of an information system and recognizes an information manufacturing process as analogous to a

product manufacturing process. Since traditional product quality is a well-explored concept, researchers have attempted to use proactive product quality management models claiming 'fitness for use' as the principle for recognising good quality data and poor quality data. The 'fitness for use' approach is based on the general definition of quality introduced by Juran (1962). In the case of products, fitness for use is evaluated with reference to product specification, which contains customer expectations expressed in terms of different orthogonal dimensions. Similarly, Wang and Strong (1996) defined quality dimensions for data in such a way that they represent customer expectations in using data

ISO 8000 (ISO 2012), which is the emerging standard for data quality, defines quality as "degree to which a set of inherent characteristics fulfils requirements". The definition emphasises that data should adhere to the characteristics that fulfil user requirements. This definition compliments the fitness for use principle while specifically emphasising as the characteristics of data that make them fit for satisfying user requirements.

In literature, authors have used the two terms "data" and "Information" interchangeably, and the following definitions can be seen for information quality.

"Information quality is the characteristic of information to meet or exceed customer expectations" *(Kahn and Strong 1998)*

"Information quality is the characteristic of information to be of high value to its users" (Lesca and Lesca 1995)

"The degree to which information has content, form, and time characteristics which give it value to specific end users" (O'Brien 1990)

"Information quality is the characteristic of information to meet the functional, technical, cognitive, and aesthetic requirements of information producers, administrators, consumers, and experts" (Eppler 1999)

Therefore, it is clear that data quality is a subjective concept that depends on data user's requirements and expectations. In the next section, we explain the foundational concepts governing DQ and how they are used in DQ management.

2.2.1 Foundation Concepts

Wang and Strong (1996) indicate that data quality is a multidimensional concept in which the consumers of data expect data to adhere to a number of characteristics, which they refer to as DQ dimensions. They defined 15 data quality dimensions classified under four main categories viz. (1)

Intrinsic data quality (Believability, Accuracy, Objectivity, Reputation), (2) Contextual data quality (Value added, Relevancy, Timeliness, Completeness, Appropriate amount of data), (3) Representational data quality (Interpretability, Ease of understanding, Representational consistency, concise representation) and (4) Accessibility data quality (Accessibility, Access security).

Over the years, DQ researchers and DQ practitioners have defined DQ dimensions as a foundational concept in DQ management. DQ dimensions are various perspectives that data consumers expect to make data fit for use (Wang and Strong 1996) and DQ literature offer different classifications of dimensions. Batini et al. (2009) indicate that there are a number of discrepancies in the definitions of most DQ dimensions in the existing classifications, due to the contextual nature of DQ. A more detailed discussion about DQ dimensions is provided in Chapters 4 and 5 of this thesis.

Based on the ISO definition for DQ as "a set of inherent characteristics fulfils requirements", the notion of DQ dimensions can be considered as the characteristics of data that fulfils user requirements. Therefore, primarily, DQ dimensions can be considered as fundamental user requirements of data. Lack of adherence to DQ dimensions prohibits fulfilment of user requirements and thus creates a DQ problem. Therefore, DQ dimensions are a central concept to both DQ requirements and DQ problems. In other words, DQ requirements and DQ problems are two sides of the same coin.

Due to the overwhelming nature of user requirements, data quality problems are complicated. Therefore, over the last few decades, many researchers have contributed to the identification of data quality problems and generalizing them into DQ problem categories to help with their management.

Lesca and Lesca (1995) identified five categories of data quality problems, limited usefulness, ambiguity, incompleteness, inconsistency and inadequate presentation format. Whereas Garvin (1988) pointed out three types of problems in data and information biased information, outdated information, and massaged information. Biased information means the content of the information is inaccurate or distorted in the process of transformation. Outdated information is information that is not sufficiently up to date for the task. Massaged information refers to different representations of the same information so that it facilitates different interpretations.

Eppler (2006) identified 11 categories of data quality problems i.e duplicates, missing data relationships, garbling (meaningless entries), spelling errors, obsolete or outdated entries, inconsistent data formats or naming conventions, misplaced data that is saved in the wrong database, complicated query procedures, wrong data coding or tagging, Incorrect data entries because of lack of source validation, manipulation of stored data. Further Kim et al. (2003) developed a taxonomy for dirty

data, which is a logical classification of a wide range of data quality problems. Their classification resulted in thirty-two categories of DQ problems.

It should be noted that all the efforts above have focussed on surface level problems reported as data quality problems from the end-user point of view and, therefore, the authors have not used the concept of data quality dimensions in characterising these problem categories. Thus, we can see a clear disconnect between the DQ dimensions literature and DQ problems literature.

In the following section, we discuss the evolution of DQ management to understand the philosophical notions towards managing DQ and thus understand what DQ management is in the current context.

2.2.2 Evolution of DQ management

Pierce et al. (2013) explained the historical evolvement of data quality management, indicating that the information and data quality discipline has had a relatively short, but rapidly evolving, history that can be identified as having five phases:

1) Problem Recognition: The Data Cleansing Phase
2) Root Cause Detection: The Prevention Phase
3) Manufacturing Analog: The Information Product and Process Management Phase
4) Information Architecture: The Quality by Design Phase
5) Enterprise View: Information as an Organizational Asset Phase

Initially, organizations followed a reactive approach for data quality management by recognising bad data and cleansing them (Inmon 1993) with the emergence of data warehousing. Data cleansing projects are still popular in DQ management in the context of data warehousing (Kimball et al. 1998). The treatment for bad data was costly for many organizations and gradually practitioners moved towards detection of root causes for bad data with the view of prevention. Best practices in Total Quality Management (TQM) like root cause analysis, were adapted to stop the same DQ problem re-occurring. As this phase the main focus of data quality projects was to prevent future data errors and correction of existing errors (data cleansing) was a second priority (English 1999).

In the third phase, practitioners adapted the entire TQM paradigm to the data and information context. They applied manufacturing concepts to information systems and to the entire information generation process (Redman 1997; Wang et al. 1998). The approach, developed by Wang (1998), followed the viewpoint that information is the product of an information system and not a by-product. By viewing data and their sources as raw materials in a warehouse, the software applications as the manufacturing process, and the final outputs (processed data) as the products, then the full range of TQM principles

could be applied to information systems. This evolution resulted in the formulation of the Total Data Quality Management (TDQM) process (Huang et al. 1998; Lee et al. 2002; Wang 1998). Similarly, approaches developed by English (1999) and Redman (1997) focused on defining, managing, and improving the organizational processes through which data is created, captured, stored, delivered, used, and retired. These approaches were ideal examples for practitioner-driven TDQM approaches at this stage.

The most important development of applying the disciplines of product and process management to DQ is that they brought into consideration the users (customers) of data and information (Wang and Strong 1996). Therefore, DQ was treated with a broader perspective with the introduction of a broader range of quality dimensions relating to usage of data, such as usefulness and value of information. As a result of this third phase in DQ management, a number of data quality frameworks came into existence with a specific focus on TQM principles. Some notable approaches are given in Table 2.1.

The next phase of DQ management began with efforts to deliver more proactive solutions for data quality management by incorporating DQ requirements in data models and in the information architecture. This choice is backed on a well-known principle of software development where the earlier in the development process that a problem is discovered, the less effort is required to correct it. Similarly, in DQ management, information systems professionals such as business analysts, solution architects, and database experts began to consider DQ requirements in the process of designing the organizational technical landscape (Loshin 2001).

In the fifth phase and current phase, there is growing need to consider data and information as an organizational asset. Therefore, DQ principles and practices are becoming a critical part of the organizational strategy (Redman 2008) and quality data is considered as a potential revenue generation asset rather than a cost centre for organizations.

Through this evolution in DQ management, over the years, many researchers and industry practitioners have developed a number of approaches\methodologies for DQ management. In the following section, we present an overview.

2.2.3 Methodologies for data quality management

Over the past three decades through the evolution of DQ management, the DQ research community, as well as practitioner communities, have developed approaches and methodologies to manage DQ. In this section, we discussed the prominent approaches along with their key elements and steps.

Total data quality management (TDQM) proposed by Wang (1998) is focused on four steps required to ensure quality of data, viz (1) clearly articulate the information product (IP) in business terms; (2) establish an IP team consisting of a senior executive as the TDQM champion, an IP engineer who is familiar with the TDQM methodology, and members who are information suppliers, manufacturers, consumers, and IP managers; (3) teach IQ assessment and IQ management skills to all the IP constituencies; and (4) institutionalize continuous IP improvement. The main focus of this methodology is to recognize the IPs and IP manufacturing process in the organization and build management capabilities around the process to prevent bad quality data\information.

Data warehouse quality (DWQ) proposed by Jeusfeld et al. (1998) focused on eight main concepts in formulating a strategy for DQ in a data warehouse, namely measurable objects in data warehouse, quality goal, a quality query to check quality, quality dimension, quality measurement, a metric unit, quality domain and quality range. The methodology is focussed on developing a meta- system using the concepts to measure and monitor DQ in a data warehouse. It uses 12 DQ dimensions to define quality goals and metrics to measure the achievement of goals. The method is a reactive one because data cleansing is the main strategy of managing DQ.

Total information quality management (TIQM) is another prominent methodology, which was developed by English (1999) for data warehouses and later enhanced to overall organizational DQ management (English 2009). This methodology emphasises six actions that should be enacted by an organization to achieve quality data. They are: (1) Implementing and sustaining an effective information quality environment, (2) Assessing information product specification and information architecture quality, (3) Assessing information quality, (4) Measuring the costs and risks of poor quality information, (5) Improving information process and (6) Data correction and controlling data redundancy. This methodology emphasises the systematic development of a case for data quality that rationalizes the requirement for having good quality data. It is a process driven methodology in which the six aspects provide guidance for a holistic process for DQ management.

AMIQ (a methodology for information quality management) was developed by Lee et al. (2002) and includes a model of DQ dimensions, a questionnaire to measure organizational DQ, and analysis techniques for interpreting the DQ measures. The main focus of the methodology is to analyse the quality of existing data and, thereby, find the gaps with regards to a confirmed specification for DQ. The methodology guides the identification of the gaps in roles and processes that cause poor quality data. The results of the techniques are used for determining the best area for IQ improvement activities.

DAQ (Data Quality Assessment) by Pipino et al. (2002) is focussed on three main steps viz. (1) performing subjective and objective data quality assessments, (2) comparing the results of the assessments, identifying discrepancies, and determining root causes of discrepancies, and (3) determining and taking necessary actions for improvement. Pipino et al. (2002) suggested that it is necessary to have both subjective (user perception oriented) and objective (database oriented) quality measurements to manage data quality. The main focus in the first step is to determine which data should be considered for DQ management since all organizational datasets may contain quality problems to some extent. The approach used DQ dimensions developed by Wang and Strong (1996) to determine the metrics for data quality A quadrant of subjective vs objective measurement of quality of a dataset is used to determine whether to invest in quality improvement initiatives. The methodology is focused on eliminating the root causes of DQ problems and to systematically reduce DQ errors, it is, therefore, a proactive methodology.

IQM (Information Quality Management) by Eppler and Muenzenmayer (2002) is a four-step methodology to ensure DQ in the internet and web context. The four steps include (1) Measurement planning, (2) Measurement configuration, (3) Measurement and (4) Follow-up activities. Measurement planning refers to the identification of relevant information quality criteria (dimensions) through interviews with stakeholders and defining qualitative and quantitative indicators for the criteria. Measurement configuration refers to the weighting of indicators according to strategic priorities and setting up alerts (dashboards) for indicators. Then measurements are done through monitoring data instances (using tools) and conducting surveys among the users. Finally, follow-up activities are performed by cleansing data and assigning responsibilities to continuously perform the four steps depending on the outcome of the current iteration of the steps. This methodology is a reactive approach for DQ because of its prime focuses on identifying and fixing problems in data.

ISTAT is a DQ approach developed by the Italian National Bureau of Census (ISAT 2004) and is designed for a distributed structure of organizations that have strong inter-organizational informational dependencies. In Italy, public administration is organized in three geographical levels, Central, Regional and Peripheral, each managing its own data autonomously. The ISTAT methodology focuses on the most common types of data exchanged among the different levels of the public administration. The methodology consists of four steps: (1) The assessment phase (focused on the quality issues from integration point of view in the central databases owned and managed by ISTAT), (2) The global improvement phase (focused on record linkage problems among national databases and designing solution on processes including the decisions to make, buy, or adapt existing

solutions for data linkage), (3) Internal DQ Improvement activities on databases owned and managed by local administrations (DQ assessment using statistical tools, standardization of data collection and transformation processes, Implementation of DQ solutions) and (4) Inter administrative activities (standardization of information exchange formats among organizations, organizational process redesign and implementation to support DQ).

The AMEQ (A Methodology for Information Quality) developed by Su and Jin (2006) is designed to assure the quality of product information in product manufacturing environments. The methodology consists of five phases for measuring and improving product information quality (PIQ) in a systematic way throughout the manufacturing process. Phase1- Establish IQ environment- assesses the cultural readiness of an organization, using the Information Quality Management Maturity Grid, (a template to conduct interviews for key managerial roles). Then the quality dimensions of product information are also defined according to their relevance for different business operations. Phase2-Define PIQ- specifies the information product where each information product is associated with a corresponding business process, modelled by means of an object-oriented approach (OOA). In the AMEQ methodology, eight types of objects are modelled (human resources, information resources, enterprise activities, resource inputs, resource processes, resource outputs, performance measures, and enterprise goals). In this phase, a model of measurement methods is also produced. Phase3-Measure PIQ- focuses on the measurement activity. Phase4-Analyze PIQ- investigates the root causes for PIQ problems by analysing the quality dimensions that have received a low score. Finally, Phase5-Improve PIQ- introducing both technical solutions and business processes to sustain long-term improvement focus of PIQ.

COLDQ (Cost-Effect Of Low Data Quality) by Loshin (2004) provides a classification of operational, tactical, and strategic economic impacts that had to be considered to manage data as an organizational asset. The COLDQ methodology provides a data quality scorecard that enables avoidance of poor quality costs due to the adoption of improvement techniques. Loshin (2010) further improved the methodology with detailed guidelines for practitioners to develop a data quality program in their organizations. The methodology consist of five cycles: (1) Asses: Data quality assessment to identify and measure how poor data quality impedes business objectives (Business case for DQ) (2) Define: Definition of business related data quality requirements (rules and performance targets) through a comprehensive DQ requirements analysis (3) Design: Design of quality improvement methods and processes (4) Deploy: Implementation of quality improvement methods and processes (5) Monitor: Monitor data quality against defined targets. The methodology includes guidelines for each cycle.

DaQuinCIS (Data Quality IN Cooperative Information Systems) by Scannapieco et al. (2004) presents a methodology for data quality that is focussed on a systematic process of (1) Quality analysis where the requirements for DQ is established, (2) quality assessment where the current DQ is measured, (3) quality certifications to certify good quality data and (4) quality improvement solutions. Further, it provides constructs to represent data, a list of DQ properties and constructs to represent them and the associations between data and quality metadata.

QAFD (Quality Assessment of Financial Data) by De Amicis and Batini (2004) is specific for DQ in the financial context through a systematic process of five phases: (1) Variable selection where the most relevant financial variables are selected, (2) Analysis of quality requirements for the variables, (3) objectively measurement of DQ, (4) Subjective measurement of quality and (5) Comparison of objective and subjective measurements and propose improvements.

CDQ (Complete Data Quality) (Batini et al. 2008; Batini and Scannapieco 2006) consists of three main phases: (1) State reconstruction: Identification of organizational units roles and processes of data usage and predation of descriptions of data flows and processes (2) Assessment: Identification of DQ requirements and measurement of quality and (3) Improvement: Identification of root causes for errors and development of data driven and process-driven solutions.

Batini et al. (2009) indicate that most of the methodologies in literature have their own DQ dimensions, but not every methodology has a systematic way of analysing DQ requirements (Table 1). This is a limitation of the methodologies since DQ requirements are fundamental to the process of DQ management by specifying what to manage. Thus, a lack of understanding of DQ requirements precludes adequate management of DQ in organizations.

Methodology	DQ Dimensions defined	DQ Requirement analysis performed	Measurement of quality	DQ improvement strategies
TDQM	Yes	No	Yes	Yes
DWQ	Yes	Yes	Yes	Yes
TIQM	Yes	Yes	Yes	Yes
AIMQ	Yes	No	Yes	Yes
CIHI	Yes	No	Yes	Yes
DQA	Yes	No	Yes	Yes

IQM	Yes	No	Yes	Yes
ISTAT	Yes	No	Yes	Yes
AMEQ	Yes	No	Yes	Yes
COLDQ	Yes	Yes	Yes	Yes
DaQuinCIS	Yes	No	Yes	Yes
QAFD	Yes	Yes	Yes	Yes
CDQ	Yes	Yes	Yes	Yes

Table 2.1: Methodologies and DQ requirements analysis –adapted from (Batini et al., 2009)

As explained in section 2.2.1, the situations where the DQ requirements are not met becomes the DQ problems and, therefore, DQ requirements provide an insight into what needs to be managed to prevent DQ problems. In other words, every DQ requirement has to be recognised and a solution has to be developed to meet the requirements to achieve the quality of data.

In the field of information systems requirement analysis and modelling is a well-established concept in developing software solutions to address organizational problems. Bourque et al. (1999) point out that requirements analysis is critical to the success of a software solution. Machado et al. (2005) emphasise that a requirement model is essential to explain and share the views in human mind about complex problems, in designing and implementing successful solutions. Thus, we posit that a comprehensive requirement model is a key to developing solutions for DQ problems. Therefore, in the next section, we investigate into the concepts behind requirements modelling with a view to understanding what is meant by a DQ requirement model.

2.3 Requirements modelling

Requirements models are used in domains like software development (Booch et al. 2000), database design (Chen 1976) and business process management (Omg 2008) to systematically analyse the requirements of end users in successfully designing the solutions. In this section, we present a synopsis of requirements modelling literature relevant to our research.

2.3.1 Fundamentals of requirements modelling

In information systems and computer science domains, conceptual models are often developed using modelling grammars, for example, UML ERM etc. Conceptual modelling is the process of describing

some aspects of the physical and social world around us for the purpose of understanding and communication (Mylopoulos 1992). The term conceptual model may be used to refer to the representation of real world abstractions in the form of an artefact which is formed after a conceptualization process in the mind. Kung and Soelvberg (1986) identified four objectives of developing a model in IS domain.

1. Enhance an individual's understanding of the representative system
2. Facilitate efficient conveyance of system details between stakeholders
3. Provide a point of reference for system designers to extract system specifications
4. Document the system for future reference and provide a means for collaboration

In IS research, the process of conceptual modelling has been examined from different perspectives. Insights about the underlying knowledge for developing conceptual models in the context of IS can be found in (Mylopoulos 1992). Mylopoulos (1992) suggests that there are four types of knowledge related to Information Systems development viz. (1) Subject world (which is the represented domain), (2) Usage world (which is the environment where the system is used), (3) Development world (which is the process and environment where the system is developed), and (4) System world (which is the information system itself).

Wand et al. (1995) describe the role of a conceptual model within the view of systems development, consisting of analysis, design, and implementation by relating to the knowledge categories explained above. *Analysis* transforms a perceived real-world system into a conceptual model of that system, (i.e. creation of models of the *subject* and *usage worlds*). *Design* transforms the conceptual model of the *subject and usage world* into a model of the *information system world*. In particular, the system interfaces are designed based on the *usage world*. Finally, *implementation* transforms the model of the *information system world* into an implemented information system, which is a machine-executable representation of *system world.*

Burton-Jones (2014) points out that the conceptual modelling techniques that have been proposed over the years tend to fall into one of two categories. The first category is focussed on providing constructs to model *substance and form* in the real world which is termed as *data modelling technique* or *semantic modelling technique*. Whereas the second category is focussed on constructs to model *possibility and change* in the real world which are called *process modelling techniques*.

Stachowiak (1973) points out that a model needs to possess three features. Mapping feature: A model is based on an original; Reduction feature: A model only reflects a relevant selection of an original's properties; and Pragmatic feature: A model needs to be usable in place of an original with respect to

some purpose. The first two features explain that a model is a relevant projection of original phenomena of a system, a process etc. whereas the third feature explains the usability perspective of the model.

As per all of the above viewpoints, it is apparent that the researchers have emphasized the importance of several orthogonal perspectives related to conceptual models. In summary, they can be classified as (1) The purpose and the role of the model (2) The content of the model in terms of the embedded knowledge in it.

Further, it should be noted the knowledge embedded in a model is derived from the domain of focus (explicit/formal knowledge about the subject domain) as well as some form of tacit knowledge demanded by the task at hand. For example, in the context of DQ, a model of DQ requirement should include explicit/formal knowledge about data quality and some tacit knowledge of the users relative to how they use the knowledge to accomplish the task at hand.

2.3.2 What is a Meta-Model?

Conceptual modelling languages like ERM, UML, BPMN etc. are based on meta-models. Even though there are different definitions explaining what a Meta-Model is, in general, a Meta-Model is a model that defines the constructs for expressing a model (OMG 2004). The purpose of a model is to represent some knowledge in such a way that the stakeholders can understand it and use it for a task at hand. In Information Systems and Computer Science domains, the stakeholders range from system users, system analysts, designers, software engineers etc. and each category of stakeholders may have their own terms and concepts in representing knowledge. For example, a system user would like to refer to employees as "employee data" whereas a software engineer would like to refer to it using "employee class". Meanwhile, a business analyst would like to refer to it as a business artefact called "employee". When there are more diverse groups involved in sharing and translating more diverse varieties of information, the problems standing in the way of representing this knowledge together within a single system increase orthogonally. Even within the same category of users, there can be discrepancies of understanding due to lack of uniformity in representations. For example, different databases may use identical labels but with different meanings; alternatively, the same meaning may be expressed via different names.

Hence, due to the diversity of the IS/CS domains, a Tower of Babel problem has arisen when it comes to knowledge representations and this situation has paved the way for many formalisations (Smith 2008). Meta-Modelling is one such formalization, which is focussed on creating consensus when creating models in a domain.

There are many definitions for Meta-Models. For example:

"Meta-Model is a Model of models" (Booch et al. 2000).

"*A model is an instance of a Meta-Model"* (OMG 2004).

"*A Meta-Model is a model that defines the language for expressing a model*" (Booch et al. 2000).

"*A Meta-Model is a model of a language of models"* (Favre 2004).

"*A Meta-Model is a specification model for which the systems under study being specified are models in a certain modelling language"* (Seidewitz 2003).

It is clear that a Meta-Model is used in the process modelling to reach some form of uniformity in the models developed in a particular domain, which facilitates shared understanding among stakeholders. Modelling languages are based on a particular Meta-Model that defines their constructs which are then used to create models.

The object Management Group (OMG) (Omg 2008) has defined a standard for model-driven engineering called Meta Object Facility. This standard specifies a four-layered architecture for modelling of real world abstracts (see Figure 2.1). It provides Meta-Meta-Model at the top layer (called the M3 layer) and this is the language used to build Meta-Models called M2 layer. M2 is the language used to build M1 models which are the representation of real world abstractions. M0 layer is the real system. As per this specification, a modelling language (eg: UML, BPMN) is a representation of the Meta-Model at M2 Layer.

As per this infrastructure, a Meta-Model can be considered as an instance of a Meta-Meta-Model and model can be considered as an instance of a Meta-Model (as per the above definition is given in (OMG 2004)). Kühne (2006) have investigated, under which circumstances a model might be granted "Meta-Model" status and found that still there is a lack of consensus about Meta-Models among the authors. Hence, what exactly is a Meta-Model is not yet a question with a clear answer. Thus, we use the OMG's Meta-Modelling infrastructure to understand the concept, owing to its capacity in systematically specifying the well-known modelling languages like UML and BPMN (Omg 2008). Therefore, we consider the meta-meta-model as the starting point of conceptualization and the conceptualization should be done sequentially by instantiating the immediate layer above.

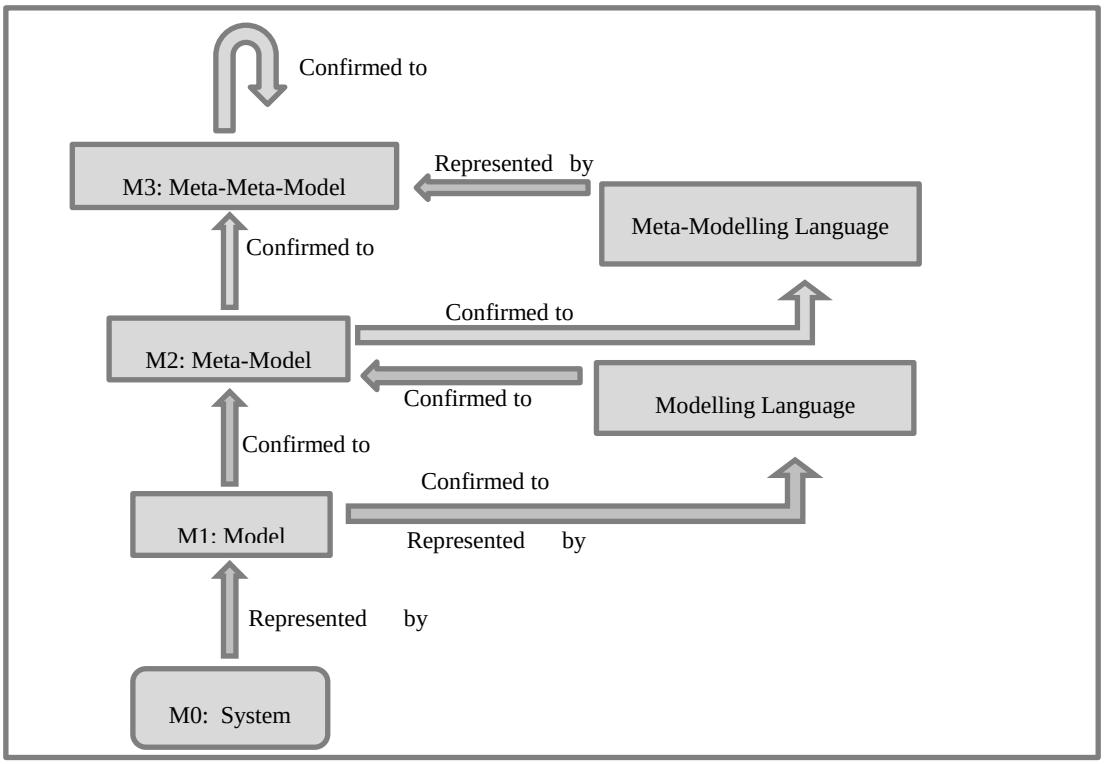


Figure 2.1: OMG's Meta-Modelling infrastructure

2.3.3 Building blocks of Meta-Modelling

Guizzardi (2007) emphasized that the process of conceptualization of a domain begins with recognizing the domain concepts, followed by creating constructs to formulate a Meta-Model of a modelling grammar for the domain.

Olivé (2007) summarised over 220 academic publications on conceptual modelling spanning across over half a century and found fourteen aspects that are useful in understanding conceptual modelling in the IS domain. Olivé (2007) has categorised the aspects of modelling into two categories "Structural Modelling" and "Behavioural Modelling" where the domain concepts and their relationships belong to structural modelling and domain events and related aspects belong to behavioural modelling.

Further, Jackson (2009) has presented some insight into the different types of models (Analytical models and Analogical models) and emphasised what need to be considered in creating models, like domain concepts and relationships, and the roles of these relationships.

From the related literature, it is clear that domain concepts are the building blocks of conceptualization. Thus, the domain concepts are coupled with a set of axioms to create constructs of a Meta-Meta-Model which is the starting point of conceptualization as per (Omg 2008). The

axioms in this context are called the production rules as per Burton-Jones et al. (2009), or context conditions as per Guizzardi (2007), which are valid and suitable for expressing real world abstractions using the constructs. Rosemann and Green (2002) suggest that domain concepts can be used with production rules of a suitable existing modelling grammar (ER, BPMN etc.) in producing a meta-model and then the resultant meta-model will have more or less the same flavour of the modelling grammar. Therefore, a suitable grammar is required considering the nature of the domain so that once the domain concepts are projected on the grammar, it will result in a meta-model that can represent the domain (Rosemann and Green 2002; Rosemann and Zur Muehlen 1998; Scheer 2009).

Another important aspect of conceptual modelling is design patterns. Design patterns are reusable design scenarios and help to reduce the complexity of the process of developing conceptual models (Riehle and Züllighoven 1996). Design patterns have been successfully used in software engineering (Gamma et al. 1995) business process modelling (Van Der Aalst et al.2003) and many other disciplines. In the following section, we discuss the applicability of design patterns in modelling DQ requirements.

2.3.4 Use of patterns in requirements modelling

Recent studies have pointed out that analysing and defining system requirements is a time-consuming and error-prone process in the process of software development and thus it is useful to develop patterns of requirements (Franch 2013). Therefore, a catalogue of software requirements patterns can support the elicitation, validation, documentation and management of requirements. Hoffmann et al. (2012) developed twenty software requirement patterns and the pre-defined requirement templates that can be used to specify trust requirements in recommender system development projects.

Even though the notion of patterns in requirements modelling is still in its infancy, patterns have been defined and successfully applied in various disciplines. Therefore, in order to understand a rationale behind a pattern based approach, it is worthwhile investigating some of these pattern-based approaches. For example, in the context of building architecture, Christopher Alexander defined a pattern as "a three-part rule, which expresses a relation between a certain context, a problem, and a solution" (Alexander 1979). In the context of software development, Riehle and Zullighoven define a pattern as "the abstraction of a concrete form which keeps recurring in specific non-arbitrary contexts" (Riehle and Züllighoven 1996), whereas Gamma et al. (1995) define a pattern as "the solution to a recurring problem in a particular context". While multiple definitions exist, the definitions imply that, in general, *a pattern is a solution to a recurring problem*. In the context of DQ management, a pattern can be defined as a solution to a recurring DQ problem. Thus, we posit that

re-usable reusable DQ requirement patterns can be developed to support the elicitation, validation, documentation and management of DQ requirements.

2.4 Existing approaches for DQ requirements modelling

In section 2.2.3 we revealed that only five DQ methodologies have a requirement analysis phase while others have overlooked this task. Out of the five methodologies, none of them suggest an approach to model DQ requirements. Further in overall DQ literature, only a few attempts exist to model DQ requirements. In this section, we present them in brief and analyse their strengths and limitations.

The approach by Tu and Wang (1993), extended the Entity-Relationship model (ER) with quality characteristics. This work was further improved by Storey and Wang (2001) and Wang et al. (2000). These attempts can be considered as fundamental approaches to include quality concerns in the database schema. The extension was done by way of adding a new attribute for each entity to represent the quality of attributes (Figure 2.2). For instance, to express the DQ dimension "Completeness" for the attribute "Address" of entity "Person" a new attribute "AddressQualityDimension" is added to the entity "Person".

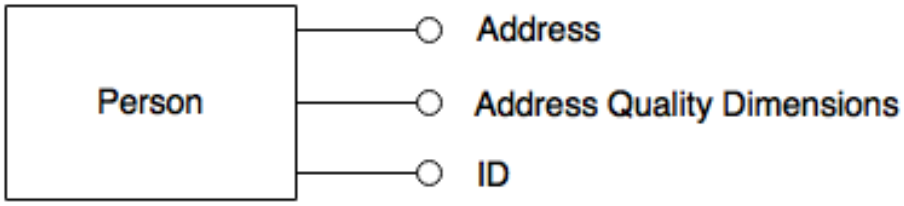


Figure 2.2: Example of quality dimension as an attribute of an entity

Further, the authors have suggested extending this approach through having two separate entities to represent DQ dimension and DQ measure for each entity in the ERD (Wang and Veda 1998). For example, the entity "Class" and attribute "Attendance" in an original data schema can be extended to a data quality schema by having two additional entities "DQ dimension" and "DQ measure" (Figure 2.3). The entity "DQ dimension" has a many to many relationships "ClassAttendense" with class-attendance attribute combination and for each such entity-attribute combinations a relationship is introduced with the corresponding "DQ measure" entity.

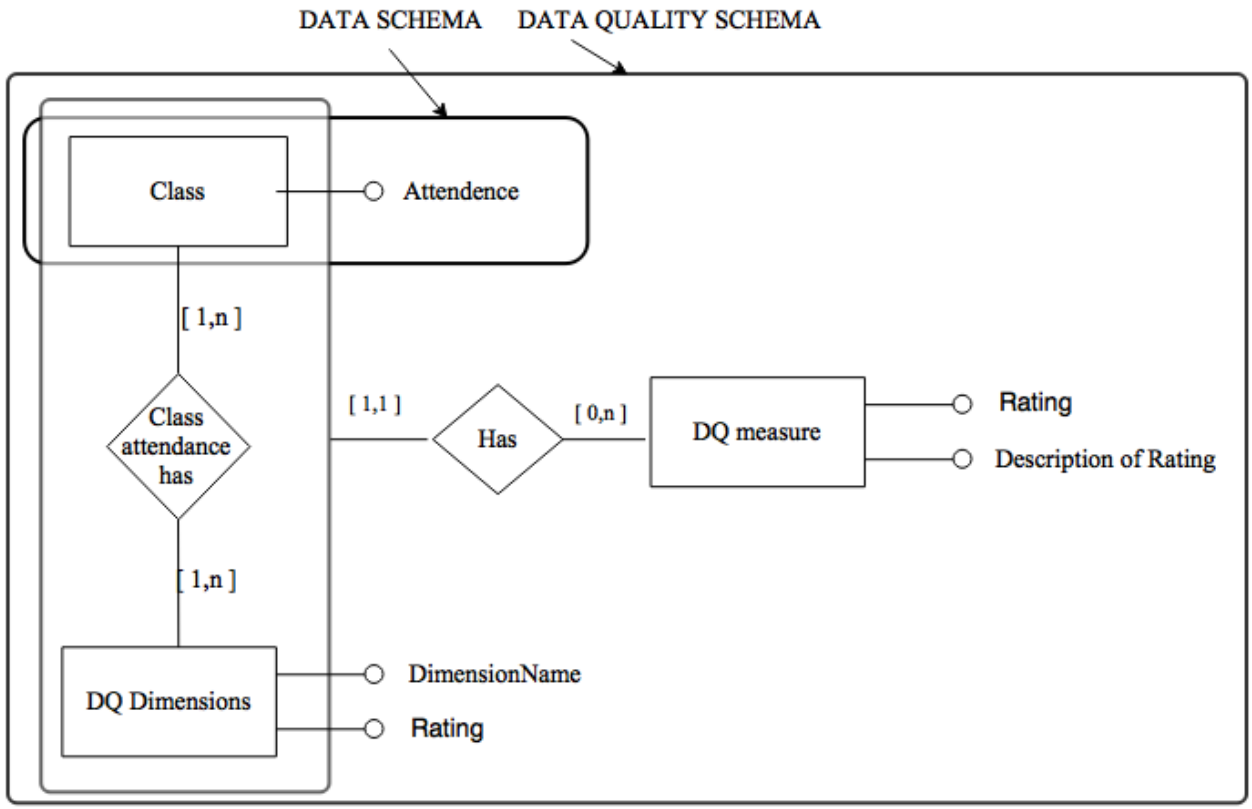


Figure 2.3: Example of DQ dimension as a separate entity in ER diagram.

The main drawback of this approach is that, at the ER modelling level, entities are not yet normalised and mapped to relational schema. Therefore, the diagram is distant from the logical database. Thus, it is hard to make use of this model for practical management purposes since the actual database is built on the normalized version of the ER. On the other hand, the majority of root causes for data quality problems are found at the Information Systems (IS) level (technical layer), Business Rules level and Business Process level (English 2009; Kim et al. 2003; Loshin 2011) and, hence DQ management functions are implemented at these operational levels rather than at the conceptual data modelling level. From a management point of view, it is not practically useful to formalize DQ requirements at a conceptual level (ER level), which is distant from the operational level.

The second notable attempt to model DQ requirements is IP-MAP (Information Product Map) (Shankaranarayanan and Wang 2007; Shankaranarayanan et al. 2000). The model considers the principle that data can be seen as a product of a manufacturing activity (Wang 1998). IP-MAP is a graphical Meta-Model (with its own constructs) designed to comprehensively describe how an information product (such as an invoice, customer order or prescription) is assembled in a business process (Figure 2.4). From a management perspective, IP-MAPs are designed to help analysts visualize the information production process, identify the ownership of process phases, understand

information and organizational boundaries and estimate and assess quality metrics associated with the production process. While this is a comprehensive meta-model, there is no evidence to indicate that it is popular among DQ practitioners. As per Caro et al. (2012), IP-MAPS did not become popular among DQ practitioners since they had to be maintained separately to other existing models like business process models or software conceptual models drawn using other meta-models (eg. BPMN. UML, EPC). Hence, IP-MAPS became an isolated view that could not be used simultaneously with other domains and ultimately became an obsolete practice.

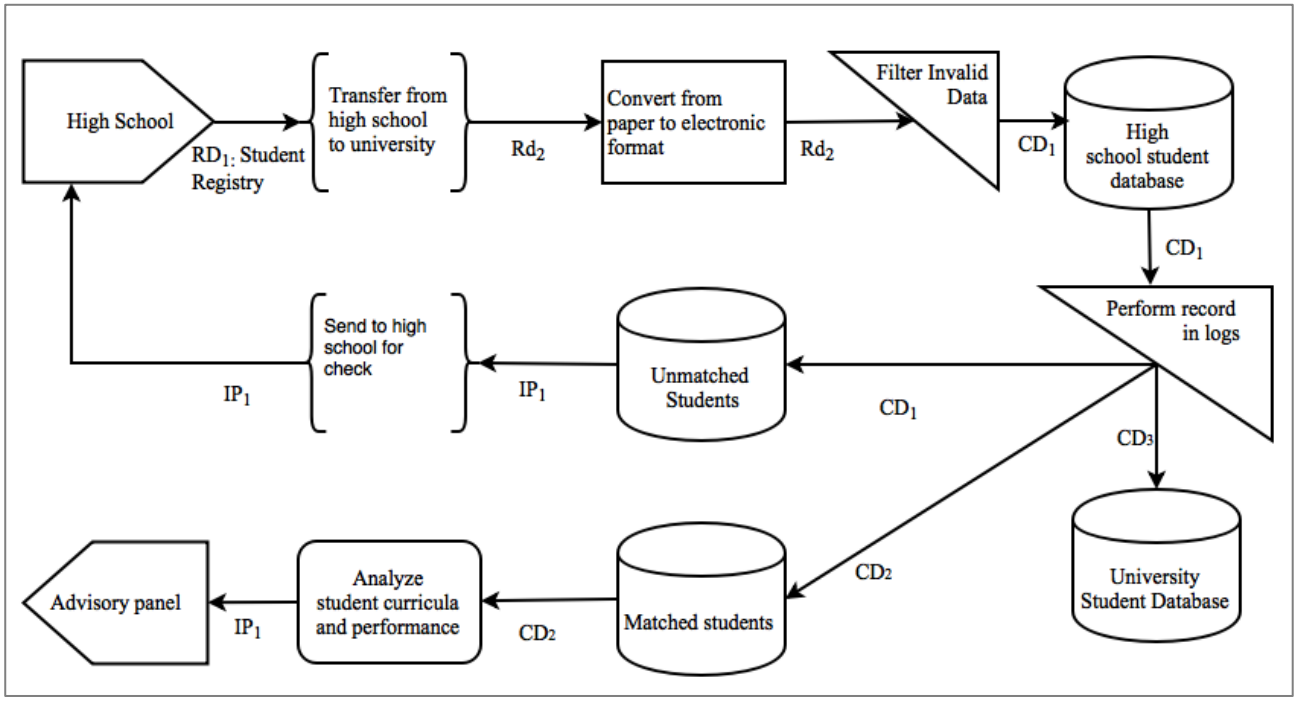


Figure 2.4: Example IP-MAP

Pierce (2002) and Scannapieco et al. (2002) have merged IP-MAPs with EPC and UML to make them usable for software engineers and business users. But to the best of our knowledge, there is no evidence to indicate that these extensions to UML and EPC are used among the practitioners, especially in the context of DQ management. The main reason could be the complexity of the diagrams due to the combination of multiple notations (Pham Thi and Helfert 2007).

Serrano et al. (2009) attempted to bring together the concepts introduced by IP-MAP along with BPMN to map the IP-MAP activities related to the DQ management with BPMN. The main goal of this attempt was to extend BPMN's capability to represent necessary DQ requirements for business processes and complement the expressiveness of BPMN by providing necessary extensions already in IP-MAP to BPMN. One of the main limitations of this work is that it considers only the DQ

requirements at the business process levels while leaving behind the concerns of business rules and software application level, which has a great impact on DQ.

In all these approaches to merge IP-MAP constructs with existing modelling notations (BPMN, UML, EPC), a common pitfall was that the resultant business process models, with extensions for DQ concerns, made the diagrams complex and were far removed from practical usage of their purpose. It is because that separation of concerns is an important aspect when specifications are created for management purpose, DQ management concerns embedded in a model created for business process management purpose or software engineering purpose will not fulfil either of the purposes.

In addition, when DQ management is considered as a separate management discipline in the organizational context, IP-MAP approach has further deficiencies. The notion of IP-MAP representation focuses on the information management process. It provides a "process oriented" view of data manufacturing and also highlights some proactive measures (eg. validation checks) to maintain the quality of the data at various stages within that process. Therefore, DQ as a management discipline requires a more detailed specification of DQ management functions in organizations (eg. DQ goals and Objectives drilled down into more measurable parameters to evaluate them, and suitable controls to enforce DQ management). In general, the "process oriented view" proposed by IP-MAP and its extensions to other modelling notations do not provide the capacity to relate the desired processes to specific data objects. We argue that a "data oriented view" would be more appropriate to represent the required knowledge to manage DQ since the data objects of concern can be linked to DQ goals, objectives, respective measurement criteria, thresholds and management controls to provide a holistic view about the entire DQ management activities.

A third approach towards developing a DQ Meta-Model can be found in Becker et al. (2007). In this approach, the authors recognize several DQ domain concepts (see Figure 2.5) that can be used in describing DQ management and explain the relationships between them. The main drawback of this work is that it does not provide any justification for selecting the domain concepts nor any validation. Due to its lack of theoretical underpinning or validation, its suitability cannot be assessed.

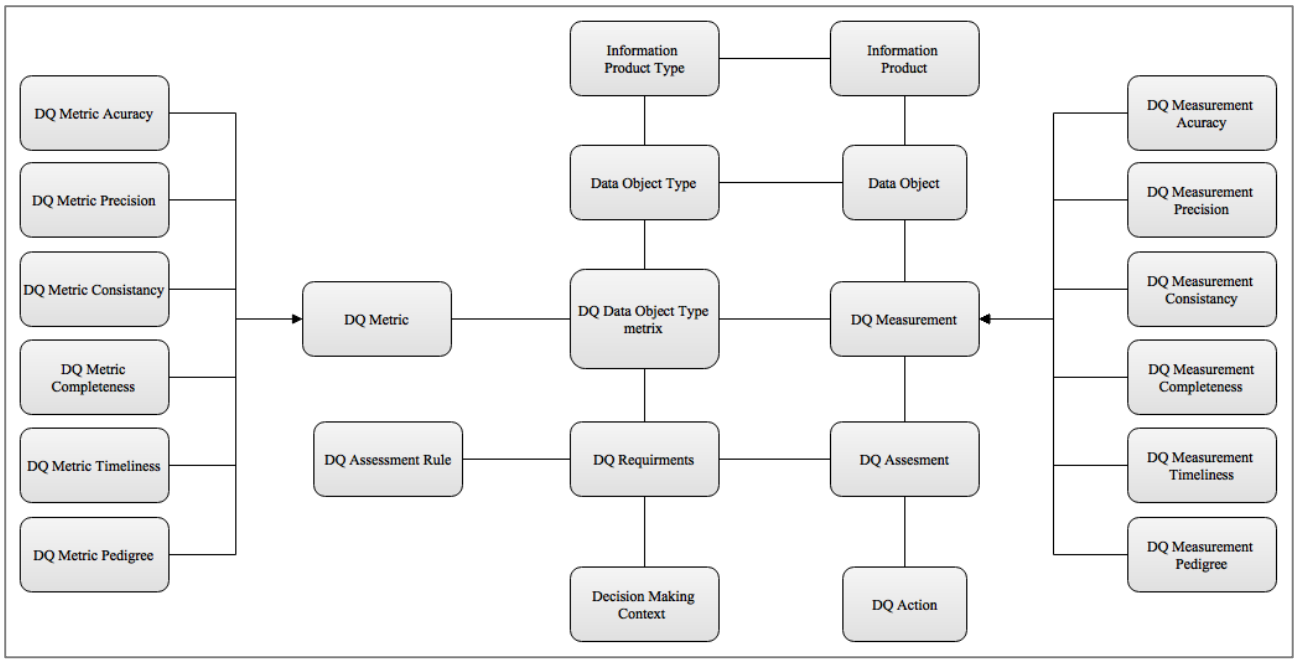


Figure 2.5: Basic DQ Meta-Model (Becker et al., 2007)

2.5 Research gap

Data quality has been researched over the past several decades, but when it comes to DQ requirements modelling, only a few approaches exist to model DQ requirements. These Meta-Models have the following limitations and weaknesses,

- Overloaded notations lead to complex diagrams which cannot be used to guide DQ management process. Therefore, these approaches have become "yet another notation" in modelling DQ requirements.
- Existing DQ Meta-Models are deficient in the coverage of the domain concepts and hence they are not capable of representing DQ requirements comprehensively.

Therefore, a necessity exists for a meta-model of DQ requirements that helps to develop DQ requirement models that balance notational complexity with adequate representation of all requisite DQ requirement concepts.

In section 2.2 we presented a number of methodologies developed by researchers and industry practitioners to manage DQ. As shown in Table 2.1 (section 2.2.3), DQ requirements analysis has been overlooked. Thus, no prominent methodology exists for analyse DQ requirements and comprehensively model them for management purposes. DQ requirements are fundamental to DQ management and in many existing DQ methodologies

Therefore, a necessity exists for a repeatable and tested methodology to analyse DQ requirements in organizations and elicit them to develop DQ models.

Recall that DQ requirement are fundamental to DQ management and the quality of data depends on how successfully the DQ requirements are met, we conducted an industry survey on how well the DQ management is performed in the industry (Jayawardene et al. 2012; Sadiq et al. 2011a). The study revealed that only 30% of the respondent claimed that DQ concepts have been well implemented in their organizations. Given the importance and challenges in defining the factors specify *what to manag*e and *how to manage* in DQ management, it is apparent that DQ requirement analysis and modelling is an essential for effective DQ management.

2.6 Chapter summary

In this chapter, we investigated existing literature on data quality relevant to this research. In particular, we perform a comprehensive study on existing DQ methodologies, literature on conceptual modelling and Meta-Models which is necessary to conceptualize a DQ requirement. Then finally we analysed DQ requirement modelling efforts in literature and their limitations emphasising on a research gap in modelling DQ requirements.

Chapter 3

METHODOLOGY

3.1 Overview

In this chapter, we illustrate the overarching methodology of this research. We systematically explain the steps of the research while referring to the relevant research methods that we used in every step with appropriate justification.

3.2 Positioning the research

Jenkins (1985) argues that there are a large number of research methodologies that are applicable for information systems research given the diversity of the field, while the best methodology must be determined within the context of the research objectives. Myers (1997) argues that both qualitative and quantitative research methods are appropriate for information systems research and suggest that qualitative research can be performed under three paradigms, viz. Positivist, Interpretive and Critical depending on the context of the research problem. Simon Herbert in his book *The sciences of the Artificial* (Simon 1996) argues that in the fields like engineering medicine business and architecture, the researchers are concerned about *how things might* be rather than *how things are* and, in this context the author argues that the design is a special perspective in manmade artefacts. Therefore, owing to the importance of the design of research artefacts, design science has become a prominent research method in information systems when the research objective is to develop an artefact to resolve a particular research problem of concern.

The models, methods, or instantiations produced as a result of research are described as artefacts by Hevner et al. (2004), and they argue design science as a suitable approach for developing such artefacts. After analysing a sample of design science research papers in leading IS journals Gregor and Hevner (2011) revealed that the artefacts developed using design science include both process and product artefacts that are applied in organizational settings. For example, product artefacts can be found in Roser et al. (2006) where the authors produce models for cross-organizational business process coordination and, in Purao et al. (2003) where the authors developed analysis patterns to improve conceptual designs. On the other hand process artefacts can be found in Rosenkranz and Holten (2011) where the authors build the Variety Engineering Method (VEM) as a process artefact to analyse, diagnose, and design information flows similarly, in Valverde et al. (2011) the authors develop a specific software engineering method as a process artefact to compare traditional and component-based models of systems in systems re-engineering.

As explained in chapter-1, the primary aim of this thesis is to develop a repository of DQ patterns, which is a product artefact to facilitate the modelling of data quality requirements of an organization. Therefore, this study uses design science as the principal research methodology for this study.

3.2.1 Design science

Design science (DS) has been used as a research methodology in IS research during the last 25 years (Iivari 2007). Simon (1996) argued that natural sciences and social sciences try to understand reality while design science tries to create things that serve human purposes. The contributions to design science by authors (Hevner et al. 2004; March and Smith 1995; Nunamaker et al. 1990; Walls et al. 1992) have improved the credibility of design science as a research methodology in IS domain. Peffers et al. (2007) argue that a DS methodology should meet three objectives, viz. it is consistent with prior literature, it provides a nominal process model for doing DS research, and it provides a mental model for presenting and evaluating design science research in IS. They further explain that the DS process includes six steps: (1) problem identification and motivation, (2) definition of the objectives for a solution, (3) design and development, (4) demonstration, (5) evaluation, and (6) communication.

Hevner et al. (2004) propose seven guidelines for design science research. These include: (1) Design as an artefact: Design science research must produce a viable artefact in the form of a construct, a model, a method, or an instantiation, (2) Problem relevance: The objective of design-science research is to develop technology-based solutions to important and relevant business problems, (3) Design evaluation: The utility, quality, and efficacy of a design artefact must be rigorously demonstrated via well-executed evaluation methods, (4) Research Contributions : Effective design-science research must provide clear and verifiable contributions in the areas of the design artefact, design foundations, and/or design methodologies, (5) Research rigor: Design science research relies upon the application of rigorous methods in both the construction and evaluation of the design artefact, (6) Design as a search process: The search for an effective artefact requires utilizing available process means to reach desired ends while satisfying laws in the problem environment and (7) Communication of research: Design science research must be presented effectively both to technology-oriented as well as management-oriented

The researchers emphasize that in design science research, rigour and relevance are important factors and have to be managed sufficiently (Kock et al. 2002; Offermann et al. 2009; Rosemann and Vessey 2008). While numerous arguments exist about rigor vs relevance (Rosemann and Vessey 2008), Hevner (2007) propose that rigour and relevance of design science should be managed iteratively

throughout the research process and he identifies three cycles in design science research projects where the above seven guidelines can be used effectively and efficiently (Figure 3.1)

The *Relevance Cycle* bridges the contextual environment of the research project with the design science activities. The *Rigor Cycle* connects the design science activities with the knowledge base of scientific foundations, experience, and expertise that informs the research project. The central *Design Cycle* iterates between the core activities of building and evaluating the design artefacts and processes of the research. Hevner (2007) argues that these three cycles must be present and clearly identifiable in a design science research project.

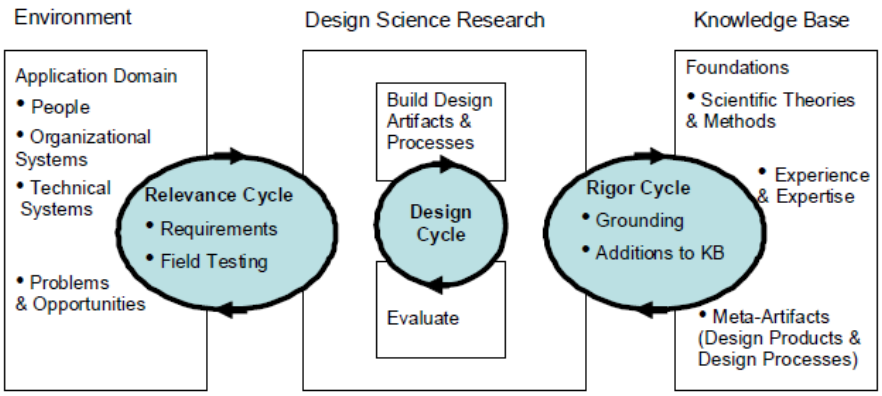


Figure 3.1: Three cycle view in DS research (Hevner, 2007)

The following section explains how these guidelines were followed in this research to develop the artefacts.

3.3 Research Design

The primary goal of this research is to design a repository of DQ patterns to model DQ requirements. As a secondary goal, we adapt an existing requirements engineering approach to support DQ requirements analysis from literature. The research was designed giving emphasis on the design science guidelines (Hevner et al. 2004; Peffers et al. 2007) and we considered the three cycle view of applying the guidelines as suggested by Hevner (2007) (see figure 3.1).

Hevner et al. (2004) suggest designing artefacts as a search process where the design will be achieved through searching solutions to sub-problems that constitute the main problem. Then heuristic problem-solving strategies are used to design solutions to each sub-problem so that the design of the final artefact will be built systematically. In chapter1, we presented five main research objectives to address the main research problem established in chapter 2. Therefore, in designing the research we identified the necessary activities/processes to achieve the five main objectives that resulted in eight

main processes viz. (1) Industry survey, (2) Conceptualization of DQ requirements, (3) Refactoring of DQ dimensions, (4) Card sorting study, (5) Development of DQ patterns, (6) Descriptive evaluation, (7) Adaptation of a requirements analysis methodology and (8) Applicability check. (More details about the rationale for the eight processes is given below). We believe that the three design cycle approach will enable us to relate each of the above processes, either to DQ literature (knowledgebase) or to the real world DQ practitioner scenarios (environment) so that the final design has a theoretical significance as well as a practical significance. Thus we argue that the three cycle view is a better approach for the design of this research. Figure 3.2 summarises the eight processes with regards to the three cycle view of design.

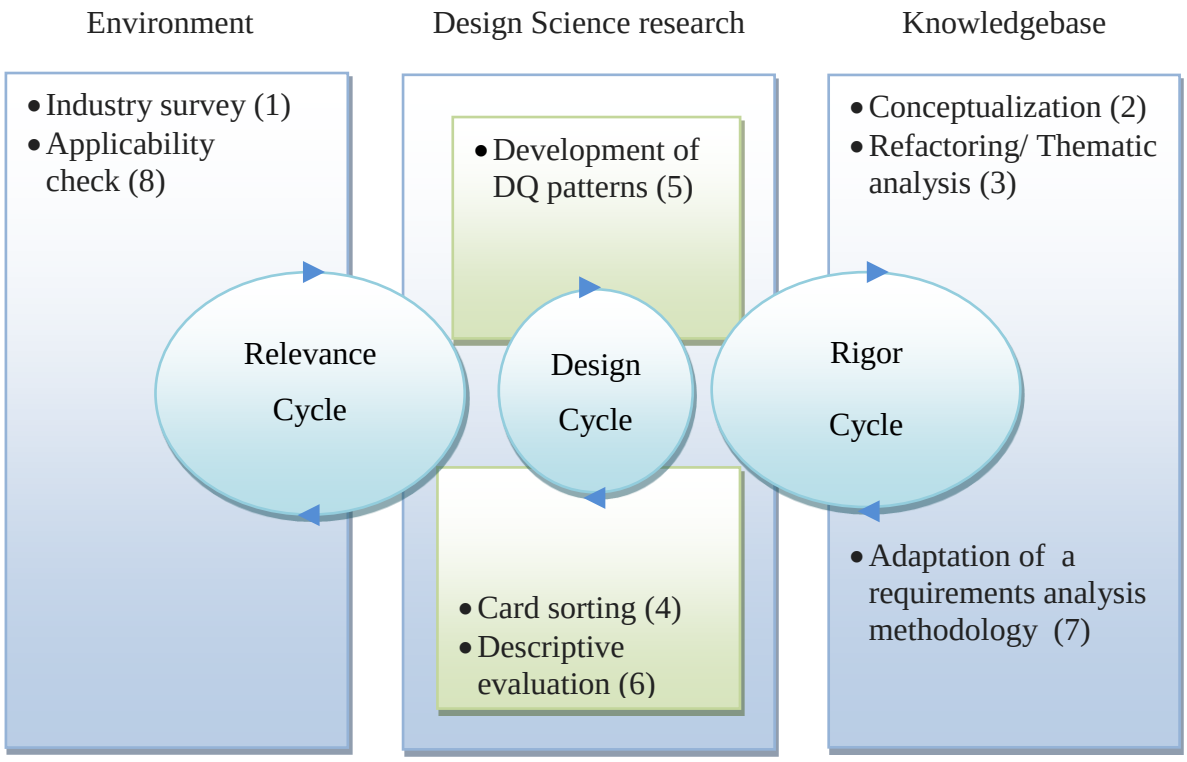


Figure 3.2: Design of the research based on three cycle view

In the following sections, we discuss each process in brief, and in the order of execution as explained in Figure 3.2, explaining how each step contributed to the research design in terms of rigor and relevance.

3.3.1 Industry survey

While the research problem was identified through a thorough literature review, we also conducted an industry survey to investigate how successfully the data quality concepts are being used in the industry, and what concepts contribute more effectively towards good quality data. The study revealed that good quality depends on whether the organizations recognize quality requirements of

its data that are critical for functioning, and, whether they implement proactive solutions to maintain the requirements. In other words the, organizations should design proactive solutions for DQ and thus need a clear specification of *what to manage* and *how to manage* (Jayawardene et al. 2012). Therefore, in addition to the literature review, the survey further confirmed that there is a necessity for clear specification of DQ requirements, and established the relevance of the research with regards to the environment as suggested by Hevner et al. (2004).

3.3.2 Conceptualisation of DQ requirements

In chapter-2, we revealed that in existing DQ requirements modelling approaches, the coverage of domain concepts is not sufficient to comprehensively define a DQ requirement to help developing management solutions. This research gap paved the way for our first objective of conceptualising a DQ requirement through the identification of requisite domain concepts. Therefore, a literature review was performed to identify the relevant concepts to represent a DQ requirement. Then, based on conceptual modelling literature we synthesized the domain concepts to develop a Meta-Meta model (Omg 2008). A Meta-Meta model is a preliminary stage of conceptualization used to develop a Meta-Model that in turn used in developing DQ requirement models in the real world (Chapter 2, Figure 2.1). As per Hevner et al. (2004) this step can be considered as strengthening the rigor cycle where the existing knowledge bases are used for the development of the artefact (More details in Chapter 4)

The second objective of the research is to develop a repository of re-usable DQ patterns. As explained in literature review, the lack of shared understanding about the concept DQ dimension is a main barrier towards achieving this objective. Therefore, a systematic refactoring of the concept of DQ dimensions was necessary to develop the shared understanding, which is the next process of this research.

3.3.3 Refactoring of DQ dimensions

Owing to the numerous number of different classifications of DQ dimensions available in the literature, we established that there is no consensus about this concept among researchers or practitioners (Jayawardene et al. 2013b). Since a shared understanding of a concept is a necessary prelude for conceptualisation (Guizzardi 2007), we performed a consolidation of existing definitions of DQ dimensions with a view to producing a shared understanding. As proposed by Hevner (2007), the design of an artefact has to be grounded on sound theoretical base to ensure the rigor of the design (rigor cycle). Thus, in producing a shared understanding, we used credible sources of DQ dimensions covering a wide range of literature produced by both academic researchers and industry practitioners.

In this task we selected 16 sources from the literature on DQ dimensions covering the perspectives of industry practitioners of DQ, market leaders of DQ management tools, data quality standards, DQ frameworks of real world organizations and academic publications. The selection process of the literature is explained in chapter 5 in detail.

In this consolidation, we used *thematic analysis* as the data analysis technique (Braun and Clarke 2006) to consolidate different definitions provided by different authors and develop a consolidated classification of DQ dimensions.

3.3.3.1 Thematic analysis

Thematic analysis is a qualitative data analysis technique that emphasizes pinpointing, examining, coding and recording patterns (themes) of data considering the semantics, so that a categorization of data can be done (Fereday and Muir-Cochrane 2006). As per Guest et al (2011), thematic analysis is the most frequently used qualitative data analysis technique in recent years since it support building in-depth insights about discrete qualitative data. In our analysis, an in-depth investigation is required about different views of authors in defining DQ dimensions and, hence we argue thematic analysis a suitable data analysis technique in this research. Similar studies can be found in recent literature where a large corpus of viewpoints are analysed using thematic analysis to build shared understanding that support our argument (Jones et al. 2011, Fereday and Muir-Cochrane 2006).

Braun and Clarke (2006) has developed a set of guidelines to perform effective thematic analyses that include the following phases.

Phase 1: familiarising with data: Since the semantics play a vital role in thematic analyses, it is essential to immerse yourself with the depth and breadth of the content. Immersion involves repeated reading of the data, reading the data in an active way searching for meanings and commonalities etc.

Phase 2: generating initial codes: Coding refers to identifying features of data based on the semantic content or latent that appears interesting to the analyst. While coding can be done manually or with the help of software tools, in both cases it is important to organize the codes in a way it supports easy browsing.

Phase 3: Searching for themes: The codes have to be carefully examined and collated, ordered or categorised into overarching themes. For this, we can use tables, mind-maps or any other visual representation to support arguments about the semantics of the codes. The themes recognised at this phase are called candidate themes, which mean they may form useful themes.

Phase 4: Reviewing themes: The candidate themes are further examined and see if the collated extracts under each theme are coherent enough to develop a consistent theme. If the extracts are not coherent, the extracts should be split into more coherent themes (in this case new candidate themes will occur in the analysis) or they should be spread across other candidate themes (in which case the original candidate theme will disappear from the analysis).

Phase 5: Defining and naming themes: In this case all candidate themes will be examined for the essence of what each theme is all about. Since each candidate theme contains coherent extracts, (codes) it is possible to provide a meaningful name and a definition for the theme.

Phase 6: Producing the report: The results of thematic analysis should be presented with sufficient trail of evidence so that the rationale for each theme is visible. Therefore, the themes can be presented at the different level of aggregation to facilitate comprehension.

We used the above guideline by Braun and Clarke (2006) in analysing the existing definitions in literature for DQ dimensions and developed a new classification of DQ dimensions which includes the essence of all the viewpoints held by different authors that we considered in this analysis.

In coding and analysing qualitative data, biasedness of the coder/analyser can tarnish the reliability of the findings (Auerbach and Silverstein 2003). Therefore, a dual coder approach has been suggested by many researchers where two or more coders independently code and analyse data and, conclusions are arrived after eliminating the differences through discussions (Auerbach and Silverstein 2003; Richards 2014). Therefore in our thematic analysis we used dual coder approach to ensure the reliability of findings. More details about this thematic analysis are provided in Chapter 5.

Once a new classification of DQ dimensions is developed, the next task is to evaluate the effectiveness of the terminology used in the definitions. Then we used a card sorting study as the tool to validate the definitions and terminology used in the new classification, which we outline as follows.

3.3.3.2 Card sorting study

The result of the thematic analysis produced a new classification of DQ dimensions to build a shared understanding. Each theme identified in the analysis was provided with a new definition and a name to represent the theme. As Hevner (2007) argues that during the design cycle it is important to evaluate the artefact. Therefore, in building this new classification, it is important to ensure that the new definitions used in the classification provide the same meaning (intended meaning) to all the stakeholders. Therefore, it was necessary to ensure (1) the clarity of the definition and (2) the clarity

of the name given to the definition of the new classification. Therefore, we sought an expert judgement to ensure the suitability of definitions and the terminology.

In literature, Delphi technique (Brown 1968; Powell 2003), a focus group (Morgan 1997), and Card sorting (Moore and Benbasat 1991) has been used by researchers to incorporate expert judgement in building consensus about knowledge. In more recent years, card sorting studies have been used to obtain user experience to understand how users classify and structure the contents in designing of user interfaces (Spencer 2009). Nawaz (2012) argues that one strength of the card sorting study is that, it uses natural thought process (mental model) of users in categorising and sorting knowledge, based on the semantics (Nawaz 2012). Therefore, due to its support towards natural processing of semantics in human mind, we posit that the card sorting is a suitable technique to validate the clarity of definitions and terminology of the new classification. More details about this validation are provided in section 5.4 of chapter 5.

3.3.4 Development of DQ patterns

As explained in section 2.3.2, Omg (2008) specifies four layers of modelling where Meta-Meta model (M3) is the starting point of conceptualization. Then comes the Meta-Model (M2) which is an instantiation of the Meta-Meta model. In chapter 4, we establish the constructs of the Meta-Meta model where DQ characteristic is the central construct and in chapter 5 we develop instances for DQ characteristic and the other related constructs that is required for this instantiation. Therefore, the Meta-Meta model developed for a DQ requirement as explained in section 3.3.2 is instantiated at this stage. The refactored DQ concepts were used for this instantiation, where every instance of the Meta-Meta model produced a DQ pattern that consists of a DQ requirement (violation of which is a DQ problem) and a solution space which can be used to prevent a DQ problem. More details about this instantiation is provided in chapter 6. The solution space was designed referring to literature on DQ improvement strategies most specifically considering DQ practitioners' contributions towards DQ improvements (English 2009; Loshin 2010; McGilvray 2008; Redman 2008). Thus, through instantiation, we produced a repository of DQ patterns.

Once a repository of DQ patterns were developed, the third objective of the research is to evaluate the completeness of the repository in representing DQ requirements. For this purpose we designed a descriptive evaluation as outlined below.

3.3.5 Descriptive evaluation

Hevner et al. (2004) suggest that design artefacts have to be rigorously evaluated for their utility in addressing the problem of concern. Therefore, it is necessary to evaluate the DQ patterns to ensure whether the DQ patterns are sufficient (complete) in representing any DQ requirement of structured data. Descriptive evaluation is a methodology proposed by Hevner et al. (2004) to evaluate design artefacts where evidence from relevant research in literature is used to build a convincing argument about the artefact.

As discussed in section 2.2.1, DQ problems and requirements are two sides of the same coin. Therefore, our validation focuses on DQ problem perspective as well as DQ requirement perspective. We used sources from the literature on DQ problems and sources from the real world on DQ requirement to perform the validation. By referring to existing literature of DQ problems, we were able to assess if the previously established DQ problems have a corresponding pattern in our pattern repository. In other words, we were interested in determining if the existing patterns are sufficient to address problems discussed in the literature. By referring to industry-based data quality rule repositories we checked if every data quality rule/data related business rule have a corresponding pattern in our collection. More details about this evaluation process are given in Chapter 7. Thus, in this evaluation, we referred to literature bases to ensure the rigor of the artefact while ensuring the relevance of the artefact by referring to industry perspective as suggested by (Hevner 2007).

The fourth objective of the research is to develop a requirements modelling approach to facilitate the use of the patterns to model DQ requirements in an organization. Hence we design the next step as follows.

3.3.6 Adaptation of a Methodology for DQ Requirement Analysis and Modelling

The DQ patterns developed in this research provide a base to model DQ requirement. Then the next question is how to use the patterns in modelling DQ requirements. As discussed in section 2.2.3, most of the DQ methodologies have overlooked the perspective of DQ requirement analysis. But in the literature of information systems, requirements engineering is a well-established area of research with prominent methodologies for effectively capture system requirements. Therefore we analysed existing requirements engineering methods with the aim to analize which methodologies suits better for using pre built requirement patterns in given organizational context. Then, we selected an appropriate existing methodology from requirements engineering literature and adapt it to cater for DQ requirement. By referring to literature we provide necessary justifications for our selection of the

methodology and more details about selecting and adapting of a suitable methodology is given in Chapter 8.

The fifth objective of the research is to evaluate the applicability of the DQ patterns by using them in real-world organizations. We designed an applicability check to fulfil the objective.

3.3.7 Applicability Check

The relevance of design artefacts to practice has been an important consideration in information systems research (Rosemann and Vessey 2008). Hevner (2007) emphasizes that a design artefact should successfully satisfy the need of the environment, thus, they can be used to solve problems in the real-world.

In our research, the main artefact is the repository of DQ patterns. In order to perform the requirements analysis and modelling, we anticipate the need for a requirements modelling methodology based on the effective use of the patterns. Thus, we have also adapted a methodology to use the patterns for DQ requirements modelling (namely KAOS4DQ, see chapter 8) which is the secondary artefact of this research.

In literature Technology Acceptance Model (TAM) (Venkatesh and Davis 2000) is widely used to evaluate the research artefacts in information systems domain. In this approach, perceived usefulness and perceived ease-of-use are evaluated with regards to the artefact in concern. Rosemann and Vessey (2008) proposed three dimensions to study the applicability of research artefacts, viz. importance, accessibility, and suitability. In this framework, importance refers to whether the artefact meets the need of practice, accessibility refers to how well the practitioners can understand the artefact (the presentation) and suitability refer to how well the artefact fits into the environment. Further, the authors argue that an artefact has to be deemed as important first to make it suitable for the environment, thus, importance and suitability are overloaded (Rosemann and Vessey 2008).

In consideration of the above viewpoints in literature we were motivated to conduct an empirical study to conduct an *Applicability Check*, that is, to study the importance, accessibility and suitability of the patterns for DQ requirements analysis and modelling:

1. How useful DQ patterns are in analysing and modelling DQ requirements (importance)
2. How well the data quality users connect with the artefact (accessibility)
3. How accurately the DQ requirement was elicited and modelled (suitability)

We designed an empirical study to conduct the Applicability Check based on the above considerations. We conducted two studies, in two organizations. The protocol for the applicability

check is given in Appendix-A and more details about the design of applicability checks are given in Chapter 9. As discussed above, the research was designed and conducted based on the design science guidelines and the roadmap of the research is provided in Figure 3.3 below.

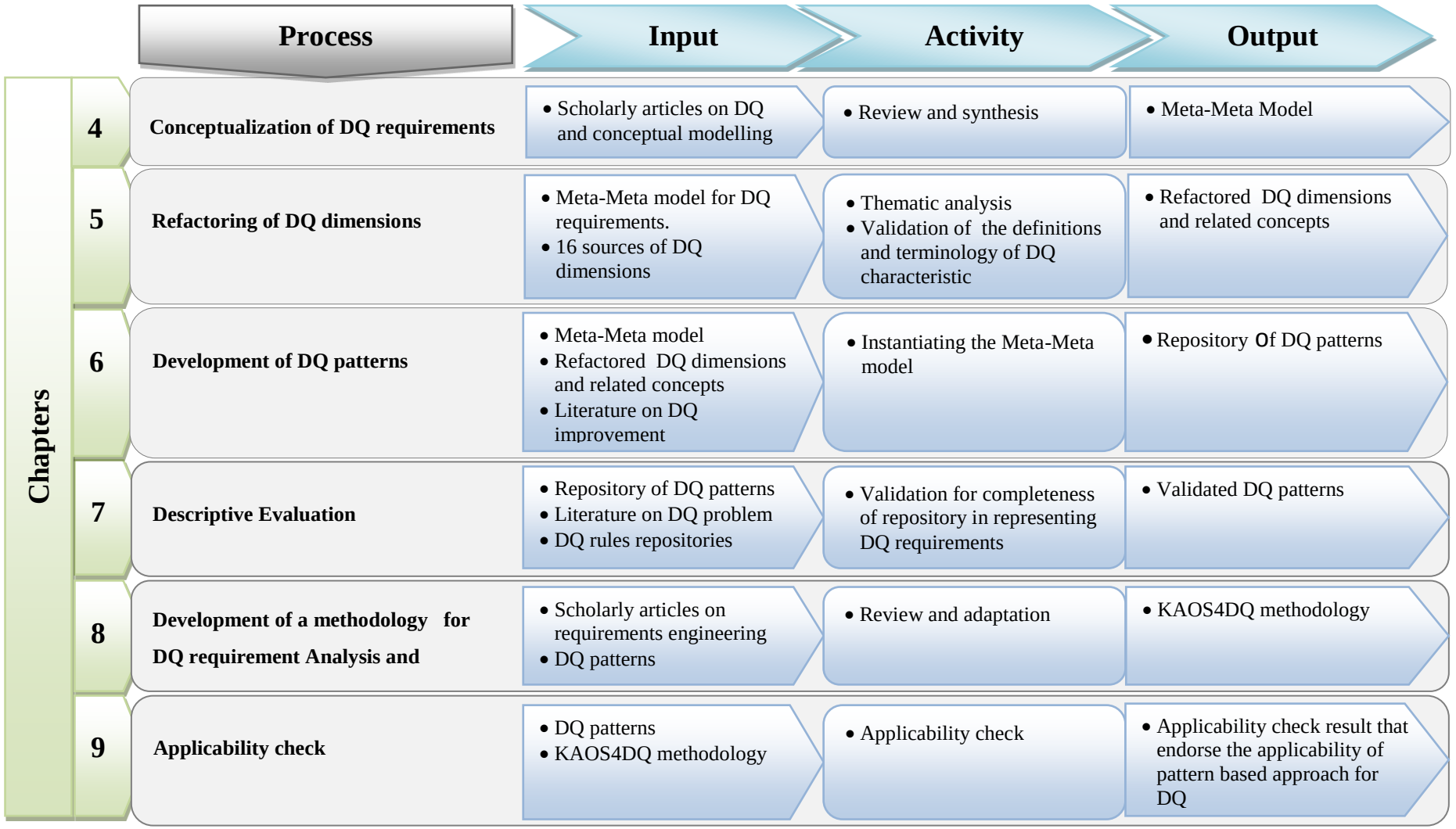


Figure 3.3: Roadmap of the research

3.4 Chapter Summary

In this chapter, we presented the overall design of the research following the design science methodology. The roadmap of the research is described in figure 3. The primary research artefact is the repository of data quality patterns developed through an iterative approach spanning across the rigour, design and relevance cycles proposed in design science. The patterns are evaluated for their sufficiency by using a descriptive evaluation, for their clarity by using a card sorting study and finally for their applicability by using an applicability check in the industry.

Chapter 4

CONCEPTUALIZATION OF DATA QUALITY REQUIREMENTS

4.1 Overview

This chapter presents a meta-meta-model to conceptualize DQ requirements. Meta-modelling is a well-recognized approach for conceptualization within the information systems and computer science body of knowledge. As explained in chapter-2, a meta-meta-model is the initial conceptualization of domain concepts, while a meta-model is an instantiation of the meta-meta-model that builds a shared understanding in a domain among stakeholders. Therefore a meta-meta-model for DQ requirements will provide the initial conceptualization of DQ concepts so that in the next step a meta-model can be built to develop a shared understanding of the knowledge related to a DQ requirement. In this chapter, we systematically investigate the DQ concepts required to build a meta-meta-model for a DQ requirement.

4.2 Foundational concepts of a DQ requirement

The ISO define quality in general as "a set of inherent characteristics fulfils requirements" (ISO 2000). Similarly, quality of data can be defined with reference to a set of characteristics of data that fulfils data users' requirements. Thus it is apparent that DQ characteristics can be considered as a fundamental concept in expressing a DQ requirement. On the other hand, the lack of adherence to a DQ characteristics prohibits fulfilment of user requirements and thus creates a DQ problem. Therefore, DQ characteristics are a central concept to both DQ user requirements and DQ problems. In other words, DQ user requirements and DQ problems are two sides of the same coin. (Figure 4.1)

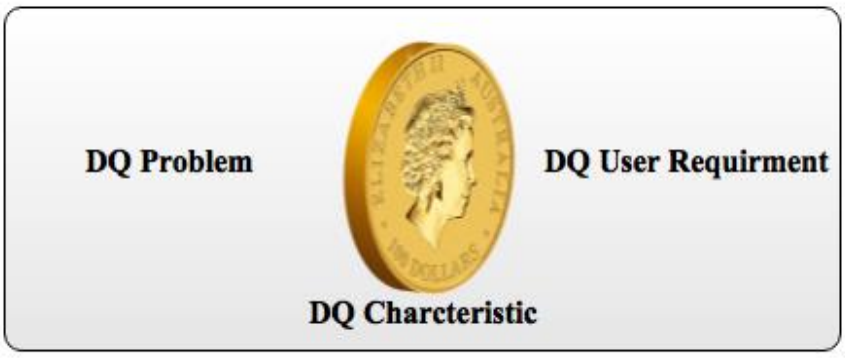


Figure 4.1: DQ requirements, problems, and Characteristics

Wang and Strong (1996) explain that data should have 16 characteristics to meet user expectations in performing tasks. However in literature, these characteristics are often referred to as DQ dimensions (Batini et al. 2009). The word "dimension" is defined as "a measurable extent of a particular kind, such as length, breadth, depth, or height" (Dictionaries 2010). In other words, a dimension is an aspect of an object or phenomenon where measurements can be made. Therefore, in DQ literature the term DQ dimension is used with the objective of measuring the quality of data rather than expressing DQ requirements. Despite the ambiguity in terminology, we argue that the notion of DQ dimension/DQ characteristic can be used intuitively for defining DQ requirements as well as quantifying data quality.

It should be noted that there are instances in literature that DQ dimension is used as a higher level abstraction of DQ characteristics. In the DQ methodology used in the Canadian Institute for Health Information (CIHI) (Long and Seko 2005), the authors have used the three terms *DQ criteria (metric)*, *DQ characteristic* and *DQ dimension* at different levels of abstractions in their DQ framework. In the first level, 86 basic DQ metrics are defined to practically measure the quality of data. Then these metrics are aggregated using composition algorithms into 24 DQ characteristics at the second level. They define these DQ characteristics as a meaningful way to present the fitness for use of data from data users' point of view. For example, each DQ characteristic is a specific aspect of data (such as over coverage of data, under coverage of data) that need to be maintained to perform a particular task (statistical calculation). At the third level, the 24 DQ characteristics are further aggregated into five DQ dimensions (accuracy, timeliness, comparability, usability, and relevance) and used for management presentation purposes so that the organization's overall DQ can be presented using five dimensions (Figure 4.2). It should be noted that in this case, the term DQ criteria (metric) is an ultimate measurement, whereas DQ characteristic and DQ dimensions are two level of abstractions used for presentation and management purposes.

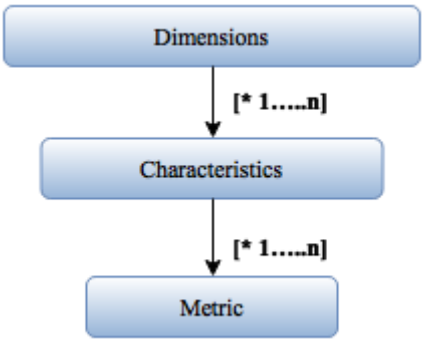


Figure 4.2: Structure of DQ dimensions vs Characteristics

We observe that such an abstraction is necessary for managing the quality of data between different levels of an organization. For example, a data user at the lowest level who would be interested in whether data values for the raw material price is given for a precision of two decimal places to calculate the costing of a raw material stock. Whereas the warehouse manager would be interested in knowing how accurate the stock values are, which include a wide variety of other characteristics of accuracy like the accuracy of entering raw material data to the system, and the consistency between the stocks in the system and the actual stocks in the warehouse etc.

Therefore we establish that an abstraction of DQ characteristics as a DQ dimension and they are high-level perspectives like accuracy, completeness, currency, usefulness etc, whereas one DQ dimension can contain more than one DQ characteristic. For example, the dimension accuracy can have the characteristics precision, accuracy to reference source etc. that are defined based on the data users' requirements.

The following section explains the concept of DQ characteristic in detail.

4.2.1 Data quality characteristics

The notion of quality characteristics is apparent in many quality domains such as quality of products and services. For example, Garvin (1987) defines eight characteristics of product quality, *viz.* performance, features, reliability, conformance, durability, serviceability, aesthetics, and perceived quality (See Table 4.1).

Characteristic	Definition
Performance	The product's primary operating characteristic (such as acceleration, braking distance, steering, and handling of an automobile)
Features	The ``bells and whistles'' of a product (such as power option and a tape or CD deck of a car)
Reliability	The probability of a product's surviving over a specified period of time under stated conditions of use
Conformance	The degree to which physical and performance characteristics of a product match pre-established standards
Durability	The amount of use one gets from a product before it physically deteriorates or until replacement is preferable
Serviceability	The speed, courtesy, and competence of repair
Aesthetics	How a product looks, feels, sounds, tastes, or smells

Perceived quality	The subjective assessment of quality resulting from the image, advertising, or brand names.

Table 4.1: Product quality dimensions (Garvin 1987)

From this classification, it is evident that the characteristics lead to a measurable perspective of the product itself. The underlying idea is that once the specification for the product is created using these characteristics, product quality can be measured by evaluating the extent to which the prescribed values for the characteristics are achieved. It should be noted that some of these perspectives are declarative in nature, explaining the product precisely (performance, features, durability, reliability, conformance etc.); i.e. they explain the inherent or representational nature of the product independent of its users. Others, on the other hand, describe users' measures (perceived quality, serviceability, aesthetics) facilitating a judgment of the product that depends on its usage.

Similarly, Russell and Taylor (2003) define the characteristics of service quality as time and timeliness, completeness, courtesy, consistency, accessibility and convenience, accuracy, and responsiveness (see Table 4.2). In this classification the dimensions are defined using the *declarative* perspective to explain the service (completeness, accuracy, time and timeliness) as well as the *perceptional* perspective, facilitating the user's judgment of the service (courtesy, consistency, accessibility and timeliness, responsiveness) when the service is consumed.

Characteristic	Definition
Time & Timeliness	Customer waiting time, On-time completion
Completeness	Customers get all they ask for
Courtesy	Treatment by employees
Consistency	Same level of service for all customers
Accessibility and convenience	Ease of obtaining service
Accuracy	Performed correctly every time
Responsiveness	Reaction to special circumstances or requests

Table 4.2: Service quality dimensions (Russell and Taylor 2003)

We observe that these studies on product and service quality consider characteristics that satisfy user requirements in both the declarative and usage perspectives. Therefore we argue that the two perspectives can be applied to DQ domain as well. Adapting product quality principles to data domain was first proposed by Wang (1998) considering that "information is processed data". Wang (1998) argues that information is analogous to products and data is analogous to raw materials in a typical

product manufacturing process where the author recognizes an information manufacturing process analogous to a product manufacturing process (Wang 1998). Thus, we argue that it is justifiable to adapt product and service quality features to data domain to further improve the concept of DQ characteristics.

In the next section, we discuss how these declarative and usage perspectives resonate within the data quality domain and how they can be used in a similar manner to characterize data quality requirements.

4.2.1.1 Type of data quality characteristics

Declarative and usage perspectives can be used for reasoning about data quality characteristics, in which case data quality characteristics can be classified into two distinct perspectives similar to products and services as explained above., viz:

Declarative Perspective (D): Focuses on user independent characteristics of data that are concerned with the definition of data. The representation of the physical reality as data is characterized in the form of meta-data, schema, and the operational conditions of organizations (business rules). These characteristics of data explain data itself and can also be considered as inherent characteristics of data. Hence these characteristics of data can be implemented ***by design*** and assured through the information system itself, independent of system users and usage (a task at hand). It should be noted that the DQ problem pertaining to declarative characteristics can be found by querying the database.

An example for a DQ characteristic of declarative type is;

Meta-data compliance: data should comply with its Meta-data.

Usage Perspective (U): Focuses on user dependent characteristics of data that leads to effective usage of data. These characteristics emerge during the use of data in performing a particular task or process. Hence, the characteristics can be defined referring to a task or a process, and may also vary from performance indicators of the task to perceptional judgments about the fitness for use in accomplishing the task. It follows that these usage perspectives can be implemented ***by conformance*** to the policies procedures and principles (eg. TQM), standards (eg. ISO), best practices or any technology/tool which facilitate data usage. It should be noted that the DQ problem pertaining to usage characteristics cannot be found by querying the database, instead, they are uncovered due to a failure or an underperformance of a task performed using data.

An example for a DQ characteristic of usage type is;

Understandability: data is understandable

Thus we establish that a DQ characteristic is a fundamental notion behind a DQ requirement, and there are two types of DQ characteristics that distinguish their implementation in the information systems landscape either by design or by conformance. In the following sections, we describe a fundamental concept that helps to further describe a DQ characteristic.

4.2.2 Granularity of DQ characteristics

When reasoning about the DQ characteristics, we argue that it is important to consider at which data granularity level a characteristic can be applied. The lack of such consideration would make it difficult to use the characteristic as a basis for defining measurements of quality. Industry practitioners have pointed out that, to manage the quality of data objectively, a clear specification of what data to consider for quality management is required (English 2009; Loshin 2001). These data may span across the atomic level to various levels of aggregations like reports. However, this consideration is often lacking in research that uses quality characteristics as a basis for assessing data quality (Batini et al. 2009; Eppler and Muenzenmayer 2002; Lee et al. 2002; Pipino et al. 2002). While we agree that defining granularity at the dimension level (Figure 4.2) is difficult due to the high-level nature of the dimensions, we argue that specifying clarity on the granularity level is feasible at the characteristic level and, will allow us to distinguish subtle differences between DQ characteristics within the same dimension.

Even and Shankaranarayanan (2005) provide insight into granularity levels of data by considering a hierarchy of data as data items (elements), data records, datasets, databases and organizational database collections when measuring data quality. In studying data quality characteristics, we argue that the DQ characteristics should be defined referring to the granularity of data so that a particular DQ characteristic can be applicable only in one data granularity level. For example *Meta-data compliance* is a DQ characteristic applicable in data element level, while *Completeness of records* is a DQ characteristic applicable at a higher granularity level such as a record or a collection of records. Further, we argue that granularity depend on the type of the characteristic (i.e. declarative (D) or usage (U). Accordingly, in our work we consider three granularity levels of data:

Data element (E): An attribute of a real world entity.

Data record (R): A collection of attributes that represents a real-world entity in a database.

Information object (IO): A collection of any arbitrary records or elements used to accomplish a task

We argue that declarative characteristics should be primarily defined on data elements and records since they are related to data definitions as explained in 4.2.1.1 above, while usage characteristics are defined on any abstraction of data elements and records retrieved from a single or multiple relations as required by the usage (task at hand).

For example, the characteristic *Meta-data compliance* (data values comply with its Meta-data) is applicable on data elements (E), the characteristic *Uniqueness* (data should be uniquely identifiable) is applicable on data record level (R), and the characteristic *Understandability* (data is understandable) is applicable on information object (IO).

When reasoning about DQ characteristics, another important aspect is how these characteristics are maintained in data. Therefore in the following section, we discuss the implementation form of the DQ characteristics.

4.2.3 Implementation form of DQ characteristics

Batini et al. (2009) point out that quality improvement strategies have been defined in every DQ methodology. We observe that in most data quality management frameworks developed by industry practitioners (English 2009; Loshin 2011; McGilvray 2008; Redman 1997) they have developed strategies to proactively prevent the occurrence of bad data. Owing to the axiom that prevention is better than cure, in DQ management the cost of bad data is very much higher than the cost involved in implementing preventive mechanisms of bad data (English 1999). Therefore more prominence has to be given to prevent bad data in the systems rather than performing expensive data cleansing tasks. Since bad data is a result of not maintaining the required quality characteristics in data, it follows that proactive mechanisms should be designed and implemented to maintain the quality characteristics of data.

As defined in section 4.2.1.1 we argue that the implementation form of a data quality characteristic is dependent on its type (D/U). Thus, a declarative characteristic can be defined and implemented ***by design*** of information systems and they can be maintained independent of the data users. Therefore we propose a ***rule-based approach*** to implement declarative characteristics that include**,**

1) Database schema level rules; and/or
2) Application program level rules

Defining schema level rules is a well-established practice in design of databases (Elmasri 2008) Therefore, rules can be implemented to maintain the declarative DQ characteristics for a particular data object. For example, DQ characteristics like *Meta-data compliance* and *Uniqueness of records*,

can be implemented at database level rules which is known as domain constraint and key constraint respectively (Elmasri 2008). Further implementation of application program level rules is also a well-established practice in designing software systems (Ross 1997; Halle and Ronald 2001). For example, a DQ characteristic like *business rules compliance* can be implemented by design of the information systems.

In contrast, usage characteristics are dependent on data users as we defined them in section 4.2.1.1. Thus, how well data is being used in a task is the concern with regards to the usage characteristics. Therefore, the productivity in data usage in performing a task has to be improved to implement a usage characteristic by considering,

1) how to improve the process of data usage and
2) what capabilities and resources are necessary to optimise the data usage

Therefore we argue that a ***process based approach*** is needed to maintain usage characteristic that includes,

1) Processes: Implementation of benchmark practices such as conformance to policies, procedures, or principles, best practice or scientific theory that is agreed to be a benchmark;
2) Capabilities and resources: Implementation of technological solutions to facilitate data usage

Therefore, to maintain a usage characteristic, the data users have to follow the benchmark practices or use the technological solutions. Thus user adherence/conformance is important in maintaining the characteristic. For example, the DQ characteristic *understandability* (of a report) can be maintained by establishing a practice to have a legend for abbreviations in every report (English 2009), *data access control* can be implemented through system security tools (Loshin 2001).

Once the DQ characteristics have been implemented, the next concern is to check the status of data quality. In section 4.2 above, we pointed out that the concept of DQ characteristic is also referred to as DQ dimensions in literature and used in defining measurements for data quality. In the following section, we explore into the measurement aspect of DQ.

4.2.4 Metrics of DQ characteristics

Batini et al. (2009) analyse existing literature on DQ methodologies over a decade and illustrate that DQ assessment is considered as a mandatory element in all DQ methodologies. In DQ assessment, quality of data is evaluated using metrics. Data quality metrics quantify the extent to which data quality characteristics are maintained in a particular data element, record or information object.

Batini et al. (2009) found that all DQ methodologies define DQ dimensions (DQ characteristics) and each DQ dimension is measured by one or more DQ metrics. The study revealed that, while some of the researchers have recognised DQ dimensions, they haven't defined explicit metrics for all dimensions. For example, DQ dimensions like accuracy, completeness, and consistency have metrics defined by many researchers, while it is significantly lower for the time-related dimensions like timeliness and currency and almost all other dimensions (Batini et al. 2009). It should be noted that from DQ measurement point of view, a dimension without a metric is meaningless and useless. Heinrich et al. (2007) observe that, in practice, most data quality metrics are developed on an ad-hoc basis to solve specific problems and thus are often affected by subjectivity. Thus we need to formalize the concept of DQ metrics in such a way it relates to a DQ characteristic and thereby relates to a DQ requirement.

4.2.4.1 Fundamentals of quality metrics

COBIT (IT Governance Institute 2007), a framework for information technology governance, identifies two types of metrics to measure system quality: outcome measures and performance indicators. Outcome measures indicate "whether the goals and objectives have been achieved". These metrics can be measured retrospectively after the event or outcome and, therefore, are also called lag indicators. In contrast, performance indicators assess "how well the goals and objectives are likely to be met" (IT Governance Institute 2007). They can be measured before the event or outcome is clear and, therefore, are called lead indicators or performance drivers. They measure the availability of appropriate capabilities, practices and skills, and the outcome of underlying activities that are necessary to achieve the goals. Similarly to COBIT, ISO 9001:2008 (ISO 2000) an industry standard for quality management, defines two kinds of metrics for any goal: verification metric and validation metric. The verification metric measures "to what extent the intended outcome has been achieved", while the validation metric measures "to what extent the right actions are being taken to produce a desired outcome" (ISO 2000).

Analogous to the above viewpoints, in data quality management we can define the verification and validation metrics as follows:

Verification metrics measure the extent to which a characteristic has been maintained in the existing data. A verification metric has a threshold value, which is a target value to achieve the expected quality level.

Validation metrics measure the degree to which the required mechanisms have been established or implemented to maintain a specific quality characteristic of data. In other words, these metrics assess the capabilities required for maintaining good quality data.

It should be noted that verification metrics provide the status of current data quality and validation metrics provide the organization's maturity in regards to achieving data quality. The organization's maturity towards a specific data quality dimension's characteristic refers to whether the required capabilities exist in the organization to maintain that characteristic. Hence these two types of metrics complement each other and provide management with the capability to decide on how to design data quality improvement activities.

4.2.4.2 Metrics for declarative type DQ characteristics

For declarative data quality characteristics, which are implemented using rules, the number of violations of rules can be considered to be a verification metric to measure quality. Hence, we define the generic form of a verification metric for declarative characteristics as:

The number of non-conforming data values found per month /per thousand records due to exceptions to, or violations of, the rules that define the declarative characteristic.

For example, *uniqueness* is a declarative characteristic of data records and the following measure can be used as a verification metric.

The number of duplicate records reported per thousand records in patient data

The generic form for a validation metric of a declarative characteristic can be defined as:

The extent to which required rules have been identified and implemented to maintain the declarative characteristic in concern.

For example, the following measure can be used as a validation metric for the uniqueness of data.

The extent to which uniqueness rules have been defined and implemented to avoid duplication of patient data records.

4.2.4.3 Metrics for usage type DQ characteristics

For usage characteristics, a surrogate measure is required to capture the performance of the task that the data is being used for. Hence we define a generic form of a verification metric for a usage characteristic as:

The number of tasks failed or negatively affected per month due to the lack of conformance to the usage characteristic or

The number of complaints received per month due to the lack of conformance to the usage characteristics

Both statements above measures the same phenomenon while in practical circumstances the latter would be more convenient. It should be noted that for a usage characteristic, there can be more than one verification metrics since all of them are surrogate measures, and a combination of measures may provide better results (English 2009; Loshin 2001)

For example, *data freshness* is a usage characteristic and the following measure can be used as a verification metric.

The number of telephone calls to customers failed in business promotions due to outdated contact numbers

For a usage characteristic, we define the generic form of a validation metric as:

The extent to which, required capabilities and processes have been implemented to improve the data usage of a task

An example validation metric would be, how mature the process of refreshing customer contact details is.

4.2.5 Thresholds for metrics

Once the metrics are defined, they should be evaluated from time to time to check the quality of data. For this evaluation, a threshold value also should be defined to compare the existing state of quality with the expected level of quality. In other words, a threshold is a target that the metric should meet to reach an acceptable level of quality. Therefore if the metric value reaches the threshold, it indicates that the quality of data is acceptable and if the metric value falls below the threshold it indicates that the quality of data is not acceptable.

It should be noted that verification metrics are defined based on numeric scale (number of…..) and thus the threshold should be a number. Whereas the threshold for validation metric should be a milestone of implementing required rules / processes to maintain the characteristic. Therefore it can be a percentage of the solution implemented by then.

4.3 Synthesising the Meta-Meta-Model for DQ requirements

As per IEEE standard, a requirement is defined as a condition or capability needed to solve a problem or achieve an objective (IEEE 1990). Therefore, it is apparent that a requirement is a hybrid concept that involves a problem and a solution. In section 4.2 we explained that a DQ requirement is the inverse of a DQ problem and DQ characteristics are a central concept to both requirement and problem. In this case, a requirement refers to simply a "user requirement". With the insight gained from IEEE definition for a requirement, it is clear that a DQ requirement should contain a solution to the problem as well. Thus a DQ requirement is a hybrid concept that contains a DQ problem and a solution to solve the problem (Figure 4.3: Hybrid DQ requirement).

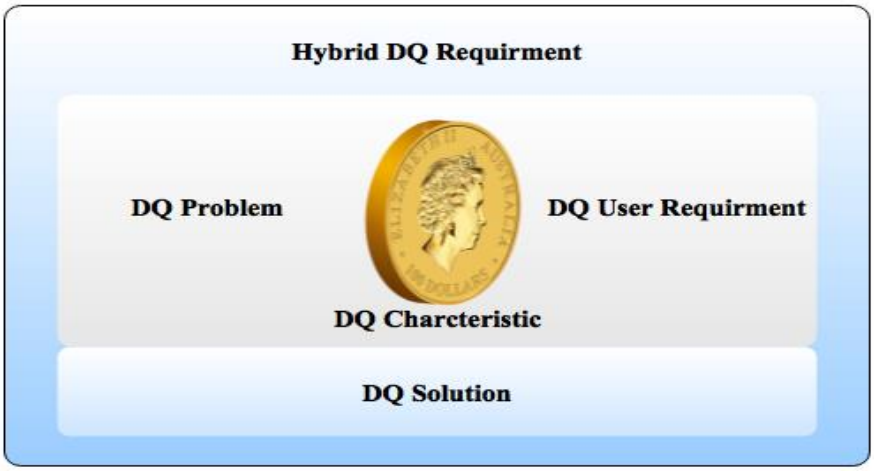


Figure 4.3: Hybrid DQ requirement

We argue that a DQ characteristic along with the other related concepts explained in section 4.2 describe a DQ problem and a solution.

4.3.1 Representation of DQ Problem

A DQ problem can be explained by a **DQ characteristic** since non-adherence to a DQ characteristic is a DQ problem. For example, consider the following DQ characteristic.

Meta-data compliance: data should comply with its Meta-data.

On the one hand, the definition express a user requirement, while on the other hand, the negative form of the definition represents a DQ problem i.e. *data do not comply with its meta-data*. Therefore we posit that DQ characteristics can be used as a representation of a DQ problem.

DQ dimension provides the overarching perspective that a DQ characteristic and thus a DQ problem belongs to. A DQ characteristic belongs to only one DQ dimension.

DQ characteristic type explains the nature of the characteristic (declarative/Usage). On the one hand, the type reveals whether the user involvement is necessary to maintain the characteristic or not while on the other hand, it reveals whether the DQ problems can be detected without user involvement or whether it needs user involvement to detect the problem (section 4.2.1.1). A DQ characteristic has only one type.

Granularity represents a hierarchy of data objects that a DQ characteristic is applicable on, or in other words, the hierarchy of data objects the DQ problem is based on. A DQ characteristic has only one data granularity level

Verification metric is a measurement about how well the DQ characteristic in concern has been maintained with regards to a data object at any given point in time. On the other hand, it provides an indication of the frequency of the DQ problems pertaining to the characteristic. The **verification metric threshold** provides an indication of what is the expected level of quality, or on the other hand the acceptable level of tolerance for DQ problems in the data object with regards to the characteristic in concern. It should be noted that a DQ characteristic has one or more verification metrics and each verification metric has only one threshold.

Therefore we establish that DQ characteristic, DQ dimension, DQ characteristic type, granularity, verification metric, and verification metric threshold represents useful information about a DQ problem.

4.3.2 Representation of a DQ solution

The characteristic type, in turn, explains the **implementation form** of the characteristic. In other words, what form of activities is needed to prevent DQ problems pertaining to the characteristic. Declarative characteristics take the form of rule-based approach where DQ rules will be implemented in the front-end or back-end of the information system to maintain the characteristic. Usage characteristics take the form of process-based approach where necessary DQ processes will be implemented to maintain the DQ characteristic.

DQ rules are database constraints and any kind of automated business rules associated with data creation data manipulation and data processing tasks. DQ processes are specially designed processes to improve the human productivity in data creation data manipulation and data usage.

Validation metric is a measure of the extent to which the solution has been implemented to maintain the DQ characteristic. Solutions are designed and implemented step by step and it takes the time to become a matured solution. Especially the process to improve human productivity takes a certain

amount of time to mature depending on its nature. Therefore the **validation metric threshold** demarcates the percentage of the solution implemented by then.

Therefore, we establish that the concepts implementation form, validation metric, and validation metric threshold represents the solution to a DQ problem.

As pointed out in chapter-2 (section 2.9) in this chapter, every model has a (1) purpose and a role, (2) it contains some embedded knowledge which in turn supports the purpose and role of the model. Hence we should also focus on the purpose, role and embedded knowledge of final models created based on the meta-meta-model of a DQ requirement. In the case of modelling a DQ requirement, the purpose and role are to provide an explicit and in-depth understanding to the stakeholders about the DQ requirement to design solutions to maintain the characteristic in a particular data object. In section 4.3 above we explained how DQ characteristics and other related concepts provide a comprehensive representation of a DQ problem and a solution. Thus they can be used as constructs for the meta-meta-model of a DQ requirement. Therefore the next step is to develop a meta-meta-model for a DQ requirement by using a suitable modelling grammar for the domain (Rosemann and Green 2002). We selected ERM (Entity Relationship Modelling) as the suitable grammar for this task. By using ERM grammar for this meta-meta-model we can maintain its compatibility with data model (ER models) and thus with logical data models (databases) and other tools like data catalogues and data dictionaries where DQ requirements added to

Figure 4.4 shows the assimilation of the above constructs in ER grammar, which is the proposed meta-meta-model for a DQ requirement. The meta-meta-model contains two representations relevant to DQ management.

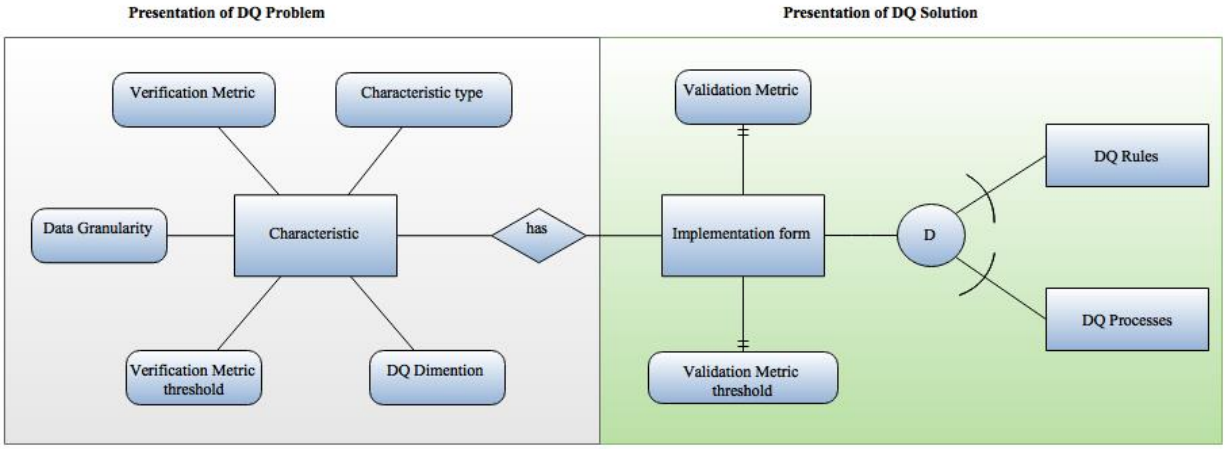


Figure 4.4: meta-meta-model for DQ requirements modelling

4.4 Towards a meta-model for a DQ requirement

As per OMG (Omg 2008), a meta-model is an instantiation of a meta-meta-model (see 2.3.2 for more details). Therefore, the meta-meta-model described in Figure 4.4 has to be instantiated to develop a meta-model. For this instantiation, we need to get instances for each concept used in the meta-meta-model. Thus in the next chapter, we investigate into each concept and develop a shared understanding about instances of each concept systematically so that a meta-model for DQ requirements can be developed.

4.5 Chapter Summary

In this chapter first, we pointed out that data quality requirements can be expressed using data quality characteristics and ten other domain concepts viz. DQ characteristic type, Data granularity, DQ dimension, Implementation form (DQ rules, DQ processes), verification metrics, verification metric threshold, validation metric and validation metric threshold. Second, we developed a meta-meta-model for a DQ requirement by synthesising the domain concepts using an appropriate modelling grammar of ERM

Chapter 5

REFACTORING DATA QUALITY DIMENSIONS

5.1 Overview

In this chapter, we present an analysis of published data quality dimensions available in the literature. Through a systematic review of research and practitioner literature, we identify previously published classifications of data quality dimensions. Then conduct an analysis and consequent consolidation, addressing overlapping and inconsistent definitions. Lastly, we refactor the concept of data quality dimension as per the meta-meta-model introduced in chapter 4, thus producing instances of DQ requirements. Further, we report on a card sorting study that was conducted to evaluate the consolidated definitions empirically.

5.2 The curse of dimensionality in data quality

In chapter 4, we revealed that there was an ambiguity in terminology used to refer to the concept of DQ dimension. Therefore, we established that DQ characteristics as the central concept in defining data quality, which is in turn related to the concepts metrics, granularity and implementation form in defining a DQ requirement. In this context, we used the term DQ dimension as a higher level abstraction of DQ characteristics used for management purposes.

Over the last two decades, researchers and practitioners have suggested several classifications of data quality characteristics. For example, the classifications of Wang and Strong (1996), Redman (1997), English (2009), Loshin (2001), Price and Shanks (2005b), Stvilia et al. (2007). However, over the course of time, many of the definitions for different data quality characteristics have overlapped, and same definitions for the same characteristic has developed conflicting interpretations. Thus, we observe that DQ characteristics have been discussed by many authors and have regressed into a level of disparity that does not support a shared understanding of the core knowledge of the discipline (Jayawardene et al. 2013b)

Despite the numerous classifications, only a few studies have embarked on a consolidation of these diverse viewpoints. Among those, Scannapieco and Catarci (2002) consider six classifications of DQ characteristics and discuss the correspondence among characteristics that share the same name. Eppler (2006) provides an analysis of several classifications of data quality characteristics and recognizes sixteen mutually exclusive characteristics. While useful, the coverage of the study does

not consider a balance between academic and practitioner contributions hence the consolidation is incomplete. Further, the basis for selection (or exclusion) of the classifications is not clear.

Shared understanding is an essential prelude in conceptualisation and therefore, to develop DQ patterns, it is essential to building a shared understanding of this central concept. In this chapter, we undertake a study of existing classifications of data quality characteristics. Importantly, our study spans both academic as well as industry practitioner contributions. We believe that such an analysis is essential to create a shared understanding of the data quality characteristics to define DQ patterns.

In section 5.3 we explain our approach for creating a consolidated set of data quality characteristics and in section 5.4 the approach and results of the empirical validation undertaken to evaluate them. Finally, in section 5.5 we present the validated set of characteristics followed by a summary in 5.6

5.3 Approach for consolidation of Data Quality Characteristics

This study relies on a systematic literature review and qualitative research methods. This section outlines the approach followed in selecting the appropriate materials for the study; analyzing and consolidating the data quality characteristics, and the process of empirically validating the outcome.

5.3.1 Selection of sources

In qualitative research, there is no strictly defined sample size (Baum 2003; Patton 2005) while small samples are usually selected since the aim is to perform an in-depth and a detailed study (Miles, Miles, and Huberman 1994; Patton 2005). Coyne (1997) and Patton (2005) describes that all sampling done in qualitative research are "purposeful sampling". As per Patton (2005), Purposeful sampling is very powerful since it leads to information-rich cases for an in-depth study. Marshall (1996) refers to purposeful sampling as "judgement sample" where he describes as the most common and most intellectual strategy used in academia based on the researchers' practical knowledge of the research area (Marshall 1996).

We are motivated to ensure broad consideration of data quality literature from both academic research as well as contributions from industry practitioner community. Thus we selected a purposeful sample of sources considering five different perspectives:

- Industry practitioners who are prominent in the industry, have been involved in large data quality projects and have contributed to the DQ body of knowledge by publishing. Relevant sources within the practitioner perspective were identified by examination of citations in

public forums and professional training programs by professional bodies such as DAMA4 and IAIDQ5. Within these sources we identified the following contributions as a representative set: (Redman 1997), (English 2009), (McGilvray 2008), (Loshin 2001), (Kimball and Caserta 2004).

- Market leaders of DQ management tools, as identified by Gartner's Magic Quadrant (Friedman 2012). These include SAP (Gatling 2007), IBM (Byrne 2008), and Informatica (Loshin 2006).
- Data Quality standards, as identified by ISO 8000 - a standard for data quality (ISO 2012).
- Organizations that have recognized the importance of DQ, and developed in-house frameworks for DQ management. Although many organizations conduct DQ projects, few make relevant material available publicly with a sufficient level of information suitable for analysis, which is a limitation of considering this perspective. In our search, however, we identified the Bank of England (Lyon 2008) and Health Information and Quality Authority (HIQA 2011), the latter representing an international study on DQ practices of health care organizations in England, Wales, Canada and New Zealand.
- Academic research covering diversified viewpoints about DQ characteristics: In our earlier work (Sadiq et al. 2011c) we analyzed DQ research contributions over the last two decades and created a bibliographic database of over 1400 publications. We used this resource to identify publications that focus on data quality dimensions/characteristics/criteria using a keyword search. Consequently, we identified 36 publications with a sufficiently deep focus on DQ characteristics. Based on citation analysis, we observed that the most prominent classification of DQ characteristics was developed by Wang and Strong (1996), with the majority of subsequent classifications using a subset of these originally defined characteristics based on the contextual needs. On this basis, we selected the original work by Wang and Strong (1996) and three other classifications (Eppler 2006; Scannapieco and Catarci 2002; Stvilia et al. 2007), all of which are consolidation efforts of other classifications available in the literature to achieve an adequate coverage from academic research. Further, we also selected (Price and Shanks 2005b) a classification based on the semiotic perspective of data, which provides a contrasting view compared to other classifications that tend to be based on the product perspective of data. Since the purpose of the analysis is to get diversified

[4] Data Management Association (DAMA) http://www.dama.org.au/
[5] International Association for information and data Quality http://iaidq.org/

viewpoints about DQ dimensions, therefore contrasting viewpoints were sought and included as relevant

Based on the above perspectives we selected 16 publications, which represent a broad and diverse scope of the analysis.

5.3.2 Analysis of Data Quality Characteristics

In the first stage of the analysis, the 16 publications (or relevant parts thereof, in the case of books) were loaded into NVIVO[6]. The text was reviewed and individually coded by two researchers. Each coder independently coded the relevant text in NVIVO, creating a node for each DQ characteristic and its definition. The coding structures were then consolidated between the two researchers to arrive at a final coding that identified 129 distinct terms as DQ characteristics, after resolving coding disagreements through discussion. For some of the DQ characteristics, there were multiple definitions given by different authors so that altogether there were 189 definitions found for the 129 representative terms which we identified as DQ characteristics. It became clear that these terms and definitions had some overlaps and conflicts. Some authors have used the same term (as a DQ characteristics) to refer to contrasting aspects of data quality while some other authors have used different terms to refer to the same/similar aspect of data quality. Hence, it was apparent that there were many common themes among the 189 definitions, and a need arose to consolidate these definitions to reach a consensus. Accordingly, the next step was to classify these definitions into thematic clusters to produce a set of consolidated DQ characteristics.

In section 4.2.1 we established that every DQ characteristics is either a declarative characteristic (D) or a usage characteristic (U). Further, we explained that the concept of DQ characteristics is related to data granularity, implementation form and different types of metrics. Hence, our next task was to apply this underpinning to each of the 189 definitions and consolidate them into a clear set of themes.

Thematic analysis is fundamentally focused on identifying themes in qualitative data, and it is the most common form of data analysis in qualitative research (Guest et al. 2011). Braun and Clarke (2006) suggested six steps to perform a thematic analysis from the scratch using raw data. These steps include (1) Familiarizing with data, (2) Generating initial codes, (3) Searching for themes, (4) Reviewing themes, (5) Defining and naming themes and (6) Producing the report.

[6] NVIVO is a qualitative data analysis tool designed for analyzing rich text-based and/or multimedia information, where deep levels of analysis of data are required. http://www.qsrinternational.com/products_nvivo.aspx

In our case, each definition of DQ characteristics explicitly referred to a specific aspect of data quality and, hence constituted a theme itself. Hence, step two and three could be combined.

It should be noted that the ultimate purpose of this analysis is to develop a shared understanding of DQ characteristics to represent a DQ requirement and in chapter 4, we have established the domain concepts required to represent a DQ requirement. Hence, in this analysis, we also considered these domain concepts in coding the definitions to developing themes that conform to the meta-meta-model developed in chapter 4.

Therefore first we coded each definition for the following aspects:

1) The main theme of the definition that is the DQ characteristic described in the definition.
2) Type of the DQ characteristic (Declarative/Usage)
3) The applicable granularity level of the characteristic (Element, Record, Information object)
4) Implementation form (By design as rules / By conformance to processes)
5) Possible metrics for the characteristic

The above coding process was performed using NVIVO and all the coded information was organized under each node created for each original definition for data quality dimension. Two researchers individually performed this analysis to avoid researcher bias in coding. In particular, for each theme/characteristic, the researchers individually coded the definitions as being related to a usage perspective (U), a declarative perspective (D) or neither (X). The aim of this task was to refine the list of DQ characteristics by eliminating those that do not represent characteristics of data for quality assurance purposes. The rest of the details, i.e. granularity level, implementation form and possible metrics, were also recorded by the researcher for the next phase of the analysis. Sometimes these details were not explicitly mentioned in the definition given by the authors themselves, and the researchers had to use the other related explanations the authors had provided with the definition to elicit the details.

The independent ratings of D/U/X assessment were evaluated using Cohen's Kappa, with a result of 0.81, indicating high confidence of rater agreement (Carletta 1996). Coding disagreements were then discussed between the two researchers until consensus was reached. In this analysis, out of 189 candidate themes/characteristics, only three did not fall into either a declarative or a usage perspective, indicating that they are neither characteristics of data itself nor a view on data usage. These are 'Efficient use of memory' and 'Use of storage' (defined in Redman (1997) and Loshin (2001) respectively). These definitions are focused on the utilization of disk space and memory space

of computers, ‘Stewardship’ (Loshin 2001), which is focused on assigning the people responsibility for data, and represents a management function rather than a declarative or usage perspective of data quality.

As a result of this step, we identified 186 candidate themes. The next step was to create a thematic map considering the conflicts and overlaps of each theme. In this task, two researchers independently consolidated the themes and created two thematic maps. These thematic maps consist of similar candidate themes clustered together which formed main themes. For example, the following 4 candidate themes were clustered together to form a theme for a DQ Characteristic.

Candidate themes:

1) *A given data element has a full value stored for all records that should have a value.*
2) *Data element is always required to be populated…….*
3) *Completeness refers to the expectation that certain attributes should have assigned values in a data set.*
4) *Determine the extent to which data is not missing. For example, an order is not complete without a price and quantity.*

Theme of a DQ Characteristic:

Mandatory attributes cannot be null

In creating the thematic clusters that represent DQ characteristics, in addition to the meaning of the candidate theme, we considered the available details to judge the characteristic type, data granularity, implementation form and possible metrics pertaining to each candidate theme. For example in the above four candidate themes the data granularity explained in the context is data element (E). The characteristic type explained by each candidate theme is declarative (D) as they describe an inherent characteristic of data which is a part of the definition of the data element. Therefore the rule based approach is considered as suitable for the implementation form of each theme and, also the metrics could be defined referring to the number of violations or exceptions to the rule. Since all the above aspects were common for each candidate theme, a decision was made to consolidate the four themes into a single DQ characteristic called completeness of mandatory attributes as above.

These DQ characteristics were then further clustered into higher level themes (or DQ Dimensions) and the consolidation continued up to two levels. For example the following four themes of DQ characteristics were clustered together into a DQ Dimension called *completeness* of data.

DQ Characteristics themes:

1) *Mandatory attributes cannot be null*
2) *Validity of null values in optional attributes.*
3) *Record completeness of entities*

Higher level theme: *Completeness*

Once the two researchers individually created the two thematic maps, the two researchers jointly reviewed their maps for the differences. Then the two researchers reviewed the individual thematic maps and created a consolidated thematic map by resolving the differences and disagreements of individual maps through discussions. At this stage, the two researchers created the thematic maps for two levels which consist of 33 distinct themes categorized into eight higher level thematic clusters. These 33 themes were considered as characteristics of data and the eight main clusters that they belong to, were considered as dimensions of data quality.

Once the dimensions and characteristics were identified, the next task was to define and name them. A group discussion was conducted where every theme in the consolidated thematic map was discussed and debated for a more meaningful definition and a representative term. Initially one researcher proposed a meaningful definition and a representative term for each characteristic considering its candidate themes, while two other researchers suggested improvements. The definition and name for each characteristic was finalised once the three researchers were satisfied about the final definition and the representative term arrived. For example, in the above example (*Mandatory attributes cannot be null),* the following definition and the representative term were developed.

Definition: *The attributes which are mandatory for a complete representation of a real world entity must contain values and cannot be null.*

Representative term: *Completeness of mandatory attributes*

Further based on the underlying characteristics, a representative name was identified for each high level theme (cluster of DQ characteristics) that is DQ dimensions. The same process with the three researchers was followed in deciding on the representative terms. A summary of the eight dimensions and the characteristics of each dimension is shown in Figure 5.1. The complete list of candidate themes for each characteristic in the consolidated thematic map is given in Appendix-D.

5.3.3 Methodology for validation of findings

We were motivated to evaluate our classification to ensure that the themes and respective definitions are both clear and representative. Couture (1986) explains the concept of effective ideation of written text. In this study, she revealed that the effective communication of ideas to readers depends on two meaning systems. They are logical meaning, which is realized in a discourse's propositional content (the substance of the discourse), and semiotic meaning, which is realized in the discourse's reference to meaning systems of language (linguistic choices of expression). The combination of these two kinds of meaning enables a text to convey ideas to its readers. Based on this idea we argue that our definitions will communicate effectively to the users depending on the substance they deliver and the linguistic choices that we use to deliver the substance. Hence, we conducted a validation of the definitions and the names of each characteristic for their substance (definitions of characteristics), and linguistic choices in representing the name of the characteristic.

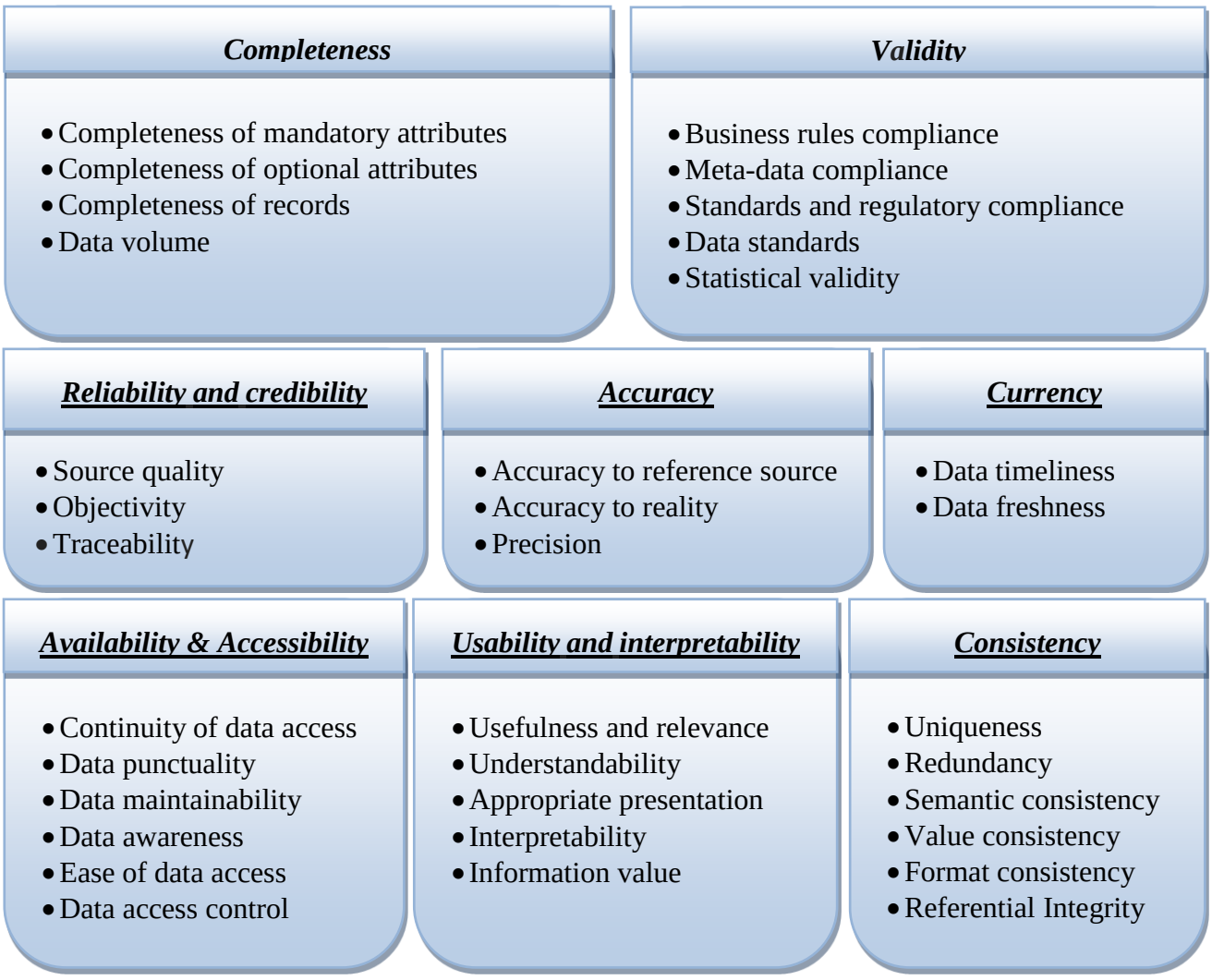


Figure 5.1: DQ dimensions and characteristics

Therefore, the validation process aimed to achieve two main goals:

1. Clarity of the definitions: To verify that the characteristic definitions represent the underlying candidate themes used to create them and the users understand them by reading the definition.
2. Clarity of the names: To verify that the names of the characteristics are adequate representations of the characteristic definitions.

For example, the dimension "completeness" has a characteristic called "completeness of records" with a developed definition "Every real world entity instance that is relevant for the organization can be found in the data". As explained in section 5.3.2 the candidate themes used in deriving this characteristic and definition were,

"Every real world phenomenon is represented in the database without omission."

"Data is complete if no piece of information is missing, anti-example: The Beatles were John Lennon, George Harrison, and Ringo Starr."

"A record exists for every real world object or event; the enterprise needs to know about."

"Monitoring for incomplete lists of eligible records or missing data items."

In the first goal, we focus on evaluating the clarity of the definitions of characteristic, and if that definition is a fair representation of the four candidate themes used to create this definition. In the second goal, we focus on evaluating if the characteristic name (e.g. data freshness) is a fair representation of its given definition.

A card sorting study was chosen as the method for our evaluation. Card sorting studies are generally used to validate constructs in research instruments (Moore and Benbasat 1991). In practice, they are also a popular method of developing navigation structures in websites, workflows and menu structures, where understanding user experience is required to achieve an optimum and natural design from a user point of view (Nawaz 2012).

In our case, the goals of our validation relate to the meaning of the definitions and terms - hence, they are related to users' cognition. In cognitive science, asymmetric dependence theory (Fodor 1987; Fodor 1990) explains how the human mind represents meanings of objects and symbols (words of a language). Fodor (Fodor 1987; Fodor 1990) argues that meanings depend on upon what is represented in the mind, or more specifically the *token* that a word creates. In other words, the tokens created initially in the mind govern the meanings of subsequently encountered things and not vice versa.

Similarly, in a card sorting study when the judges read the categories (characteristic definitions) first, each category will create a unique token, which is the meaning of the definition. When they read the cards (candidate themes) subsequently, the cards also create tokens and as per asymmetric dependency theory; these tokens depend on the tokens created by the categories and not vice versa. Hence, we can argue that a judge can place the right card into the right category only if the token created by the category (the meaning to the judge) is a unique token and the token created by the card has a dependency to the category's token that dominates the whole process of selection.

Based on this notion, we stipulate that the judges can sort the cards successfully into categories only because they can understand the definitions of characteristics and also these definitions represent the candidate themes (as they have selected the right card into the right category). If the judges fail to sort the cards successfully, it can be because the judges cannot understand the definition or the definition is not a good representation of the cards. In this case, we discussed the outcome with the judges and performed the required amendments to the definitions. Then the amended definitions were used for the second round of the card sorting study to finalize the definitions.

We used two card-sorting studies to achieve our two validation goals. In the first card sorting study, we focused on the first goal. Hence, we considered the definitions of the characteristic as the categories and the candidate themes as the cards and asked the judges to sort the cards into categories. In the second card sorting study, we focused on the second goal, and we considered the characteristic definitions as the cards and the characteristic name as the categories. Based on the results, we refined the terminology of the characteristic name.

In literature card sorting results have been analyzed and interpreted using the following two measures in conjunction (Alsaghier et al. 2011; Li et al. 2006; Moore and Benbasat 1991; Nahm et al. 2002),

(1) Inter-rater agreement-Kappa coefficient and

(2) Item Placement ratio.

Kappa coefficient is a statistical measure of agreement for multi-rater categorization of nominal variables which provides an indication of the reliability of expert opinions (Fleiss and Cohen 1973; Randolph 2005). Hence, in a card sorting study Kappa can be used to measure the overall agreement between judges in placing the cards in general across all the categories. On the other hand item, placement ratio is an indicator of how many items were placed in the intended or target category (Moore and Benbasat 1991). Hence, both these measures together provide confidence in the accuracy of placement and the reliability of the decision of placements. In other words by using both these

measures in conjunction, we can assure that judges have placed the cards in the right category consistently.

In our study, a low item placement ratio indicates weak category definitions (characteristic definitions) where the judges could not place the cards correctly, as they get low comprehension. In other words, when the clarity of the definition of a characteristic is weak, there is a tendency that the judges tend to place the cards giving a low item placement ratio. While there is no standard acceptance threshold for item placement ratio, Moore and Benbasat (1991) considered 70% and above as an accepted level in finalizing factors for a survey instrument. In contrast, in our study, the judges are involved in rather a subjective cognitive task of evaluating the comprehensibility of definitions and terms. Therefore, we used a threshold of 80% for item placement ratio to ensure the required minimum (70%) is well exceeded in light of a lack of consensus on the minimum.

Kappa index was calculated using an online Kappa index calculator (Randolph 2008). This Kappa calculator provided two Kappa values, fixed marginal Kappa - when the raters know the number of cards that should be allocated to a specific category and free-marginal Kappa- when there is no specific limit to the number of cards that can be allocated to a specific category. Hence, the researchers have to select the appropriate kappa index depending on the design of the card sorting study. In literature, Kappa value can range from -1.0 to +1.0 and there is no standard Kappa indicator for the inter-rater agreement. Previous research has shown consensus among researchers interpreting Kappa coefficient, against the following ranges: <0.00 Poor agreement (less than chance agreement); 0.00 – 0.20 Slight agreement; 0.21-0.40 Fair agreement; 0.41-0.60 Moderate agreement; 0.61-0.80 Substantial agreement; 0.81-1.00 Almost Perfect agreement (Castillo et al. 2006; Landis and Koch 1977; Ng et al. 2012). We consider the range 0.61-0.80 as our accepted threshold to conclude that there is substantial agreement among the judges in placing the cards into categories.

To achieve the first validation goal, for each dimension, a closed card sorting study was designed so that the corresponding DQ characteristics were considered as the predetermined categories and the individual definitions given by different authors were considered as the cards. The study was performed using Optimal Sort[7], and it was designed to be conducted through several iterations of online surveys until the definitions were refined. In the first iteration, there were eight separate surveys for the eight dimensions. In each survey, the characteristics belonging to the specific

[7] Optimal Sort is an online card sorting tool and it provides a user friendly interface to naturally sort the cards into categories https://accounts.optimalworkshop.com

dimension were considered as categories and the candidate themes that contributed to creating the characteristics' definitions were considered as cards.

To validate the second goal, for each dimension, a closed card sorting study was designed. In this study the definitions of the characteristics were considered as cards, and the proposed characteristic name was considered as the category.

Moore and Benbasat (1991) consider two participants as being sufficient for a card sorting study and in their survey; they had used four judges for better representation of expert opinions. Hence, in both card-sorting studies we used a minimum of four judges, all of whom were Ph.D. students researching data quality and information systems. The choice of using PhD students working in the area of data quality and information systems was made in order to utilize their expertise and understanding about the domain to make this validation credible and authentic.

5.4 Validation of DQ characteristics

As a result of the thematic analysis, we arrived at eight main clusters of data quality definitions that are termed as data quality dimensions and within each dimension, there are multiple DQ characteristics pertaining to the dimension. Figure 5.1: DQ dimensions and characteristics, shows these DQ dimensions and DQ characteristics. In this section, we present the empirical validation of DQ characteristics based on the methodology mentioned in the section 5.3.3.

5.4.1 Validation of goal 1: clarity of the definitions

A card sorting study was performed for each data quality dimension using five judges having characteristic definitions as categories and the candidate themes as cards. In data analysis for each dimension, we calculated Kappa index and item placement ratio at dimension level (total item placement ratio) as well as individual characteristic level. This allowed us to investigate the placement of items in more detail at characteristic level. The summary of the first round of card sorting for goal one is presented in Appendix-E (Table 1).

In this case, we used free-marginal Kappa since the number of cards for each category was not fixed. In this analysis out of the eight dimensions, seven of them received total item placement ratio above 80% and Kappa index above 0.61, which is within the acceptable range (see section 4.4.3). The only exception is the dimension "Currency" which received total item placement ratio of 76% with 0.3 Kappa index that is below the acceptable range. Also, it should be noted that the individual item placement ratio for the two characteristics in this dimension are 74.29% and 77.5% that is below the acceptance range of 80%. Hence, from this analysis, we conclude that the first goal stands for the seven dimensions, "Completeness," "Availability and Accessibility," "Accuracy," "Validity,"

"Reliability and Credibility," "Consistency" and "Usability and Interpretability." However, for the dimension "Currency", the first goal was not achieved, and another round of card sorting was conducted.

While total item placement ratio describes how the cards were placed within each dimension, the individual item placement ratio describes how the cards were placed on individual characteristics. It should be noted that for some dimensions, even though the total item placement ratio is above 80%, some characteristics had an individual item placement ration less than 80%. Hence, it is rational to investigate into these cases to determine if there are clarity issues of the definitions of the characteristic or, else, if the judges have placed the cards into a different category due to an ambiguity in the card definition. Hence, we conducted interviews with the judges to justify reasons for these individual cases.

The dimension "Availability and accessibility" received 82.22% for total item placement ratio and 0.64 for Kappa index while the individual item placement ratio of the two characteristics "Data punctuality" and "Ease of data access" were 76% and 73.33% respectively, which is below the acceptable range considered in this analysis. The two definitions of the respective characteristics in concern were:

Data punctuality: Data should be available at the time of its intended use.

Ease of data access: Data should be easily accessible in a form that is suitable for its intended use.

In raw data, it was evident that some cards which should go into the category "punctuality" were placed into "ease of access" which reduced the individual item placement ratio of data punctuality. In the interview with judges, we found that the cards "*Information is accessible when it is needed"* and *"The characteristic of getting or having the information when needed by a process"* were wrongly classified into "ease of data access." The reason was that the term *"when needed"* did not impress a punctuality aspect, but rather related to accessing data. Similarly, the cards "*Data are available or easily retrieved*" and "*Ease of obtaining an information object relative to a particular activity*" resulted in some ambiguity as they were classified into the characteristic "punctuality". Despite this ambiguity in the cards, the experts agreed on the two definitions given for the characteristics as clear and comprehensive. Hence, the two definitions were accepted without any change.

The two characteristics belong into "Currency" received an individual item placement ratio of 77.5% and 74.29%, which further supported the necessity for having a second card sorting round for the

dimensions. During the interviews with judges, we found that it was hard for them to distinguish between the two phrases used in the definitions "data which refers to time" and "data which is subjected to changes over time." The definitions were:

Data age: Data, which refers to time, should be available for use within an acceptable time relative to its time of creation.

Data Freshness: Data, which is subjected to changes over the time, should be fresh and up-to-date with respect to its intended use.

As a result, two new definitions were proposed, and a decision was made to perform another round of card sorting with the new definitions to check if the new definitions confirm to the first goal.

For the second round, we selected four judges who have expertise in data modeling, data-driven applications and emergency response systems where data currency played a major role. The summary results of the second round of card sorting for goal one is provided in Appendix-E (Table 2)

The results of round two illustrate that total item placement ratio, Kappa, and individual item placement ratios are well within the acceptability range and hence it convinced that new definitions conform to goal 1.

5.4.2 Validation of goal 2: clarity of the names

For this validation, a card sorting study was performed for each dimension having the names of the characteristics as categories and the definition of characteristics as cards. In this study, every category had one and only one card and hence in analyzing the results; fixed-marginal Kappa was used instead of free-marginal Kappa since the number of cards for each category was fixed.

We selected five judges who were Ph.D. students researching data quality and information systems. In addition to their expertise in the domain, language proficiency was also considered in selecting them for the study. The results were analyzed based on the same parameters as in the previous validation except that we used fixed marginal Kappa. The summary results of the first card-sorting round for goal two is provided in Appendix-E (Table 3).

Our analysis showed that all the dimensions have received an acceptable total item placement ratio and Kappa index, except for "currency". For "currency", none of the judges have placed the cards in the right category hence the total item placement ratio as well as the individual item placement ratio was 0%. Also, their agreement about the decision was unanimous, which was indicated by Kappa index of 1%. Hence, we selected different names for each definition to improve their representation. The two definitions and the names used under currency are:

Data age: Time related data should be available for use within an acceptable time relative to its time of creation.

Data freshness: Data should be kept up-to-date if it is subjected to a natural process of obsolescence over time.

Having discussed the results with the judges, we found that they saw a connection between the name "Data age" and the phrase "process of obsolescence over time" (used in the latter definition), which influenced their classification. Hence, we selected the name "Data timeliness" to replace "Data age" after consulting the judges. To check the clarity of the new names introduced, we performed another card sorting study using a different set of judges. Five judges were selected using the same criteria used as above and the summary results of the second card-sorting round for goal two are shown in Appendix-E (Table 4).

In this round, the results were very positive having 100% for item placement and a Kappa of 1. Hence, "Data timeliness" and "Data freshness" were concluded as the respective names for the two characteristics.

5.5 Validated DQ dimensions and characteristics

Following the first level classification and clustering, eight main clusters were identified, *viz.* Completeness, Availability & Accessibility, Currency, Accuracy, Validity, Usability & Interpretability, Reliability and Credibility, and Consistency. In the following discussion, each dimension and the various characteristics within the dimension is presented, along with the finalized definitions and representative terms that resulted from the empirical validation.

In Chapter 4 we established the concepts that define a DQ requirement where the DQ characteristic is the central concept. The other related concepts, (implementation form, verification metric, validation metric, and threshold) are determined by the type of DQ characteristic except for the granularity level. Then we provided formal definitions for them in section 4.2. Therefore, in the following discussion we present only the DQ characteristic, the characteristic type, and the granularity while other concepts can be defined as per the formalization accordingly.

5.5.1 Completeness

Completeness of mandatory attributes: The attributes which are mandatory for a complete representation of a real world entity must contain values and cannot be null (Element | Declarative)

In representing an entity, certain attributes are essential to convey its full meaning. For example, the title of a book is a mandatory attribute in representing a book even though the primary key attribute

of the entity book may be some other attribute like book id. The necessity of having mandatory attributes (and optional attributes) has been discussed under the terms "Value completeness" (English 2009), "Completeness" (Gatling 2007; Loshin 2001; Loshin 2006) and "Complete" (Byrne 2008). As specified by its definition the characteristic is applicable on element level and type of the characteristic is declarative since it user independent and can be maintained by implementing mandatory rules at database level or application program level to prevent null values entering in the attribute in concern.

Completeness of optional attributes: Optional attributes should not contain invalid null values (Element | Declarative)

Provided that all mandatory attributes contain some value other than null, the next question is the null values contained in optional attributes. In general, a null value is not a desirable occurrence in a database, but on the other hand, not all null values are invalid occurrences. For an example, the maiden name of unmarried women is not defined for that data instance and hence a null value is acceptable. However, it is important to recognize the unacceptable null values and treat their root causes accordingly. Since the characteristic is focussed on defining data elements, it is independent to users and, hence its type is declarative. Required rules can be implemented at data element level to detect null values in optional attributes and prompt for action. Previous works on the problem of having null values for data elements include, "ability to represent null values" (Redman 1997), "null values" (Loshin 2001), "representation of null values" (Loshin 2001), and "completeness" (Redman 1997).

Completeness of records: Every real world entity instance that is relevant for the organization can be found in the data (Record | Usage)

While the existing records in a table may be complete as far as the mandatory and non-mandatory attributes are concerned, a missing record, i.e. a required entity not being recorded in the database, is a well-recognized problem. It is evident from "completeness" (HIQA 2011; Kimball and Caserta 2004), "mapped completely" (Price and Shanks 2005b) and "record existence" (English 2009). Even though it is hard to find out what is missing, control processes have to be implemented to prevent omission of records. Further, it is necessary to establish a mechanism to escalate issues regarding missing records so that root causes can be identified and treated accordingly.

Data volume: The volume of data is neither deficient nor overwhelming to perform an intended task (Information Object | Usage)

The volume of data has to be managed in such a way that the organization does not maintain too much or too little information for a task at hand. Too much information may lead to unnecessary maintenance costs whereas too little information may not allow satisfactory completion of the task at hand. “Comprehensiveness” (Eppler 2006) “Data Coverage” (McGilvray 2008) and “Appropriate amount of data” (Wang and Strong 1996) all discuss this characteristic of data quality. Therefore necessary processes have to be implemented to ensure the required data volumes are maintained for each task, where data volume is critical in performing the task.

5.5.2 Availability & Accessibility

Continuity of Data Access: The technology infrastructure should not prohibit the speed and continuity of access to the data for the users (Information Object |Usage)

Continuous and unobstructed accessibility is essential for efficient usage of data. Facilitating the accessibility of data through an appropriate technology infrastructure has been previously discussed as: “Accessibility” (Eppler 2006), “Speed” (Eppler 2006), “Reliability” (Scannapieco and Catarci 2002) and “Accessible” (Price and Shanks 2005b). Based on the task, required capabilities should be maintained in the technology infrastructure to ensure the speed and continuity of data access.

Data maintainability: Data should be accessible to perform necessary updates and maintenance operations in its entire lifecycle (Record |Usage)

Once the data is created, it should be maintained as required through updates, consolidations, enrichments, etc. to assure its usability. Prior work on “ease of use and maintainability” (McGilvray 2008), and “maintainability” (Eppler 2006) emphasize having a systematic process for maintaining data.

Data awareness: The data users should be aware of all available data and its location (Information Object | Usage)

When the necessary data is available, stakeholders should have the knowledge required to find the necessary information objects that suit the task at hand. Prior investigations of “accessibility” (HIQA 2011) provide an insight into this aspect. Data awareness of users has to be improved on a continuous basis through establishing the necessary capabilities and mechanisms.

Ease of data access: The data should be easily accessible in a form that is suitable for its intended use (Information Object |Usage)

Stakeholder information requirements should be thoroughly analyzed to ensure necessary data is provided and does not require further processing before it can be used for a task at hand. Prior work

on “Accessibility” (Stvilia et al. 2007; Wang and Strong 1996) and “Accessibility and clarity” (Lyon 2008) emphasizes this aspect. Therefore, the information systems should have the capabilities to cater the data users’ needs.

Data Punctuality: Data should be available at the time of its intended use (Information Object |Usage)

The required data should be available on time, and otherwise, it may result in a failure or an underperformance of the tasks. “Accessibility timeliness” (English 2009), “Availability” (English 2009), “Timeliness and punctuality” (Lyon 2008) and “Timeliness” (Eppler 2006; Loshin 2006) emphasize this aspect. Therefore, it is necessary to establish a process to ensure that critical tasks receive data on time.

Data access control: The access to the data should be controlled to ensure it is secure against damage or unauthorized access (Information Object |Usage)

It is essential to ensure that data is not distorted or damaged due to usage. Thus, potential threats and risks to the data need to be continually evaluated, and access control should be implemented. The definitions of “Access Security” (Wang and Strong 1996), “Secure” (Price and Shanks 2005b) and “Security” (Eppler 2006; Stvilia et al. 2007) emphasize this security aspect.

5.5.3 Currency

Data timeliness: Data which refers to time, should be available for use within an acceptable time relative to its time of creation (Record |Usage)

Some concepts of the real world are captured with reference to time (e.g.: exchange rates, stock prices). Thus a time stamp is required in generating and using such data. The timing of the activities that generate such data, and valid period of using such data, need to be specified and systematically enforced among data users since data become meaningless without the time stamp. Prior work “currency” (English 2009; Stvilia et al. 2007), “timely” (Byrne 2008; Price and Shanks 2005b), “Volatility” (Stvilia et al. 2007), “Timeliness” (Gatling 2007; HIQA 2011; Wang and Strong 1996), and “Timeliness and availability” (McGilvray 2008) emphasize this aspect.

It should be noted that there is a difference between data timeliness and data punctuality. Data timeliness is defined only for time-related data, and its granularity level is recorded, whereas data punctuality is defined for non-time related data and its granularity is information object.

Data freshness: Data which is subjected to changes over the time should be fresh and up-to-date with respect to its intended use (Record |Usage)

Some data attributes have permanent values (e.g. date of birth) whereas others are subject to changes over time, based on events in the real world (e.g. phone number). While such data does not refer to time (as mentioned in data timeliness above), it can become obsolete over the time unless appropriate updates are carried out. This natural process of obsolescence of data is emphasized in "Currency" (Eppler 2006; Loshin 2006; Redman 1997), "Timeliness" (Kimball and Caserta 2004), "Currency/Timeliness" (Loshin 2001) and "Data Decay" (McGilvray 2008). Therefore, a systematic process should be implemented to ensure that data values represent the current reality.

5.5.4 Accuracy

Accuracy to reference source: Data should agree with an identified source (Element | Usage)

Given that data is a representation of reality, every data element should have a reference source. In the process of data creation, appropriate mechanisms should be employed to ensure data is captured from its reference source free of errors. Involvement of users (data entry staff) is necessary to maintain this characteristic unless the data entry is fully automated. Therefore, we consider this as a special case of a usage characteristic. Prior work, "Accuracy" (HIQA 2011; Loshin 2001; Loshin 2006; McGilvray 2008; Redman 1997; Wang and Strong 1996), "Accuracy to surrogate source "(English 2009) has emphasized the importance of specifying the most suitable reference source for a data element and how to effectively and efficiently capture the data from its source.

Accuracy to reality: Data should truly reflect the real world (Record |Usage)

It should be possible to trace the data to its corresponding real-world entity to support tasks in the real world (e.g.: checking off packed goods against an invoice). Prior work has emphasized that data should have an unambiguous one-to-one mapping with real world objects, as per: "Accuracy" (Eppler 2006) and "Accuracy to surrogate source" (English 2009).

Precision: Attribute values should be accurate as per linguistics and granularity (Element | Declarative)

Data values should be precise to the right level of granularity (e.g.: weight to the nearest tenth of a gram) in the case of quantitative measurements, and they should convey lexically, syntactically and semantically correct statements for text data. This idea is conveyed by "Precision" (English 2009), "Correctness" (Kimball and Caserta 2004), "Accuracy" (Stvilia et al. 2007) and "Correctness" (Eppler 2006). Data precession rules should be implemented to both numeric and text data so that imprecise data values will not be entered into data elements.

5.5.5 Validity

Business rules compliance: Data must comply with business rules (Element | Declarative)

Some attribute values are derived from business rules. Business rules can also be triggered based on attribute values. Accordingly, most business rules have an impact on data, which makes conformance to business rules an important aspect. Prior work on "Business rule validity" (English 2009), "Derivation validity" (English 2009), and "Integrity" (Gatling 2007) emphasizes this aspect. Therefore, all data related business rules have to be implemented as a central repository, so that they can be enforced systematically.

Meta-data compliance: Data should comply with its meta-data (Element | Declarative)

One aspect of attribute values is that they should conform to meta-data, which is specified as schema level rules. A significant collection of prior work has emphasized this aspect: "Value validity" (English 2009), "Conformance"(Loshin 2006), "Valid"(Byrne 2008), "Representation consistency"(Loshin 2001; Redman 1997), "Conformity" (Gatling 2007), "Conforming to meta-data" (Price and Shanks 2005b), "Accuracy/Validity" (Stvilia et al. 2007), "Definition Conformance" (English 2009) and "Understood" (Byrne 2008).

Standards and Regulatory compliance: All data processing activities should comply with the policies, procedures, standards, industry benchmark practices and all regulatory requirements that the organization is bound by (Information Object | Usage)

In addition to rules and constraints imposed by meta-data and business rules, organizations are bound by artifacts such as laws, regulations, standards, etc. These have an impact on organizational data and hence need to be consistently followed when processing organizational data. It is necessary to standardize and enforce such artifacts so that they can be used without conflict in data processing activities. "Validity" (HIQA 2011), "Data Specifications" (McGilvray 2008), "Signage Accuracy and Clarity" (English 2009) and "Semantic definition" (Byrne 2008) emphasize this aspect.

Statistical validity: Computed data must be statistically valid (Information Object | Usage)

Some tasks require that statistical validity of data is maintained (e.g.: calculation of statistical parameters, forecasts, etc.). For such data, it is important to ensure that data is collected and organized in accordance with statistical guidelines so that the data is appropriate for statistical calculations. This idea is emphasized in "Coherence" (HIQA 2011) and "Accuracy" (Lyon 2008).

5.5.6 Reliability and Credibility

Source Quality: Data used is from trusted and credible sources (Information Object | Usage)

Data that is obtained from third parties (e.g.: market analysis data) needs to be assessed regarding the quality of data sources when it is used in critical analysis. “Source Quality and Security Warranties or Certifications” (English 2009), “Authority” (Stvilia et al. 2007), “Enterprise Agreement of Usage” (Loshin 2001) and “Reputation” (Wang and Strong 1996) emphasize the importance of the quality of data sources.

Objectivity: Data are unbiased and impartial (Information Object | Usage)

The subjectivity of data collector, transmitter, or analyst, can result in data that lacks credibility for use in critical tasks. Prior work on “Objectivity” (Wang and Strong 1996) and “Presentation Objectivity” (English 2009) recognizes this issue. Thus, it is necessary to implement control processes to avoid any distortions from data collection to data usage.

Traceability: The lineage of the data is verifiable (Record | Usage)

The origin of a data instance, the process by which it arrived in a database, and its movement between databases is an important aspect of some data forensic activities (e.g.: error detection, data dependency analysis and compliance analysis). The ability to trace the evolution of data improves its credibility, as highlighted in “Traceability” (Eppler 2006) and “Verifiability” (Stvilia et al. 2007).

5.5.7 Consistency

Uniqueness: The data is uniquely identifiable (Record | Declarative)

Entities in the real world should have a unique representation in a database (i.e. there should be no duplication of records). This is a central concept in database management which is known as key constraint implemented at database schema level. Further, the same real world entity entered under different unique keys can be detected by application program level rules which are known as entity resolution. Prior work on “Uniqueness” (Loshin 2006), “Unique” (Byrne 2008) and “Mapped consistently” (Price and Shanks 2005b) highlights the importance of uniqueness.

Non-redundancy: The data is recorded in exactly one place (Record | Declarative)

An entity captured and recorded more than once in a database is known as data redundancy. The problem of data redundancy is highlighted in “Duplication” (McGilvray 2008), “Non-duplication” (English 2009) and “Uniqueness” (Gatling 2007). The heterogeneous technological landscape can result in multiple systems capturing the same data and recorded in multiple places. Therefore, a rule based can be used to consolidate such data and avoid redundancy.

Semantic consistency: Data is semantically consistent (Element | Declarative)

From the semiotic perspective of data, the meanings of labels and values of data attributes are of high importance when representing real world objects. The use of data is ultimately dependent on its meaning, and hence, labels and values of data attributes should be consistently used in the database. Prior work on "Equivalence of redundant or distributed data" (English 2009), "Understood" (Byrne 2008), and "Semantic Consistency" (Stvilia et al. 2007) emphasize this aspect.

Value consistency: Data values are consistent and do not provide conflicting or heterogeneous instances (Element | Declarative)

Attribute values need to be consistently used to avoid confusion (e.g.: "QLD" and "Queensland"). If different values are used in different databases, the outcome is heterogeneous and conflicting data. Assuring consistency regarding the content of data values through rules implemented at application program level is important. Accordingly, standardization of values is important to maintain consistency in databases – an issue emphasized in "Consistency" (Byrne 2008; Gatling 2007; Kimball and Caserta 2004; Loshin 2001; Loshin 2006; Redman 1997).

Format consistency: Data formats are consistently used (Element | Declarative)

The representation of certain types of data values (e.g.: date, address, phone number, etc.), even when the values are standardized, plays an important role in enabling understanding (both by humans and machines). Accordingly, data values should be represented in a consistent format in either at schema or application program level to stipulate their meaning because the representational format of data values is important from a semantic and pragmatic perspective. The definitions "Structural Consistency" (Stvilia et al. 2007), "Consistency and synchronization" (McGilvray 2008) emphasize having consistent formats for data values.

Referential integrity: Data relationships are represented through referential integrity rules (Record | Declarative)

A fundamental aspect of relational databases is referential integrity. Failure to maintain referential integrity rules results in data anomalies. "Referential integrity" (Loshin 2006), "Concurrency of redundant or distributed data" (English 2009) discuss this issue.

5.5.8 Usability & Interpretability

Usefulness and relevance: The data is useful and relevant for the task at hand (Information Object | Usage)

Information requirements change over time due to changes in the external environment (e.g.: market conditions) and the internal environment (e.g.: business processes). Ensuring that available data is

useful, and relevant requires continuous monitoring and evaluation in data usage process. This aspect is discussed in: “Informativeness /Redundancy” (Stvilia et al. 2007), “Interactivity” (Eppler 2006), “Relevance/ Relevancy” (English 2009; HIQA 2011; Lyon 2008; Stvilia et al. 2007; Wang and Strong 1996), “Transactability” (McGilvray 2008), “Applicability” (Eppler 2006), “Convenience” (Eppler 2006), “Naturalness” (Stvilia et al. 2007), “Completeness” (Wang and Strong 1996), “Type-sufficient” (Price and Shanks 2005b), and “Fact completeness” (English 2009).

Understandability: The data is understandable (Information Object | Usage)

A user must understand data to use it. Accordingly, data should be organized considering the level of cognitive skills of the target user group. A wide range of prior work addresses this aspect in detail: “Interpretability” (HIQA 2011; Wang and Strong 1996), “Correct Interpretation” (Loshin 2001), “Unambiguity” (Kimball and Caserta 2004), “Concise representation” (Wang and Strong 1996), “Ease of understanding” (Wang and Strong 1996), “Format precision “(Loshin 2001; Redman 1997),” Structured Valued Standardization”(English 2009), “Precise” (Byrne 2008) “Document Standardization”(English 2009), , “Representational consistency” (Wang and Strong 1996),” Presentation media appropriateness” (English 2009) “Presentation Clarity” (English 2009), “Usability” (HIQA 2011), “Clarity” (Eppler 2006), “Cohesiveness” (Stvilia et al. 2007), and “Complexity” (Stvilia et al. 2007),

Appropriate Presentation: The data presentation is aligned with its use (Information Object | Usage)

A representation of information used in one organization should be consistent with the representation of the same information in another organization in the same industry to create a harmonized understanding of facts (e.g.: a patient record used in one hospital should be understandable by any doctor in any other hospital). However, the representation should also be adaptable to different technological environments (e.g.: an electronic invoice generated in one system should be able to be used in another system). The convenience, compatibility, and flexibility of formats is discussed in “Understandable” (Price and Shanks 2005b), “Presentation Standardization” (English 2009), “Format flexibility “(Redman 1997), “Appropriateness” (Loshin 2001),” Suitably presented” (Price and Shanks 2005b), “Flexibly presented” (Price and Shanks 2005b),” Presentation Quality” (McGilvray 2008), “Presentation Utility” (English 2009), “Flexibility” (Loshin 2001) and “Portability” (Loshin 2001; Redman 1997).

Interpretability: Data should be interpretable (Information Object | Usage)

Incorrect interpretation of data leads to poor quality decisions. When human beings do the interpretation, it is important to standardize the process of interpreting data to reduce individual bias. Facilitating the process of interpretation is discussed in definitions "Comparability" (HIQA 2011; Lyon 2008), "Interpretability" (Redman 1997), "Appropriateness" (Redman 1997) and "Unambiguity" (Kimball and Caserta 2004).

Information value: The value that is delivered by quality information should be effectively evaluated and continuously monitored in the organizational context (Information Object | Usage)

Data and information are used to perform organizational tasks such that organizational goals are achieved. It is, therefore, important that the maintained data and information effectively contribute to achieving organizational goals. The value of information is emphasized in "Value added" (Wang and Strong 1996) and "Ubiquity" (Loshin 2001). In addition, these definitions, "Perception Relevance and Trust" (McGilvray 2008) and "Verifiability" (Stvilia et al. 2007) (which are also relevant to reliability and credibility) emphasize the importance of evaluating the value of information.

5.6 Summary

Though data quality dimensions is a widely researched topic, the growing number and the evolution of data quality dimensions, as well as emergence of new classifications and definitions is leading towards a lack of shared understanding which is a barrier towards formalizing DQ requirements to develop DQ patterns. In this chapter, we have addressed this problem by conducting an extensive review and consolidation of data quality dimensions literature.

We have summarized the existing definitions of data quality dimensions into eight main clusters referred to as dimensions. Within these dimensions, we recognized the main themes, which we called DQ characteristics, and provided a consolidated definition to each characteristic. The consolidated view of DQ characteristics is supplemented by related concepts like characteristic type, data granularity, verification metrics, validation metric and DQ implementation form so that a DQ requirement can be presented comprehensively.

Chapter 6

DEVELOPMENT OF DQ PATTERNS

6.1 Overview

The aim of this chapter is to instantiate the meta-meta-model for DQ requirements developed in Chapter 4 using the refactored DQ characteristics introduced in Chapter 5 and synthesize the patterns for modelling DQ requirements. This chapter explains the notion of pattern-based approaches in literature and the rationale behind extending the concept of patterns to model DQ requirements. Finally, we present the 33 data quality patterns developed in our study for modelling DQ requirements.

6.2 Developing a Meta-Model for a DQ requirement

In Chapter 6 we developed a meta-meta-model for a DQ requirement and the next step is to instantiate the meta-meta-model into a meta-model that can be used in creating real-world models for DQ requirements. The central construct in the meta-meta-model is the DQ characteristic and in Chapter 5 we identified 33 instances for DQ characteristics. Further, we developed the instances of other concepts related to each instance of DQ characteristic. In other words, we have produced 33 instances of the meta-meta-model that can be used as the constructs of a meta-model for DQ requirements.

The purpose of this conceptualization is to develop powerful constructs to model DQ requirements. As mentioned in section 2.3.4, in literature the pattern are used in various other disciplines to create models, due to the muddling power that the patterns provide in modelling complex scenarios (Franch 2013; Hoffmann et al. 2012; Rolland et al. 1998). Therefore, in defining the meta-model for DQ requirements, we want to define each construct in the meta-model as a DQ pattern so that an extra potential can be gained in modelling complex DQ requirements.

6.2.1 Data quality patterns

Patterns have been defined and successfully applied in various disciplines. For example, in the context of building architecture, Christopher Alexander defined a pattern as "a three part rule, which expresses a relation between a certain context, a problem and a solution" (Alexander, 1979). In the context of software development, Riehle and Zullighoven define a pattern as "the abstraction from a concrete form which keeps recurring in specific non-arbitrary contexts" (Riehle and Züllighoven, 1996), whereas Gamma *et. al* in (Gamma et al., 1995) define a pattern as "the solution to a recurring problem in a particular context". While multiple definitions exist, the definitions imply that, in

general, a pattern is a solution to a recurring problem. Thus, in the context of DQ management, a pattern can be defined as a solution to a recurring DQ problem. The meta-meta-model developed for a DQ requirement in chapter 4 contains clearly defined two representations (1) DQ problem and (2) DQ solution. Thus it is apparent that a DQ requirement defined by the meta-meta-model satisfies the general definition of a pattern.

In literature, the researchers have identified DQ problems as recurring problems by defining generic categories of DQ problems. (Castillo et al. 2006; Chen et al. 2014; Eppler 2006; Landis and Koch 1977; Lee et al. 2009; Nawaz 2012; Ng et al. 2012; Oliveira et al. 2005). Therefore it is clear that while there can be unique data quality problems, many are of a recurring nature regardless of the type of organization, or context of the application. Thus we argue that a DQ patterns can be used repeatedly to resolve DQ problems.

Rolland et al. (1998) have summarised two properties of a pattern explained in various contexts in the literature as follows.

(1) A pattern should be made explicit and precise so that it can be used time and time again. A pattern is explicit and precise if

    a. It defines the problem

    b. It defines a concrete solution,

    c. It defines a recurring set of situations the pattern can be applied.

(2) A pattern should be visualisable and identifiable. Visualisation may take the form of statements in the natural language, drawings ‘conceptual models and so on.

We observe that the above two properties mentioned by Rolland et al. (1998) further confirms that the DQ meta-meta model represents a structure of a pattern. In section 2.5 we discussed that overloaded notations have created complexities in existing approaches of modelling DQ requirements. In order to make the resultant DQ requirement models simple and usable, we present the patterns in the form of statements in natural language as mentioned by Rolland et al. (1998) above. Thus instantiating the meta-meta-model creates DQ patterns that can be used as a meta-model to model DQ requirements.

6.2.2 Instantiating the Meta-Meta-Model of a DQ requirement

In Chapter 5, we developed 33 instances of DQ characteristics. For each instance of a DQ characteristic, we specified a characteristic type (declarative/usage) and the applicable data

granularity level. The summary of this formalization is given in Table 6.1 below. Further, we established that every DQ characteristic has two types of metrics (validation\verification) and we provided a generic form of definitions of the metrics for both declarative and usage characteristics. Also, we established that implementation form of a characteristic depends on its type, and the declarative characteristics can be implemented using rule-based approach and the usage characteristics can be implemented using process-based approach. The summary of this formalization is given in Table 6.2: Characteristic type vs. implementation form and validation metric below.

Dimension	Characteristic	Description	Granularity	Characteristic type
Completeness	Completeness of mandatory attributes	The attributes which are mandatory for a complete representation of a real world entity must contain values and cannot be null	E	D
	Completeness of optional attributes	Optional attributes should not contain invalid null values	E	D
	Completeness of records	Every real world entity instance that is relevant for the organization can be found in the data	R	U
	Data volume	The volume of data is neither deficient nor overwhelming to perform an intended task	IO	U
Availability & Accessibility	Continuity of Data Access	The technology infrastructure should not prohibit the speed and continuity of access to the data for the users	IO	U
	Data maintainability	Data should be accessible to perform necessary updates and maintenance operations in its entire lifecycle	R	U
	Data awareness	The data users should be aware of all available data and its location	IO	U
	Ease of data access	The data should be easily accessible in a form that is suitable for its intended use	IO	U
	Data Punctuality	Data should be available at the time of its intended use	IO	U
	Data access control	The access to the data should be controlled to ensure it is secure against damage or unauthorized access	IO	U

Currency	Data timeliness	Data which refers to time should be available for use within an acceptable time relative to its time of creation	R	U
	Data freshness	Data which is subjected to changes over the time should be fresh and up-to-date with respect to its intended use	R	U
Accuracy	Accuracy to reference source	Data should agree with an identified source	E	U
	Accuracy to reality	Data should truly reflect the real world	R	U
	Precision	Attribute values should be accurate as per linguistics and granularity	E	D
Validity	Business rules compliance	Data must comply with business rules	E	D
	Meta-data compliance	Data should comply with its meta-data	E	D
	Standards and Regulatory compliance	All data processing activities should comply with the policies, procedures, standards, industry benchmark practices and all regulatory requirements that the organization is bound by	IO	U
	Statistical validity	Computed data must be statistically valid	IO	U
Reliability	Source Quality	Data used is from trusted and credible sources	IO	U
	Objectivity	Data are unbiased and impartial	IO	U
	Traceability	The lineage of the data is verifiable	R	U
Consistency	Uniqueness	The data is uniquely identifiable	R	D
	Non-redundancy	The data is recorded in exactly one place	R	D
	Semantic consistency	Data is semantically consistent	E	D
	Value consistency	Data values are consistent and do not provide conflicting or heterogeneous instances	E	D
	Format consistency	Data formats are consistently used	E	D
	Referential integrity	Data relationships are represented through referential integrity rules	R	D
Usability and	Usefulness and relevance	The data is useful and relevant for the task at hand	IO	U
	Understandability	The data is understandable	IO	U
	Appropriate Presentation	The data presentation is aligned with its use	IO	U
	Interpretability	Data should be interpretable	IO	U

	Information value	The value that is delivered by quality information should be effectively evaluated and continuously monitored in the organizational context	IO	U

Table 6.1: DQ characteristics and their types and granularity

We instantiated the meta-meta-model using the instances of the concept DQ characteristic and other related concepts as mentioned in Table 6.1 and Table 6.2. It should be noted that for implementation form we provided some guidelines as well, which can be used to design DQ rules or DQ processes. These guidelines we designed referring to literature. As mentioned in Chapter 2, there are a number of DQ methodologies developed by both academics and industry practitioners. In developing the guidelines the practitioner contributions were quite useful. Therefore most of the guidelines were developed referring to practitioner insights about DQ management.

Finally, as a result of this instantiation we developed 33 DQ patterns corresponding to the 33 DQ characteristics. These patterns are the constructs of the meta-model to model DQ requirements. Each DQ patterns provide generic definitions for a DQ user requirement\problem and a solution thus can be further instantiated according to any organizational DQ context and develop DQ requirement models.

Characteristic Type	Implementation form	Verification Metric / Threshold	Validation Metric/Threshold
Declarative	*Rule based approach*	*The number of non-conforming data values found per month/ per thousand records due to exceptions to, or violations of, the rules that define the declarative characteristic.*	*The extent to which required rules have been identified and implemented to maintain the declarative characteristic in concern.*
		Threshold is a numeric figure and the most optimum value is 0	*Threshold is a % value and the most optimum value is 100 %*

Usage	*Process-based approach*	*The number of tasks failed or negatively affected per month due to lack of conformance to the usage characteristic.* *OR* *The number of complaints received per month due to the lack of conformance to the usage characteristics*	*The extent to which, required capabilities and processes have been implemented to improve the data usage of a task*
		Threshold is a numeric figure and the most optimum value is 0	*Threshold is a % value and the most optimum value is 100 %*

Table 6.2: Characteristic type vs. implementation form and validation metric

For each DQ pattern, we have presented some guidelines for the design of its implementation form. Although providing such guidelines was not a part of the scope of our study, we suggest they will be useful in the process of DQ requirements analysis and modelling. In order to develop the guidelines, we performed a literature review on experience based DQ literature by DQ practitioners (English 2009; Kimball and Caserta 2004; Loshin 2004; Loshin 2011; McGilvray 2008; Redman 2008; Redman 1997). Further we used the published DQ management frameworks by the Canadian Institute for Health Information (CIHI) (Long and Seko 2005), the Health Information and Quality Authority in Ireland (HIQA 2011) and referred to some of the success stories of award-winning organizations for DQ recognised by the International Association for Data and Information Quality (IAIDQ 2015).

This literature review was performed using a qualitative data analysis tool NVIVO. In NVIVO we created a node for each DQ patterns and, relevant text in literature that explains the management activities applicable for a particular pattern was coded against the node. Then the guidelines were prepared based on this knowledge gathered through the coded information. We developed one or more examples to further illustrate each guideline so that the stakeholders can easily understand the guideline. Coding of literature and development of the guidelines was performed incrementally until a reasonable set of guidelines achieved for each pattern.

In the next section, we discuss the repository of DQ patterns that we developed to formally present them to the DQ stakeholders.

6.3 Repository of DQ patterns

Thirty-three DQ patterns resulted from the instantiation were organised into eight main clusters for easy reference. The eight DQ dimensions were the clusters and each pattern was listed under the

dimension of which the DQ characteristic of the pattern is based on. The full pattern repository can be accessed online with the URL http://dke.uqcloud.net/DataQualityPatterns/.

The purpose of this repository is to disseminate the knowledge of patterns among DQ stakeholders so that they can use the knowledge of patterns in designing DQ solutions in organizational settings. As suggested by Rolland et al. (1998), a pattern can be visualised using statements in the natural language, drawings 'conceptual models and so on. Therefore, in this repository we use statements in natural language to visualize the patterns using a template that contains the constructs of the Meta-Meta Model explained in figure 4.4 in chapter 4.

In this repository, the users have the facility to search the patterns using keywords, view background literature behind each pattern and also compare two patterns to study them comprehensively.

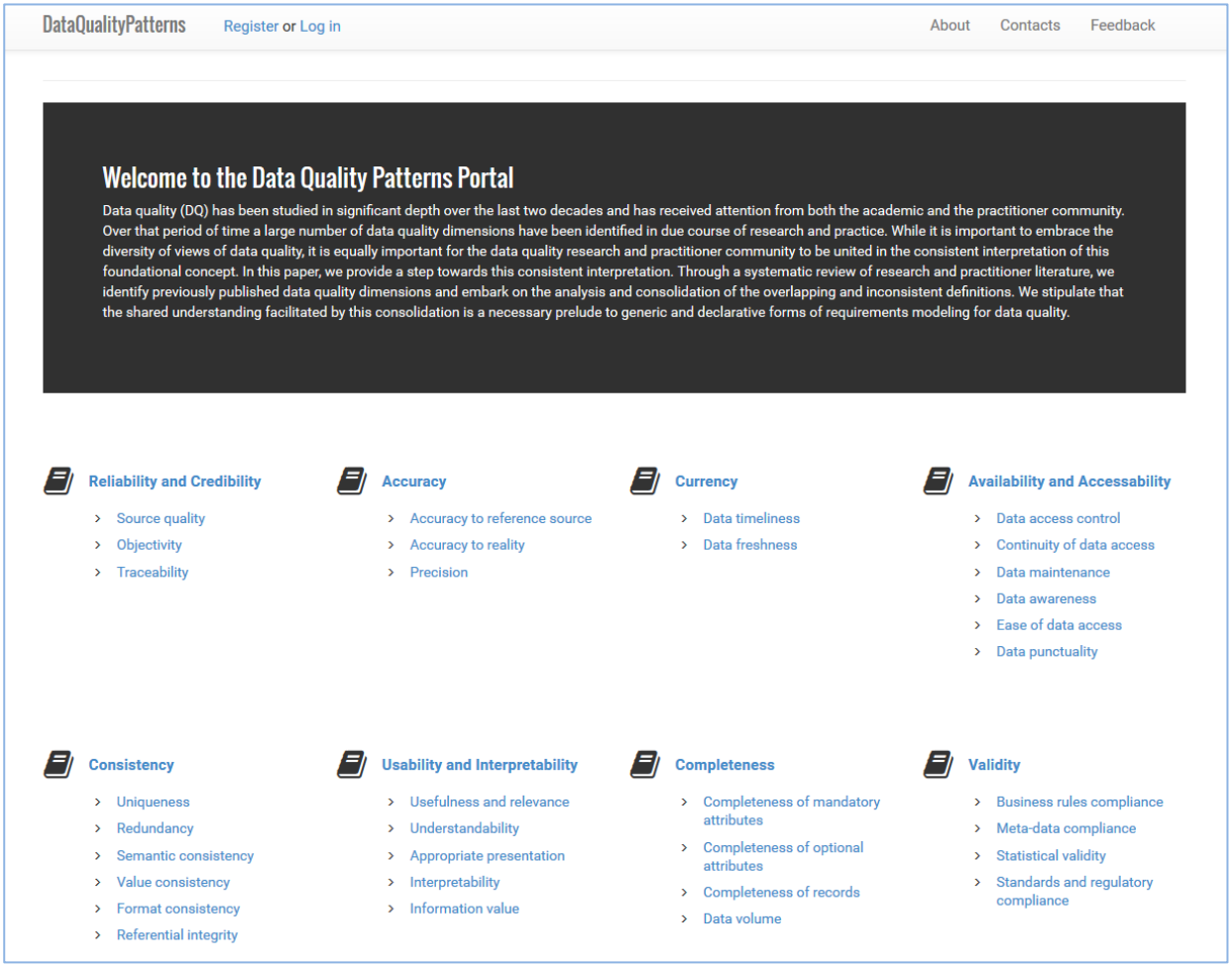


Figure 6.1: Home page of the pattern repository

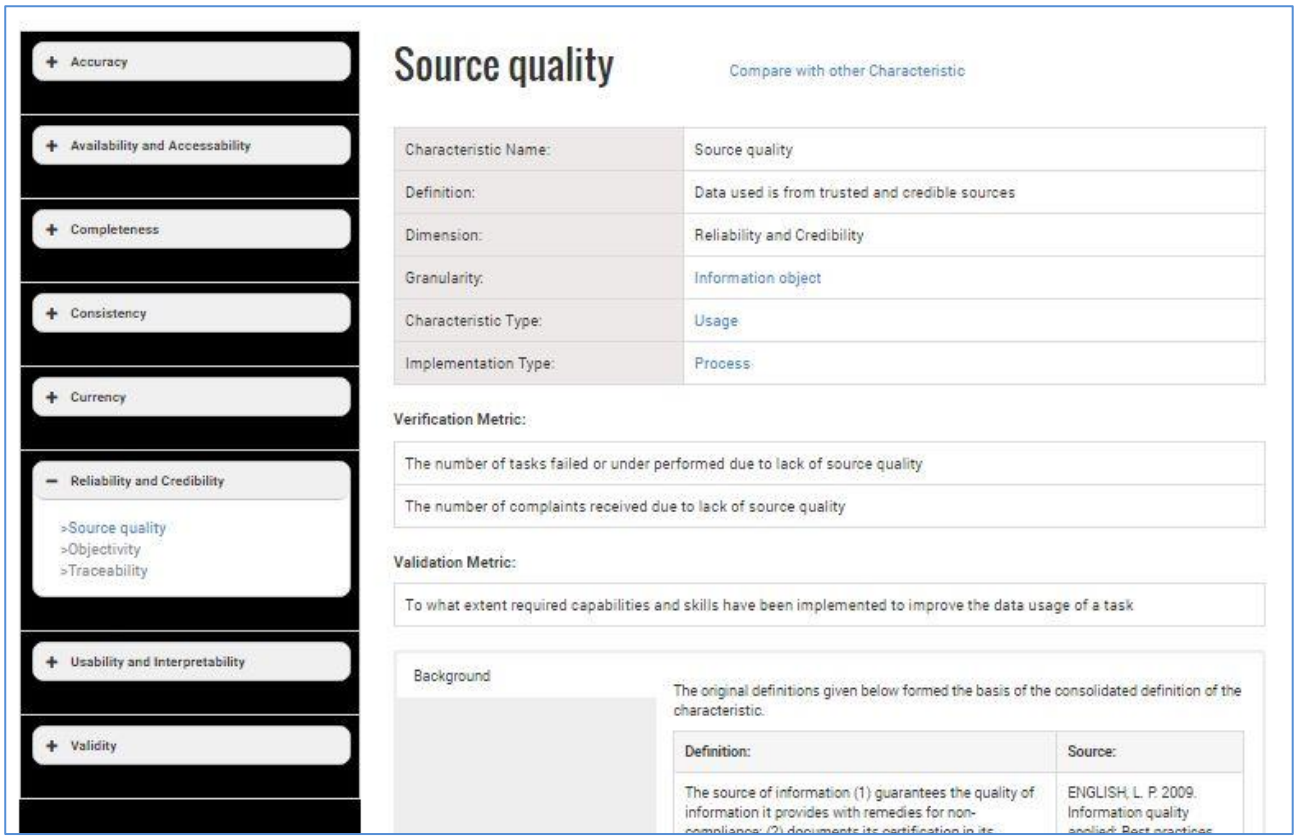


Figure 6.2: Navigation and Pattern view

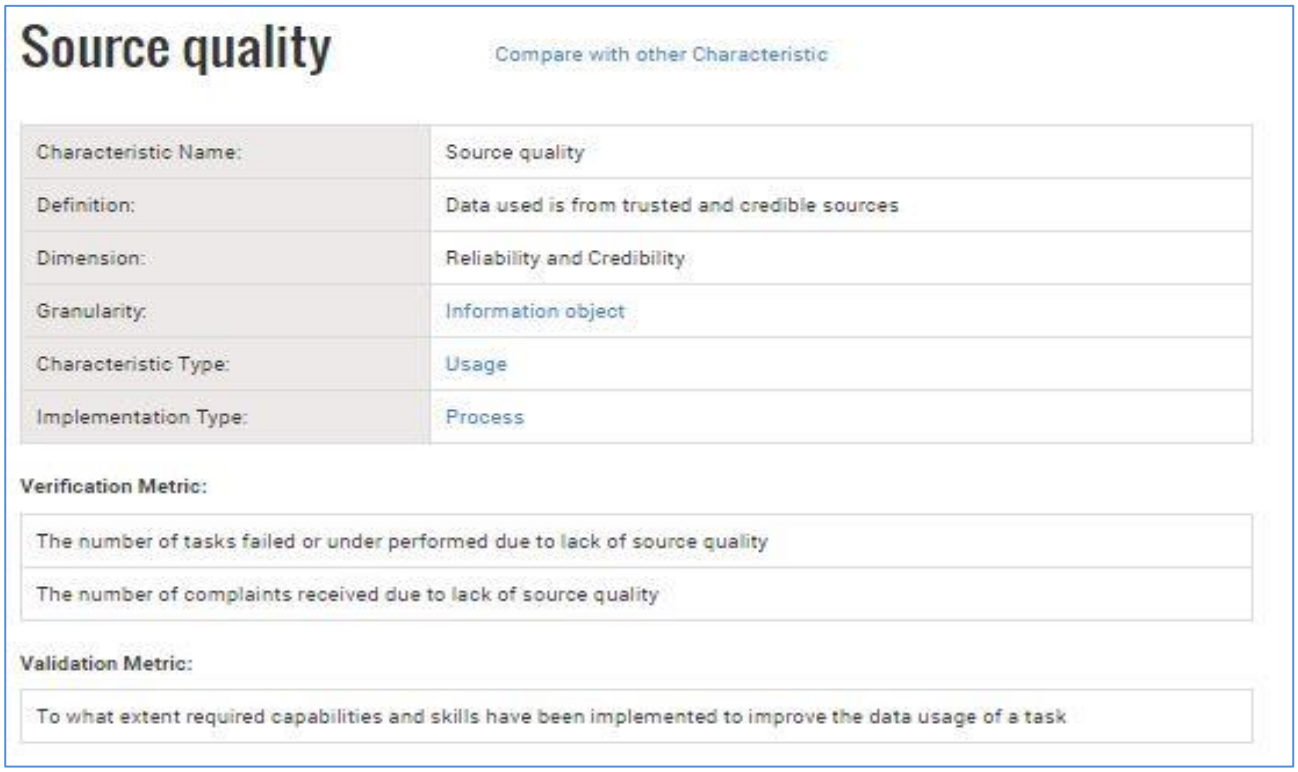


Figure 6.3: Detail View of Pattern

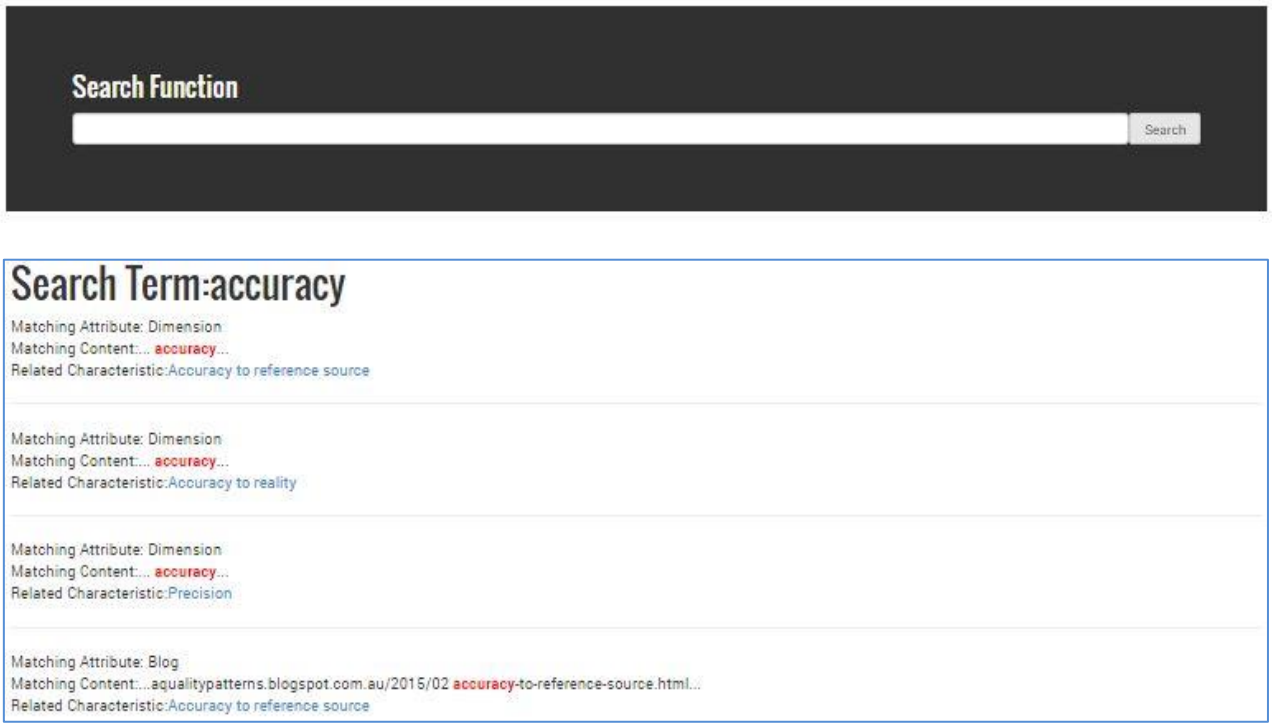


Figure 6.4: Search Patterns - Text Search

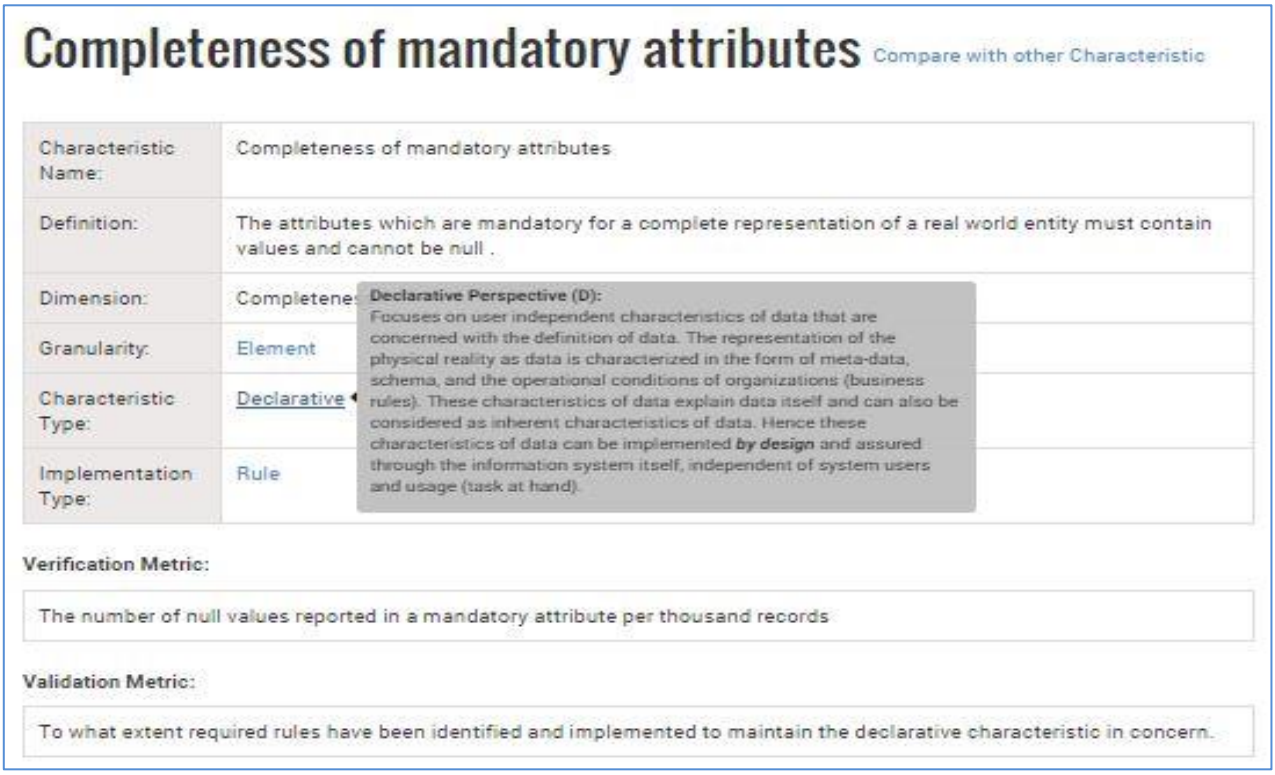


Figure 6.5: Descriptions of concepts used in patterns

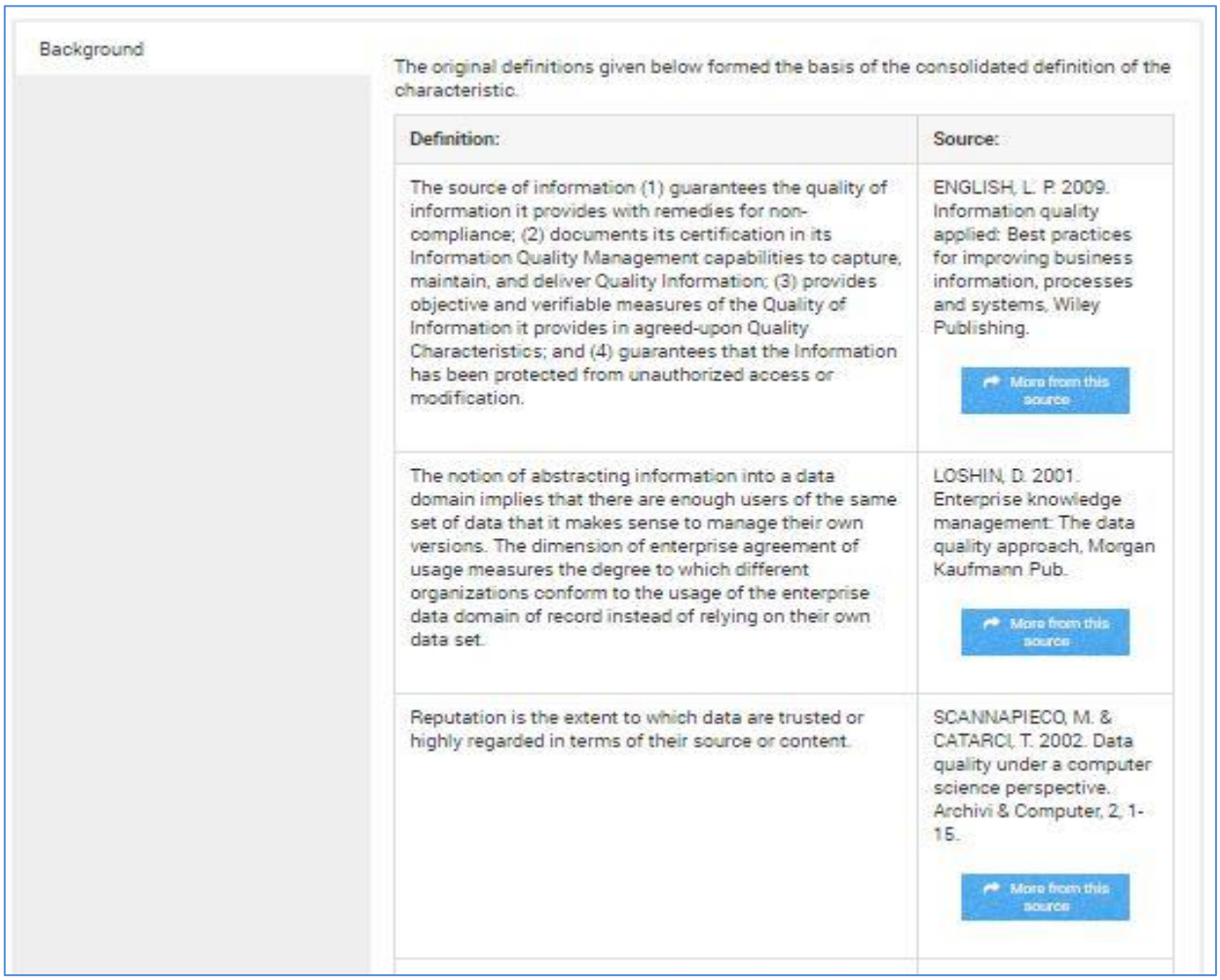
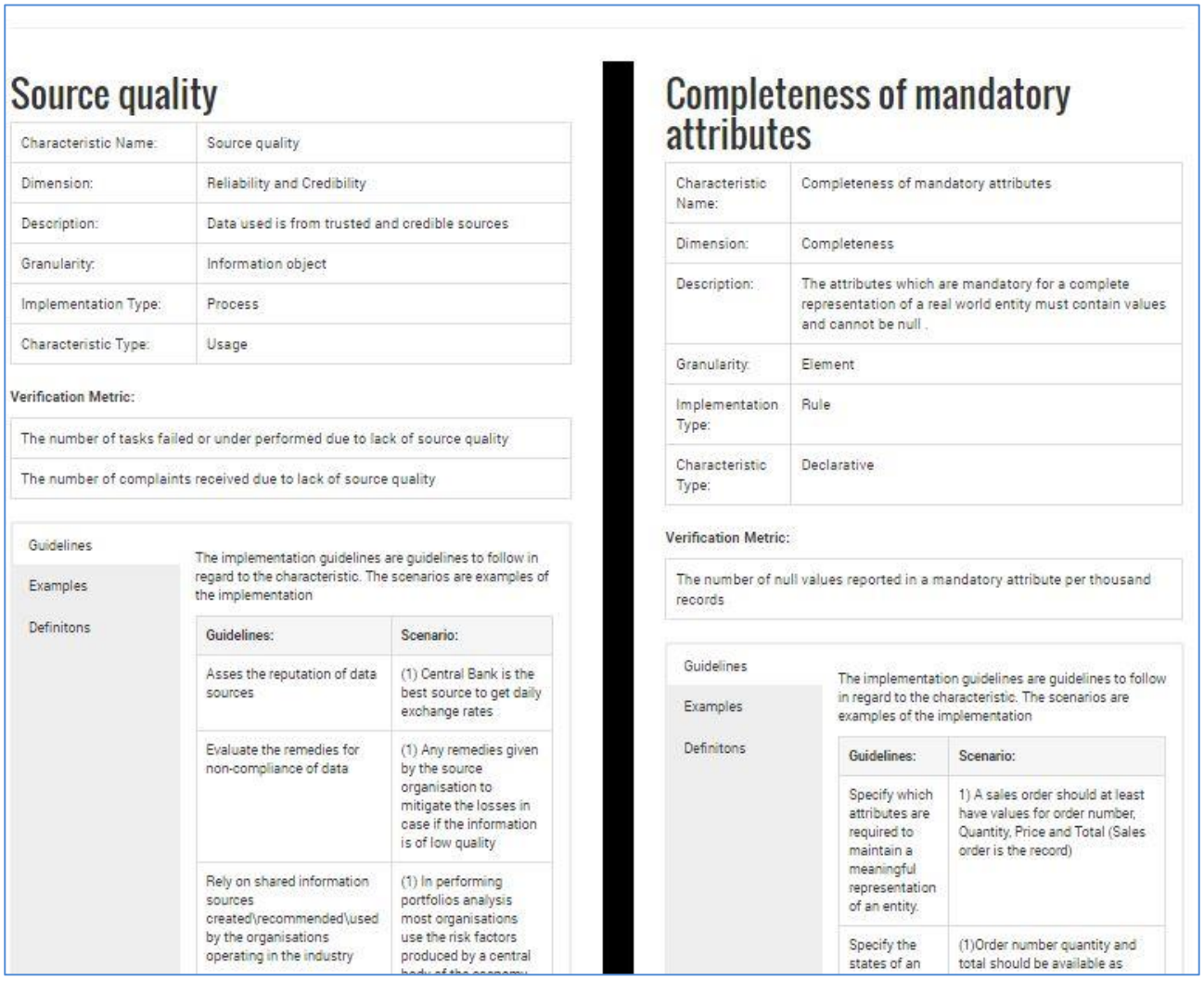


Figure 6.6: Academic Background of the pattern

The home page list the thirty-three DQ patterns with eight DQ dimensions (Figure 6.1). The users can view each pattern via the hyperlink. The patterns can be viewed in detail and navigated systematically (Figure 6.2, Figure 6.3). A searching facility is available to search the relevant patterns using the keywords (Figure 6.4). An online help is available in the form of pop-up screens to explain the concepts used in the patterns (Figure 6.5). Further, the online repository allows the users to study the background details about each and every pattern, that include references for the underlying definitions used to create each DQ characteristic (Figure 6.6)



Source quality

Characteristic Name:	Source quality
Dimension:	Reliability and Credibility
Description:	Data used is from trusted and credible sources
Granularity:	Information object
Implementation Type:	Process
Characteristic Type:	Usage

Verification Metric:

The number of tasks failed or under performed due to lack of source quality
The number of complaints received due to lack of source quality

Guidelines

Examples

Definitons

The implementation guidelines are guidelines to follow in regard to the characteristic. The scenarios are examples of the implementation

Guidelines:	Scenario:
Asses the reputation of data sources	(1) Central Bank is the best source to get daily exchange rates
Evaluate the remedies for non-compliance of data	(1) Any remedies given by the source organisation to mitigate the losses in case if the information is of low quality
Rely on shared information sources created\recommended\used by the organisations operating in the industry	(1) In performing portfolios analysis most organisations use the risk factors produced by a central

Completeness of mandatory attributes

Characteristic Name:	Completeness of mandatory attributes
Dimension:	Completeness
Description:	The attributes which are mandatory for a complete representation of a real world entity must contain values and cannot be null .
Granularity:	Element
Implementation Type:	Rule
Characteristic Type:	Declarative

Verification Metric:

The number of null values reported in a mandatory attribute per thousand records

Guidelines

Examples

Definitons

The implementation guidelines are guidelines to follow in regard to the characteristic. The scenarios are examples of the implementation

Guidelines:	Scenario:
Specify which attributes are required to maintain a meaningful representation of an entity.	1) A sales order should at least have values for order number, Quantity, Price and Total (Sales order is the record)
Specify the states of an	(1)Order number quantity and total should be available as

Figure 6.7: Side by Side Comparison of patterns

In the following section, we present the patterns in each cluster. It should be noted that for implementation guidelines we have provided the main references that we used to design the guideline.

6.3.1 Patterns of data completeness

<table>
<tr><th colspan="3">Pattern 1 (Completeness of mandatory attributes)</th></tr>
<tr><td colspan="3">Characteristic: The attributes which are mandatory for a complete representation of a real world entity must contain values and cannot be null.</td></tr>
<tr><td>Dimension: Completeness</td><td>Data granularity: Element</td><td>Type: Declarative</td></tr>
<tr><td colspan="3">Verification metric: The number of null values reported in a mandatory attribute per thousand records / per month</td></tr>
<tr><td colspan="3">Implementation form: Rules-based approach
Implementation of rules to prevent null values in mandatory attributes</td></tr>
<tr><td colspan="2">Guidelines for DQ rules
(Byrne 2008; English 2009; Loshin 2004)</td><td>Examples</td></tr>
<tr><td colspan="2">Specify which attributes are required to maintain a meaningful representation of an entity and create a validation rule.</td><td>A sales order should at least have values for order number, quantity, price and total (sales order is the record).</td></tr>
<tr><td colspan="2" rowspan="2">Specify the states of an entity where the above-identified attributes become mandatory values.</td><td>Order number quantity and the total should be available as mandatory by the time the order is created whereas price will become mandatory when the order is approved. (states :"order created" "order approved").</td></tr>
<tr><td>The product is retired and now has a product-last-available-date.</td></tr>
<tr><td colspan="2">Specify the dependencies of entities in an operational context to identify the mandatory values.</td><td>An invoice number should exist to create a gate pass.</td></tr>
<tr><td colspan="2">Specify default values where possible.</td><td>The default country is Australia for those who fill in the application from Australian IP addresses.</td></tr>
<tr><td colspan="3">Validation metric: The extent to which required rules have been identified and implemented to maintain the mandatory attribute of concern</td></tr>
</table>

Table 6.3: Pattern 1 (Completeness of mandatory attributes)

<table>
<tr><th colspan="3">Pattern 2 (Completeness of optional attributes)</th></tr>
<tr><td colspan="3">Characteristic: Optional attributes should not contain invalid null values.</td></tr>
<tr><td>Dimension: Completeness</td><td>Data granularity: Element</td><td>Type: Declarative</td></tr>
<tr><td colspan="3">Verification metric: The number of invalid null values reported in an optional attribute per thousand records/ per month</td></tr>
<tr><td colspan="3">Implementation form: Rules-based approach
Implementation of rules to prevent/detect invalid null values in optional attributes</td></tr>
<tr><td colspan="2">Guidelines for DQ rules
(Loshin 2004)</td><td>Examples</td></tr>
<tr><td colspan="2">Provide default values for each valid case of null values for the attribute of concern so that null values occur only for actually missing values which are invalid cases for the attribute of concern.</td><td>Case 1: Attribute values are not defined for a particular entity instance (e.g. maiden name of unmarried women). Such instances will get the default value “NOT DEFINED”.

Case 2: Attribute values are defined for the entity instance whereas the real value for the attribute instance is null (e.g. vehicle number of a student who does not have a vehicle). Such instances will get the default value “NOT EXISTING”.

Case 3: Attribute values are defined for the entity instance and the attribute instance should have a value (e.g. student’s date of birth).</td></tr>
<tr><td colspan="3">Validation metric: The extent to which required rules have been identified and implemented to detect invalid null values in optional attribute in concern</td></tr>
</table>

Table 6.4: Pattern 2 (Completeness of optional attributes)

<table>
<tr><th colspan="3">Pattern 3 (Completeness of records)</th></tr>
<tr><td colspan="3">Characteristic: Every real world entity instance that is relevant for the organization can be found in the data.</td></tr>
<tr><td>Dimension: Completeness</td><td>Data granularity: Record</td><td>Type: Usage</td></tr>
<tr><td colspan="3">Verification metric: (1) The number of tasks failed or underperformed due to missing records (2) The number of complaints received due to missing records</td></tr>
</table>

Implementation form: *Process-based approach* *Implementation of capabilities and processes to prevent/detect missing records*	
Guidelines for DQ processes and capabilities (English 2009; HIQA 2011; ISO 2012; Kimball and Caserta 2004)	**Examples**
Implement a process level validation mechanism to avoid the occurrence of missing records.	*(1) A buyer must record/verify an expense or asset in accordance with accepting/receiving any purchased items. (2)New applications are stored in a temporary cabinet after being entered into the system and they will be transferred to the file cabinet at the end of every week after the property manager cross-checks them with the system.*
Execute database commits upon transaction sequences in application programs and makes sure all the transactions in the sequence successfully commit and generate the required records at the end of the sequence.	*In generating the MRP, the database operations will not be committed unless all materials in the BOM are successfully executed for the MRP.*
When distributed databases are used or online data collection devices are used, ensure the synchronisation/replication of records happens successfully without distortions and omissions.	*EFTPOS transactions are replicated to the bank database and the new balance B/F in the account is created.*
Implement a periodic audit process for critical tangible objects that are recorded as data in the database.	*Tangible assets in the organization are audited annually.*
Implement a validation mechanism in data transfers considering the business rules, to monitor and ensure all records relevant to an event/transaction are transferred successfully.	*(1) Rules are applied to verify the number of records in the source and destination files.* *(2) All records relevant to a customer trip are transferred to the central database from online data stores.*
Maintain error logs for system transactions and regularly monitor them and perform relevant forensic activities to find missing records.	*A failed instance of a sales order creation is recorded in the error log.*
Validation metric: *The extent to which required capabilities and processes have been identified and implemented to prevent/detect missing records*	

Table 6.5: Pattern 3 (Completeness of records)

<table>
<tr><th colspan="3">Pattern 4 (Data volume)</th></tr>
<tr><td colspan="3">Characteristic: The volume of data is neither deficient nor overwhelming to perform an intended task.</td></tr>
<tr><td>Dimension: Completeness</td><td>Data granularity:
Information object</td><td>Type: Usage</td></tr>
<tr><td colspan="3">Verification metric: (1) The number of tasks failed or underperformed due to not meeting the right volume of data (2) The number of complaints received due to volume related issues</td></tr>
<tr><td colspan="3">Implementation form: Process-based approach
Implementation of capabilities and processes to maintain the right volume of data</td></tr>
<tr><td colspan="2">Guidelines for DQ processes and capabilities
(Eppler 2006; McGilvray 2008; Scannapieco et al. 2004)</td><td>Examples</td></tr>
<tr><td colspan="2">Define the scope of data in terms of organizational coverage to perform an organizational activity.</td><td>At least 70% of the production units should submit data to calculate total production efficiency of the company.</td></tr>
<tr><td colspan="2">Define the scope of data in terms of activities relating to any organizational task.</td><td>Pages with more than one thousand hits per day and above are considered for the analysis.</td></tr>
<tr><td colspan="2">Define the scope of data in terms of the population of data which is under investigation.</td><td>At least 10% of the population of white blood cells in the culture should be collected as samples to calculate its growth.</td></tr>
<tr><td colspan="2">Define an appropriate number of records in terms of lower and upper limits for any task.</td><td>At least six responses should be available to evaluate a tutor's skills and competency.</td></tr>
<tr><td colspan="3">Validation metric: The extent to which required capabilities and processes have been identified and implemented to maintain data volume</td></tr>
</table>

Table 6.6: Pattern 4 (Data volume)

6.3.2 Patterns of data availability and accessibility

<table>
<tr><th colspan="3">Pattern 5 (Continuity of Data Access)</th></tr>
<tr><td colspan="3">Characteristic: The technology infrastructure should not prohibit the speed and continuity of access to the data for the users.</td></tr>
<tr><td>Dimension: Availability and accessibility</td><td>Data granularity:
Information object</td><td>Type: Usage</td></tr>
<tr><td colspan="3">Verification metric: (1) The number of tasks failed or underperformed due to the lack of continuity in data access, (2) The number of complaints received due to a lack of continuity in data access</td></tr>
<tr><td colspan="3">Implementation form: Process-based approach
Implementation of capabilities and processes to maintain continuity of data access</td></tr>
<tr><td colspan="2">Guidelines for DQ processes and capabilities
(Eppler 2006; Scannapieco and Catarci 2002)</td><td>Examples</td></tr>
<tr><td colspan="2">A convenient and efficient platform should be made available to access data depending on the task at hand.</td><td>For a sales person, a web-based interface run on a smart device is more suitable to quickly access data.</td></tr>
<tr><td colspan="2">The speed of the data retrieval should be acceptable for users' working space.</td><td>(1) For an online customer care executive, speedy retrieval of information is necessary since the customer cannot be kept waiting.
(2) With the growth of the database, reporting becomes slower (anti-example).</td></tr>
<tr><td colspan="2">Continuous and unobstructed connectivity should be ensured for data retrievals.</td><td>The connection is lost while accessing reports (anti-example).</td></tr>
<tr><td colspan="2">Proper concurrency control has been implemented.</td><td>Access to data is controlled by locks.</td></tr>
<tr><td colspan="2">Technological changes in the infrastructure/system should be handled in such a way that they should not make data inaccessible.</td><td>The new version of the software does not provide access to "X out orders" since the new version does not allow the function "X out".</td></tr>
<tr><td colspan="3">Validation metric: The extent to which required capabilities and processes have been identified and implemented to maintain continuity of data access</td></tr>
</table>

Table 6.7: Pattern 5 (Continuity of Data Access)

<table>
<tr><th colspan="3">Pattern 6 (Data maintainability)</th></tr>
<tr><td colspan="3">Characteristic: Data should be accessible to perform necessary updates and maintenance operations in its entirety.</td></tr>
<tr><td>Dimension: Availability and accessibility</td><td>Data granularity: Record</td><td>Type: Usage</td></tr>
<tr><td colspan="3">Verification metric: (1) The number of tasks failed or underperformed due to a lack of data maintenance, (2) The number of complaints received due to a lack of continuity in data access</td></tr>
<tr><td colspan="3">Implementation form: Process-based approach
Implementation of capabilities and processes to ensure maintainability of data</td></tr>
<tr><td colspan="2">Guidelines for DQ processes and capabilities
(Kimball and Caserta 2004; Loshin 2009; McGilvray 2008)</td><td>Examples</td></tr>
<tr><td colspan="2">Mission critical data in the organization should be categorised and their lifecycle should be identified to plan for maintenance purposes.</td><td>A sales order is created once a customer signs a contract. Then it is updated in three instances:
(1)The delivery date and shipment date are updated once the production plan is created.
(2) The actual quantity is updated once the manufacturing is complete.
(3) The total cost is updated once the freight charges are incurred.
(4) A sales order is achieved after one year from delivery.</td></tr>
<tr><td colspan="2">A maintenance policy for mission-critical data should be developed and implemented to handle on-going systematic updates (create, read, update, delete, archive and cleanse).</td><td>Customer data is created when a customer enters into a contract, updated once the customer details change or the contact changes, and archived once the contract ends.</td></tr>
<tr><td colspan="2">When multiple versions of the same data are available through different datasets\databases create a master record and make it available across the systems.</td><td>Implementation of Master data management (MDM)</td></tr>
<tr><td colspan="2">Leverage applications and storage technology in such a way that the maintenance policies can be applied to data.</td><td>Addresses which were not updated during the last 24 months are prompted for validations.</td></tr>
</table>

Create a responsibility structure/authorization structure and a communication structure to manage the process of information generation, maintenance, and utilization.	*(1) It is the responsibility of the work study team to provide standard minute values (SMV) for a garment.* *(2) Approved SMVs should be sent to the planning department for planning purposes.*
Validation metric: *The extent to which required capabilities and processes have been identified and implemented to maintain data in its entire lifecycle*	

Table 6.8: Pattern 6 (Data maintainability)

Pattern 7 (Data awareness)		
Characteristic: *Data users should be aware of all available data and their location.*		
Dimension: *Availability and accessibility*	**Data granularity:** *Information object*	**Type:** *Usage*
Verification metric: *(1) The number of tasks failed or underperformed due to a lack of data awareness (2) The number of complaints received due to a lack of data awareness*		
Implementation form: *Process-based approach* *Implementation of capabilities and processes to maintain data awareness*		
Guidelines for DQ processes and capabilities (HIQA 2011; Long and Seko 2005)		**Examples**
Educate users on the data landscape of the organization and how to access the information when needed.		*Training programs are conducted to educate on using the systems and retrieve data when necessary.*
Provide appropriate searching tools, manuals to find the required information.		*(1) Data catalogues* *(2)"Search help " facility provided in SAP applications*
Educate users on how data is collected and what are the procedures used in data collection so that they can decide on the appropriateness of the data for use.		*Opinion poll data during festival season is not appropriate to analyse drinking habits.*
Validation metric: *The extent to which required capabilities and processes have been identified and implemented to maintain data awareness*		

Table 6.9: Pattern 7 (Data awareness)

Pattern 8 (Ease of data access)		
Characteristic: *Data should be easily accessible in a form that is suitable for its intended use.*		
Dimension: *Availability & Accessibility*	**Data granularity:** *Information object*	**Type:** *Usage*
Verification metric: *(1) The number of tasks failed or underperformed due to a lack of ease in data access (2) The number of user complaints received regarding difficulties in data access*		
Implementation form: *Process-based approach* *Implementation of capabilities and processes to maintain ease of data access*		
Guidelines for DQ processes and capabilities (Lyon 2008; OIC 2012; Stvilia et al. 2007)	**Examples**	
Routinely accessed information to continue operations should be automatically delivered to stakeholders online without their having to waste time searching for it.	*Daily exchange rates are linked into the accounting application or maintained in a dashboard on the accountant's desktop.*	
Information needed for management reporting purposes should be identified and catered for through built-in reports as much as possible where the users do not have to process data manually and create the reports.	*Order status is frequently searched information by different stakeholder groups and hence a report is made available with multiple searching criteria.*	
In preparing cross-functional reports, segregate duties to each functional unit so that relevant data can be accessed and prepared over the relevant period avoiding any bottlenecks.	*A work in progress report contains information from all production lines and each line supervisor is responsible for giving that data to the floor manager.*	
Assist users by providing tools to query the database without their having to use any specific technical knowledge and perform business analytics to bring innovation.	*Technical infrastructure supports the users to develop their own reports based on dynamic information needs without consulting technical staff (e.g. SAP queries).*	
Facilitate the user to filter the relevant information depending on the need.	*A sales report is generated with filtering criteria for customer and date range.*	
The interfaces and reports should be created for the users' convenience so the users do not have to write complex queries or process information further before usage.	*Product prices are ordered as per "relevance" or "price" to enable an e-commerce customer on a purchase decision.*	

Validation metric: *The extent to which required capabilities and processes have been identified and implemented to maintain ease of data access*

Table 6.10: Pattern 8 (Ease of data access)

Pattern 9 (Data punctuality)		
Characteristic: *Data should be available at the time of its intended use.*		
Dimension: *Availability and accessibility*	**Data granularity:** *Information object*	**Type:** *Usage*
Verification metric: *(1) The number of tasks failed or underperformed due to a lack of data punctuality (2) The number of complaints received due to a lack of data punctuality*		
Implementation form: *Process-based approach* *Implementation of capabilities and processes to maintain data punctuality*		
Guidelines for DQ processes and capabilities (Eppler 2006; Loshin 2006)		**Examples**
Standardize the timelines for the availability of information for a particular task.		*Investment product pricing data is often provided by third-party vendors. As the success of the business depends on accessibility to that pricing data, service levels specifying how quickly the data must be provided are defined and compliance with those timeline constraints enforced.*
Create efficient processes for information delivery by removing the bottlenecks in information flow.		*Billing details of a patient are gathered two hours before discharging the patient.*
Validation metric: *The extent to which required capabilities and processes have been identified and implemented to maintain data punctuality*		

Table 6.11: Pattern 9 (Data punctuality)

Pattern 10 (Data access control)		
Characteristic: *The access to the data should be controlled to ensure it is secure against damage or unauthorised access.*		
Dimension: *Availability and accessibility*	**Data granularity:** *Information object*	**Type:** *Usage*
Verification metric: *(1) The number of tasks failed or underperformed due to a lack of data access control (2) The number of complaints received due to a lack of data access control*		
Implementation form: *Process-based approach* *Implementation of capabilities and processes to control data access*		
Guidelines for DQ processes and capabilities (English 2009; Eppler 2006; Loshin 2001)		**Examples**
Periodically evaluate the security needs considering the criticality of data (value, confidentiality, privacy needs etc.) and data accessibility requirements and then update the information security policy consistently.		*(1) Employee salary is a confidential data item and hence security against unauthorised access is needed.* *(2)Master data has a high economic value to the organization and hence security is needed against unauthorised access and change.*
Continuously evaluate the risks threats and identify the vulnerabilities for data and update the information security policy.		*The frequency of security assessment for data associated with online transactions was increased due to the high volume of online transactions.*
Implementation of access controls for each piece of critical information is as prescribed by the information security policy.		*(1) An employee's salary data can be viewed only by him/herself and his or her superiors.* *(2) Master data can be created and updated only by the authorised executives.* *(3) Login credentials are required for system access*
Data is stored in secured locations and appropriate backups are made.		*(1) Databases are stored in a special server and backups are made regularly.* *(2) Documents are saved using a content management system in a file server.*
The accessibility of information is restricted using software based mechanisms.		*(1)Data encryption* *(2)Firewalls*

Restrict the accessibility of information using hardware-based mechanisms.	*Security tokens*
Validation metric: *The extent to which required capabilities and processes have been identified and implemented to control data access*	

Table 6.12: Pattern 10 (Data access control)

6.3.3 Patterns of data currency

Pattern 11 (Data timeliness)		
Characteristic: Data which refers to time should be available for use within an acceptable time relative to its time of creation		
Dimension: *Currency*	**Data granularity:** *Record*	**Type:** *Usage*
Verification metric: *(1) The number of tasks failed or underperformed due to a lack of data timeliness (2) The number of complaints received due to a lack of data timeliness*		
Implementation form: *Process-based approach* *Implementation of capabilities and processes to maintain data timeliness*		
Guidelines for DQ processes and capabilities (English 2009; McGilvray 2008)		**Examples**
Recognize the activity/event that generates the time sensitive data and specify a criterion to generate attribute values if necessary.		*Efficiency of production line:* *Activity: Line out a quality check which signifies the end of manufacturing of a product in a manufacturing line.* *Criteria: The number of products which passed the line out quality checks per given time period is the efficiency measure of the line.*
Specify the valid time period for the values of the attribute to be recorded.		*(1) The growth of the bacteria should be measured after 15 hours of culturing. (2)Efficiency should be calculated and recorded once every 10 minutes starting from the initial 10th minute of an hour (six times per hour).*
Specify the valid time period for the values of attributes to be used and implement controls to adhere to them.		*The exchange rate for the day is valid from 8 am to 8am the following day.*

Validation metric: *The extent to which required capabilities and processes have been identified and implemented to control data access*

Table 6.13: Pattern 11 (Data timeliness)

Pattern 12 (Data freshness)		
Characteristic: *Data which is subjected to changes over the time should be fresh and up-to-date with respect to its intended use.*		
Dimension: *Currency*	**Data granularity:** *Record*	**Type:** *Usage*
Verification metric: *(1) The number of tasks failed or underperformed due to a lack of data freshness (2) The number of complaints received due to a lack of data freshness*		
Implementation form: : *Process-based approach* *Implementation of capabilities and processes to maintain data freshness*		
Guidelines for DQ processes and capabilities (McGilvray 2008)		**Examples**
Identify the natural factors which make a particular data item obsolete.		*(1) A seasonal change may impact the customers' food preferences.* *(2) Customers who are students may change their addresses frequently.*
Considering the above factors plan for data refreshing activities by specifying the frequency of refreshing the data elements and adhere to the plan.		*Customer contact information should be refreshed annually.*
Identify the master data that may change over the time but may be used in the longitudinal analysis.		*The name of a customer in 2001 is XYZ (Pvt) Ltd. After a merger in 2006, its name is ABC PLC. This customer is an ongoing customer.*
For such master data maintain longitudinal versions with a time stamp in such a way that they can be linked in longitudinal analysis		*2001-2005: XYZ (Pvt) Ltd* *2006-2012: ABC PLC*
Validation metric: *The extent to which required capabilities and processes have been identified and implemented to maintain data freshness*		

Table 6.14: Pattern 12 (Data timeliness)

6.3.4 Patterns of data accuracy

<table>
<tr><th colspan="3">Pattern 13 (Accuracy to reference source)</th></tr>
<tr><td colspan="3">Characteristic: Data should agree with an identified source.</td></tr>
<tr><td>Dimension: Accuracy</td><td>Data granularity: Element</td><td>Type: Usage</td></tr>
<tr><td colspan="3">Verification metric: (1) The number of tasks failed or underperformed due to a lack of accuracy with reference sources (2) The number of complaints received due to a lack of accuracy with reference sources</td></tr>
<tr><td colspan="3">Implementation form: Process-based approach
Implementation of capabilities and processes to maintain accuracy with reference sources</td></tr>
<tr><td colspan="2">Guidelines for DQ processes and capabilities
(Long and Seko 2005; Loshin 2011; McGilvray 2008; Redman 1997)</td><td>Examples</td></tr>
<tr><td colspan="2">Establish the source for a data attribute and maintain facilities to access the correct source.</td><td>(1) The actual cost of raw material is taken from supplier invoices and not from quotations.
(2) Interest rates are taken from daily central bank statistics which are available in the finance system online.</td></tr>
<tr><td colspan="2">Establish the data capturing points in the business process without leading to any ambiguity and enforce process level validation mechanisms to ensure the process is being followed.</td><td>(1) Personal drug utilization data is captured at POS units at pharmacies and all pharmacies in the country are connected to a central system (all pharmacy data is considered).
(2) In a barcode scanning system in a production system, finished products cannot be scanned into quality checked products (finished and quality checked are the two data capturing points here).</td></tr>
<tr><td colspan="2">Implement effective techniques and efficient technological solutions (devices) in collecting data which minimizes data errors and omissions in data capturing.</td><td>(1) Barcode scanning is used to enter sales of products.
(2) Invoices are scanned into the system and prices are automatically recognized.
(3) Standard forms are used to collect patient data.</td></tr>
<tr><td colspan="2">If data is collected and transferred batch wise, establish the frequencies of data transfers/uploads</td><td>(1) All drug utilization data collected in the pharmacies are transferred to the central system at the end of every month.</td></tr>
</table>

considering the nature of the data and business needs.	*(2)Production efficiency data is transferred to monitoring systems every 30 minutes.*
Implement effective and efficient data transferring technologies/tools which do not cause distortions or omissions to data.	*The legacy system migration workbench in SAP (LSMW) facilitates smooth data transferral without creating data quality problems.*
Define and implement appropriate input validation rules to notify the data collector/operator about the erroneous values being entered or erroneous values are flagged for clear identification.	*The telephone number field does not accept on- numeric characters.*
Implement flexible data capturing interfaces to accommodate importantly but out of the way data.	*A field exists to record special comments in goods receipts note (GRN).*
Implement and enforce standardized data capturing procedures/ best practices through the data collecting system itself.	*Standard wait times are used in taking blood samples from a patient (e.g. one hour after meals).*
Establish a mitigation mechanism to handle measurement errors and ensure that acceptable error tolerance levels are established.	*Calibrate the equipment on a routine basis.*
Identify barriers for data collection or barriers for data providers and take appropriate actions to remove them.	*Maintain a log file of response failures of a web-based survey and then eliminate the root causes.*
Identify the practices which encourage data providers.	*Reward survey participants.*
Conduct regular training programs for data capturing/entering staff and educate them on possible data capturing problems and how to overcome data entry errors depending on the context.	*(1) Do not restart the scanner when it is hung up while scanning.* *(2) Repeat a telephone number in a different pattern to validate it from the source e.g. for 045 220 371 9, invalidating, repeat it as 04 52 20 37 19.*
Validation metric: *The extent to which required capabilities and processes have been identified and implemented to maintain accuracy to reference sources*	

Table 6.15: Pattern 13 (Accuracy to reference source)

<table>
<tr><th colspan="3">Pattern 14 (Accuracy to reality)</th></tr>
<tr><td colspan="3">Characteristic: Data should truly reflect the real world</td></tr>
<tr><td>Dimension: Accuracy</td><td>Data granularity: Record</td><td>Type: Usage</td></tr>
<tr><td colspan="3">Verification metric: (1) The number of tasks failed or underperformed due a lack of accuracy compared to reality (2) The number of complaints received due to lack of accuracy compared to reality</td></tr>
<tr><td colspan="3">Implementation form: Process-based approach
Implementation of capabilities and processes to maintain accuracy compared to reality</td></tr>
<tr><td colspan="2">Guidelines for DQ processes and capabilities
(English 2009; Eppler 2006)</td><td>Examples</td></tr>
<tr><td colspan="2">Continuously evaluate if the existing data model is sufficient to represent the real world as required by the organizational need and make the necessary amendments to the data model if needed.</td><td>A student who received a concession GO Card is not eligible for a concession fare if s/he terminates candidature before completion of the course. Hence the data model should have an extra attribute for "current status of candidature".</td></tr>
<tr><td colspan="2">Perform regular verification checks and audits on mission critical data to verify that every record has a meaningful existence in the scope of the reality which is useful for the organization,</td><td>(1) All customers existing in the customer master file represent a customer in the customer space open for the organization (non-customers are not in the customer file). (2)"Greg Glass" is recorded as a glass work company but in fact, they are opticians.</td></tr>
<tr><td colspan="2">Perform regular audits on mission critical data to verify that every record has a unique existence in reality.</td><td>It is difficult to find out whether “Professor Andrew" is the one from Columbia University or from the University of Queensland.</td></tr>
<tr><td colspan="2">Ensure that information available in the system is accurate in the context of a particular activity or event.</td><td>The driver details taken from vehicle registration may not be accurate for finding the person who really drove the vehicle when an accident was caused.</td></tr>
<tr><td colspan="3">Validation metric: The extent to which required capabilities and processes have been identified and implemented to maintain accuracy compared to reality</td></tr>
</table>

Table 6.16: Pattern 14 (Accuracy to reality)

<table>
<tr><th colspan="3">Pattern 15 (Precision)</th></tr>
<tr><td colspan="3">Characteristic: Attribute values should be accurate as per linguistics and granularity</td></tr>
<tr><td>Dimension: Accuracy</td><td>Data granularity: Element</td><td>Type: Declarative</td></tr>
<tr><td colspan="3">Verification metric: The number of imprecise values reported in an attribute per thousand records</td></tr>
<tr><td colspan="3">Implementation form: Rules-based approach
Implementation of rules to maintain precision of data</td></tr>
<tr><td colspan="2">Guidelines for DQ rules
(English 2009; Kimball and Caserta 2004; Loshin 2011)</td><td>Examples</td></tr>
<tr><td colspan="2">Ensure the data values are correct to the right level of detail or granularity.</td><td>(1) Price is listed at the penny or weight to the nearest tenth of a gram.
(2) The precision of the values of an attribute is determined according to some general-purpose IS-A ontology such as WordNet.</td></tr>
<tr><td colspan="2">Ensure that data is legitimate or valid according to some stable reference source such as a dictionary/thesaurus/code.</td><td>(1) The spelling and syntax of a description are correct as per the dictionary/thesaurus/code (e.g. NYSIIS Code)
(2) The address is consistent with global address book</td></tr>
<tr><td colspan="2">Ensure that the user interfaces provide the precision required by the task.</td><td>If the domain is infinite (e.g. all rational numbers), then no string format of finite length can represent all possible values.</td></tr>
<tr><td colspan="2">Ensure the data values are lexical, syntactically and semantically correct.</td><td>Examples include ‘Toni Blair’ as UK’s former Prime Minister (lexically wrong); ‘De la Mancha Don Quixote’ as a book title (syntactically wrong); “Germany is an African country” (semantically wrong).</td></tr>
<tr><td colspan="3">Validation metric: The extent to which required rules have been identified and implemented to maintain precision</td></tr>
</table>

Table 6.17: Pattern 15 (Precision)

6.3.5 Patterns of data validity

<table>
<tr><th colspan="3">Pattern 16 (Business rules compliance)</th></tr>
<tr><td colspan="3">Characteristic: Data must comply with business rules</td></tr>
<tr><td>Dimension: Validity</td><td>Data granularity: Element</td><td>Type: Declarative</td></tr>
<tr><td colspan="3">Verification metric: The number of business rule violations reported in an attribute per thousand records</td></tr>
<tr><td colspan="3">Implementation form: Rules-based approach
Implementation of rule management mechanisms to maintain business rules</td></tr>
<tr><td colspan="2">Guidelines for DQ rules
(English 2009; Halle and Ronald 2001)</td><td>Examples</td></tr>
<tr><td colspan="2">Identify data related business rules separately (business rules that determine the value of data elements and business rule that are executed depending on the values of data elements) and organize them into a separate executable data rules engine.</td><td>SAP has price routines to handle price related data
(e.g. when the element is the discount rate, a price routine (procedure) can be maintained to calculate the discount rate considering the rules).
R1: All registered customers get a discount of 6%
R2: All gold customers get a discount of 12%
R3: All purchases greater than $500 get a discount of 5%</td></tr>
<tr><td colspan="2">Implement a stewardship structure for business rules (parallel to a stewardship structure for data) and manage the changes to the rules properly.</td><td>The sales director is responsible for discounts and his approval is needed to change a discount rate. Only the sales manager can change the rules related to discounts.</td></tr>
<tr><td colspan="2">Maintain an error log to identify the problems resulting in the data rules repository where the problematic data records can be identified precisely.</td><td>Rules engines are used.</td></tr>
<tr><td colspan="2">Continuously monitor the root causes for the errors recorded in the log and take preventative actions by amending the rules, fixing the technical defects in the system etc.</td><td>Some trip data is missing for a particular journey in the GO card system and as a result, an unacceptable journey duration resulted. New rules are implemented to process such data using a different set of criteria.</td></tr>
</table>

Validation metric: *The extent to which required rules have been identified and implemented to maintain business rules compliance*

Table 6.18: Pattern 16 (Business rules compliance)

Pattern 17 (Meta-data compliance)		
Characteristic: *Data should comply with its meta-data*		
Dimension: *Validity*	**Data granularity:** *Element*	**Type:** *Declarative*
Verification metric: The number of meta-data violations reported in an attribute per thousand records		
Implementation form: *Rules-based approach* *Implementation of rules to specify and maintain meta-data*		
Guidelines for DQ rules (Byrne 2008; Loshin 2001; Redman 1997)		**Examples**
Domains should be specified by considering all possible value patterns, cases and usage needs which are applicable for a data attribute.		*It is easy to maintain the order number as a numeric value since it can be easily incremented (10000, 10001, and 10002). But it can also be defined as alphanumeric in order to distinguish special cases (e.g. 10000R is a return order of sales order 10000).*
Maintain valid values/value ranges/value lists for attributes.		*Gender can be M or F.* *The country is taken from the existing list of countries.* *The birth date cannot be a future value.* *The salary range for Level 4 employees must be between $40000-60000.*
Use number ranges for critical data elements.		*Sales orders: 10000001 to 1999999* *Purchase orders: 50000001 to 59999999*
Maintain the possible synonyms and abbreviations which could be accepted as valid values.		*Post Box , PO BOX, PO Box, BOX etc.*
Explicitly mention what values and characters are not permitted in the attribute		*Username can contain only A-Z.* *No blank spaces are allowed for a credit card number.*

Explicitly mention the minimum and/or a maximum number of characters, or any other requirements such as case sensitivity, that an attribute value should meet.	*A password should contain a minimum of 8 characters including one numeric and one capital.*
Maintain values based on specific formats as defined by the stakeholders, standards, best practices or agreements.	*Time should be in 24-hour clock format.* *The date should be in DD/MM/YYYY format.*
An appropriate measurement scale should be maintained for quantities and volumes.	*(1) Currency for price values* *(2) Kg/g/mg for weights* *(3) Litres for volumes*
Documentation for meta-data is available online for the users.	*Data dictionary, data catalogue*
Validation metric: *The extent to which required rules have been identified and implemented to maintain meta-data compliance*	

Table 6.19: Pattern 17 (Meta-data compliance)

Pattern 18 (Standards and regulatory compliance)		
Characteristic: *All data processing activities should comply with the policies, procedures, standards, industry benchmark practices and all regulatory requirements that the organization is bound by*		
Dimension: *Validity*	**Data granularity:** *Information object*	**Type:** *Usage*
Verification metric: *(1) The number of tasks failed or underperformed due non-adherence to standards and regulations (2) The number of complaints received due to non-adherence to standards and regulations*		
Implementation form: *Process-based approach* *Implementation of capabilities and processes to maintain standards and regulatory compliance*		
Guidelines for DQ processes and capabilities (English 2009; HIQA 2011; McGilvray 2008)		**Examples**
Identify the policies, procedures, standards, benchmark practices and any regulatory requirements that an information object is bound by,		*Each person's compensation criteria must be determined in accordance with the Annuities Based on Retired or Retainer Pay Law.*

Ensure that all data processing activities are well defined and documented based on specified policies, procedures, standards, benchmarks and regulatory requirements.	*The process of making a damage estimate is well defined based on industry benchmarks.*
Ensure that the application programs cater for standards and regulatory compliance.	*A software program to make damage estimates which includes all benchmark data is utilized.*
Regularly monitor the data processing activities and identify the problems and inefficiencies so that corrective and preventive actions can be taken.	*Frequent delays in time sheet approvals result in delayed payments.*
Signs should be specified, standardized and universally used.	*Signs used to demarcate different processes are standardised across the production floor.*
Relevant standards, procedures, policies, and regulations should be communicated to the users effectively.	*Staff training on organizational quality management process*
Ensure that proper conversion tables are maintained and used in converting attribute values to different measurement bases.	*Metric conversion tables are used to convert lbs to kgs.*
Validation metric: *The extent to which required capabilities and processes have been identified and implemented to maintain standards and regulatory compliance*	

Table 6.20: Pattern 18 (Standards and regulatory compliance)

Pattern 19 (Statistical validity)		
Characteristic: Computed data must be statistically valid.		
Dimension: *Validity*	**Data granularity:** *Information object*	**Type:** *Usage*
Verification metric: *(1) The number of tasks failed or underperformed due to a lack of statistical validity in data (2) The number of complaints received due to a lack of statistical validity of data*		
Implementation form: *Process-based approach* *Implementation of capabilities and processes to maintain statistical validity*		
Guidelines for DQ processes and capabilities (HIQA 2011; Lyon 2008)		**Examples**

Establish the population of interest unambiguously with appropriate justification (maintain documentation).	*Both credit customers and cash customers are considered for a survey on customer satisfaction.*
Establish an appropriate sampling method with appropriate justification.	*Stratified sampling is used to investigate the drug preference of the medical officers.*
Establish statistical validity of samples avoiding over coverage and under coverage (maintain documentation.)	*Samples are taken from all income levels in a survey on vaccination.*
Maintain consistency of samples in case longitudinal analysis is performed (maintain documentation).	*The same population is used over time to collect epidemic data for a longitudinal analysis.*
Ensure that valid statistical method are used to enable valid inferences about data, valid comparisons of parameters and generalizability of the findings.	*Poisson distribution is used to make inferences since data generating events occurred at fixed intervals of time and/or space.*
Ensure that the acceptable variations for estimated parameters are established with appropriate justifications.	*A 95% confidence interval is used in estimating the mean value.*
Ensure that appropriate imputation measures are taken to nullify the impact of problems relating to outliers, data collection, and data collection procedures and that the edit rules are defined and maintained.	*Incomplete responses are removed from the final data sample.*
Validation metric: *The extent to which required capabilities and processes have been identified and implemented to maintain statistical validity of data*	

Table 6.21: Pattern 19 (Statistical validity)

6.3.6 Patterns of reliability and credibility of data

Pattern 20 (Source quality)		
Characteristic: *Data used is from trusted and credible sources.*		
Dimension: *Validity*	**Data granularity:** *Information object*	**Type:** *Usage*
Verification metric: *(1) The number of tasks failed or underperformed due to a lack of source quality (2) The number of complaints received due to a lack of source quality*		

Implementation form: *Process-based approach* *Implementation of capabilities and processes to maintain source quality*	
Guidelines for DQ processes and capabilities (English 2009; Loshin 2004; Scannapieco and Catarci 2002)	**Examples**
Assess the reputation of data sources.	*The Central Bank is the best source from which to get daily exchange rates.*
Evaluate the remedies for non-compliance of data.	*Review any remedies given by the source organization to mitigate the losses in case the information is of low quality.*
Rely on shared information sources created/ recommended/used by the organizations operating in the industry.	*In performing portfolio analysis most organizations use the risk factors produced by a central body of the economy (e.g. a Central Bank.)*
Validation metric: *The extent to which required capabilities and processes have been identified and implemented to maintain reliability and credibility*	

Table 6.22: Pattern 20 (Source quality)

Pattern 21 (Objectivity)		
Characteristic: *Data are unbiased and impartial*		
Dimension: *Reliability and credibility*	**Data granularity:** *Information object*	**Type:** *Usage*
Verification metric: *(1) The number of tasks failed or underperformed due to biased and partial data (2) The number of complaints received due to biased or partial data*		
Implementation form: *Process-based approach* *Implementation of capabilities and processes to maintain objectivity*		
Guidelines for DQ processes and capabilities (English 2009; Eppler 2006)		**Examples**
Identify all the factors that make a particular piece of data/information biased for the intended use and take preventive actions to eliminate them.		*A written questionnaire is better than a face to face interview in getting sensitive personal data.*
Design and execute preventative actions for all possible information distortions (malfunctioning or		*A dual coder approach to code qualitative data reduces bias.*

personal bias) which may be caused by information /data collectors.	
Design and execute preventative actions for all possible information distortions (malfunctioning or personal bias) which may be caused by information /data transmitters.	*After a survey is performed, each participant is contacted individually by a party (other than the person who conducted the survey) and to randomly verify if the participant's real responses have been marked properly.*
Validation metric: *The extent to which required capabilities and processes have been identified and implemented to maintain objectivity*	

Table 6.23: Pattern 20 (Source quality)

Pattern 22 (Traceability)		
Characteristic: *The lineage of the data is verifiable.*		
Dimension: *Reliability and credibility*	**Data granularity:** *Record*	**Type:** *Usage*
Verification metric: *Records*		
Implementation form: *Process-based approach* *Implementation of capabilities and processes to maintain traceability*		
Guidelines for DQ processes and capabilities (Eppler 2006; Stvilia et al. 2007)		**Examples**
Identify the lifecycle of dynamic data in the organization.		*Data is tracked from the creation of a sales order, updates to the sales order, and its usage, to its archival.*
Maintain provenance records for events such as creation, update, transcription, abstraction, validation and transforming ownership, if the data are dynamic.		*An inventory system shows the current stocks and keeps records for all the transactions to which the stocks are subjected.*
In case multiple sources are available for the same data/information, implement a traceability mechanism to view all versions from multiple sources.		*Document management systems*
Maintain proper protocols/standards/policy to archive data.		*Every invoice is archived after 120 days of payments.*

Maintain versions of data records where necessary	*Customer versions*
Validation metric: *The extent to which required capabilities and processes have been identified and implemented to maintain traceability*	

Table 6.24: Pattern 22 (Traceability)

6.3.7 Patterns of data consistency

Pattern 23 (Uniqueness)		
Characteristic: Data is uniquely identifiable.		
Dimension: *Consistency*	**Data granularity:** *Record*	**Type:** *Declarative*
Verification metric: *The number of duplicate records reported per thousand records*		
Implementation form: *Rules-based approach* *Implementation of rules to prevent/detect duplicate record*		
Guidelines for DQ rules (Byrne 2008; Loshin 2006; Talburt 2011)		**Examples**
Implement a key for every relation so that each record (entity) can be uniquely identified.		*Establish a key constraint.*
Ensure that the unique key is not-null at any cost.		*The employee ID which is the key of the employee table is not null at any cost.*
Ensure the primary key is brief so that it supports efficient indexing and foreign keys.		*It is cumbersome to create foreign keys with a composite primary key with a large number of attributes.*
Ensure the primary key attributes take simple values free from embedded spaces, special characters or length data types.		*Customer ID: 2006 RIO 764536 is not an efficient value for a PK field.*
Implement surrogate keys when the natural key is cumbersome due to a large number of attributes in the key or due to complex values.		*Composite key: Longitude, latitude, date, and time* *Surrogate key: 1345234567*
Ensure that the values for primary keys are not re-used (after deleting a record) or changed.		*Deleted customer IDs are not used again for new customers.*

Through the use of alternative keys (other candidate keys) ensure that the same entity is not recorded twice under different unique identifiers.	*The same customer is entered under different customer IDs but the customer ABN is the same.*
Use auto-incrementing values in primary key fields whenever possible.	*The customer ID is an auto-incrementing number.*
Use validation rules based on entity resolution algorithms to detect and merge the duplicate records when the same entity is recorded under more than one key.	*Record linkage tools*
When using bar codes, standardise the barcode generation process to ensure that bar codes are not reused (reprinted.)	*UPC standard*
Validation metric: *The extent to which required rules have been identified and implemented to maintain uniqueness*	

Table 6.25: Pattern 23 (Uniqueness)

Pattern 24 (Non-redundancy)		
Characteristic: *Data is recorded in exactly one place.*		
Dimension: *Consistency*	**Data granularity:** *Record*	**Type:** *Declarative*
Verification metric: *The volume of redundant data as a percentage of total data*		
Implementation form: *Rules-based approach* *Implementation of rules to prevent/detect redundant data*		
Guidelines for DQ rules (English 2009; Gatling 2007; McGilvray 2008)		**Examples**
Design the database schema eliminating the causes for redundancies		*All customers are in the customer table and all quotation holders too go to the customer table (with a specific customer type)*
Ensure that there are no redundant records when there are multiple databases.		*The organization has different customer bases maintained in different databases. However, each customer is available in only one database.*
Ensure that the same entity is not originally captured more than once in the systems for different purposes.		*The medical insurance system refers employee bank details from the payroll.*

Ensure that all temporary tables are refreshed periodically and systematically.	*A backup for the employee table is created as a temporary employee table for a specific purpose and it is still in the database.*
Validation metric: *The extent to which required rules have been identified and implemented to maintain non-redundancy*	

Table 6.26: Pattern 24 (Non-redundancy)

Pattern 25 (Semantic consistency)		
Characteristic: *Data is semantically consistent.*		
Dimension: *Consistency*	**Data granularity:** *Element*	**Type:** *Declarative*
Verification metric: *The number of semantically inconsistent data reported per thousand records*		
Implementation form: *Rules-based approach* *Implementation of rules to maintain semantic consistency*		
Guidelines for DQ rules (Byrne 2008; English 2009; Stvilia et al. 2007)		**Examples**
Ensure the labels for data attributes are consistent between all relations and databases used within the organization.		*All orders placed by the customers are called "sales order" in all tables.*
Ensure the labels for data attributes are consistent between the organization and the external parties dealing with the organization.		*All orders placed by the customers are called "sales order" in the ERP database and CRM databases.*
Ensure that data labels are consistent between database and application program interfaces.		*All orders placed by the customers are called "sales order" in the database as well as ERP system interfaces.*
Continuously educate users on the semantics of terminology used for data elements and business processes and how it is important to adhere to the semantics in achieving organizational goals.		*Training sessions are held for end users of the information systems.*
Ensure that a data dictionary/data catalogue is available for the users to refer to for the semantics of data.		*Users can access the data dictionary from application program interfaces.*

Maintenance of online help facilities with standard vocabularies in application programs.	*The data dictionary provides technical data as well as the semantics of data.*
Validation metric: *The extent to which required rules have been identified and implemented to maintain semantic consistency*	

Table 6.27: Pattern 25 (Semantic consistency)

Pattern 26 (Value consistency)		
Characteristic: *Data values are consistent and do not provide conflicting or heterogeneous instances.*		
Dimension: *Consistency*	**Data granularity:** *Element*	**Type:** *Declarative*
Verification metric: *The number of inconsistent data values reported in an attribute per thousand records*		
Implementation form: *Rules-based approach* *Implementation of rules to prevent/detect null values in mandatory attributes*		
Guidelines for DQ rules (Gatling 2007; Kimball and Caserta 2004; Loshin 2001)		**Examples**
For critical data elements provide standard classifications (values lists) for data entry interfaces and restrict arbitrary values across the system.		*Country and city are taken from a standard list.* *Generally accepted industry classifications are used to analyse customers industry-wide (Education, Banking & Finance, Medical, Manufacturing etc)*
When data elements are combined for specific identification, management or accounting purposes, standardize such combinations and use them across the system.		*Customer and sales order are combined for identification purposes* *Costs of wastage are managed through associating them with the individual orders which incur them.*
Define data attributes in such a way that data values are atomic and hence consistency can be maintained for any form of aggregation or consolidation.		*A name is divided into first name, middle name, and last name.*
Maintain consistency in using a unit of measures across different tables and different databases.		*The sales price is in $ in the sales table and accounts receivable ledger.*

Validation metric: *The extent to which required rules have been identified and implemented to maintain value consistency*

Table 6.28: Pattern 26 (Value consistency)

Pattern 27 (Format consistency)		
Characteristic: *Data formats are consistently used*		
Dimension: *Consistency*	**Data granularity:** *Element*	**Type:** *Declarative*
Verification metric: The number of inconsistent data formats reported in an attribute per thousand records		
Implementation form: *Rules-based approach* *Implementation of rules to prevent/detect null values in mandatory attributes*		
Guidelines for DQ rules (English 2009; Redman 1997; Stvilia et al. 2007)		**Examples**
Maintain standardised and consistent formats and/or masks for data elements across tables and databases.		*Telephone number:* *country code/area code/number* *Address: house number, street, suburb, state, country*
Maintain standardised data structures across tables and databases so that the data elements will be consistent		*The customer record has the same structure in all systems in which it is being used.*
Maintain consistent and compatible encoding and decoding standards across different applications.		*ASCII, UTF-8, XML*
Validation metric: *The extent to which required rules have been identified and implemented to maintain format consistency*		

Table 6.29: Pattern 27 (Format consistency)

Pattern 28 (Referential integrity)		
Characteristic: *Data relationships are represented through referential integrity rules.*		
Dimension: *Consistency*	**Data granularity:** *Record*	**Type:** *Declarative*
Verification metric: *The number of referential integrity violations per thousand records*		

Implementation form: *Rules-based approach* *Implementation of rules to maintain referential integrity of data*	
Guidelines for DQ rules (English 2009; Loshin 2001; Price and Shanks 2005a)	**Examples**
Implement foreign keys across tables through database integrity rules or maintain data integrity through application program level rules	*Implementation of foreign keys*
Implement proper validation rules and automated suggestions of values based on popular value combinations, to prevent incorrect references for foreign keys.	*The attribute of customer_zip_code of the customer relation contains the value 4415, instead of 4445. Both zip codes exist in the zip code relation.*
Implement validation rules for foreign keys of relevant tables in case of data migrations.	*Error logs are generated for foreign key violations.*
Implement proper synchronizing mechanisms to handle data updates when there are concurrent operations or distributed databases.	*Establish lock mechanisms for data objects while data is being updated.*
Ensure the consistency of the data model when changes are made to the process model (software.)	*The data dictionary provides the FDs and CFDs.*
Validation metric: *The extent to which required rules have been identified and implemented to maintain referential integrity*	

Table 6.30: Pattern 28 (Referential integrity)

6.3.8 Patterns of data usability and interpretability

Pattern 29 (Usefulness and relevance)		
Characteristic: *Data is useful and relevant for the task at hand.*		
Dimension: *Usability and interpretability*	**Data granularity:** *Information object*	**Type:** *Usage*
Verification metric: *(1) The number of tasks failed or underperformed due to the lack of usefulness and relevance of data (2) The number of complaints received due to the lack of usefulness and relevance of data*		
Implementation form: *Process-based approach* *Implementation of capabilities and processes to maintain usefulness and relevance*		

Guidelines for DQ rules	**Examples**
Define the content of the information object based on the user requirements (as required by the task at hand), considering the context and all other compliance requirements, so that the information is relevant and legitimate.	*The customer invoice should contain information for the customer to understand his liability, for the delivery person to understand the point of delivery and for the tax department to verify the applicable tax amount.*
Regularly monitor the changes to the internal operational environment (business process changes etc.), find out the new information requirements emerging due to the changes and provide for them by amending the information structures.	*The time stamp became an important attribute for goods receipts notes (GRNs) when Lean manufacturing started as all raw materials are expected to be received by six hours before production (GRN-record and the time stamp attribute).*
Regularly monitor the changes in the external environment to find out new information requirements which emerge due to such changes and provide for such data needs.	*Competitors' rates have become important to price existing products during a recession period since the traditional costing method does not give a competitive price.*
Regularly check with knowledge workers to find out how their operations and decisions can be performed better with new data available to them and provide for such data in the information system.	*An hourly working progress report is useful in identifying the bottlenecks in production lines, then balance the lines accordingly.*
Monitor and measure the user satisfaction about the information provided.	*User satisfaction survey*
Validation metric: *The extent to which required capabilities and processes have been identified and implemented to maintain objectivity*	

Table 6.31: Pattern 29 (Usefulness and relevance)

Pattern 30 (Understandability)		
Characteristic: *Data is understandable*		
Dimension: *Usability and interpretability*	**Data granularity:** *Information object*	**Type:** *Usage*
Verification metric: *(1) The number of tasks failed or underperformed due to the lack of understandability of data (2) The number of complaints received due to the lack of understandability of data*		

Implementation form: *Process-based approach* *Implementation of capabilities and processes to maintain understandability*	
Guidelines for DQ processes (English 2009; Eppler 2006; McGilvray 2008)	**Examples**
Ensure that appropriate signs and language is used to strengthen the readers' understanding of the information object.	*Poor, good and excellent is more suitable than 1, 2 and 3 as ratings to compare two factors.*
Avoid any possibility of ambiguity in understanding data with the inclusion of footnotes, legend etc.	*Footnote: Total price includes GST.*
Provide supplements to understand the content of non-text and non-numeral information (e.g. images).	*A location in a plan can be identified by the coordinates.*
Ensure that data are concisely represented without being overwhelming.	*Focus on one topic.*
Convenient and user-friendly (more natural) formats are used for structured attributes like date, time, telephone number, tax ID number, product code and currency amount.	*U.S. phone number formats [+1(555)999-1234]*
Appropriate fonts and styles are used to improve the clarity of the content.	*Headings and totals are formatted in bold.*
Validation metric: *The extent to which required capabilities and processes have been identified and implemented to maintain understandability*	

Table 6.32: Pattern 30 (Understandability)

Pattern 31 (Appropriate presentation)		
Characteristic: *The data presentation is aligned with its use.*		
Dimension: *Usability and interpretability*	**Data granularity:** *Information object*	**Type:** *Usage*
Verification metric: *(1) The number of tasks failed or underperformed due to the lack of appropriate presentation of data (2) The number of complaints received due to the lack of appropriate presentation of data*		
Implementation form: *Process-based approach* *Implementation of capabilities and processes to maintain appropriate presentation*		

Guidelines for DQ processes (English 2009; Loshin 2004; McGilvray 2008; Redman 1997)	**Examples**
Ensure that universally accepted standard formats are used to maintain the compatibility of information across organizations and across time.	*A patient's diagnostic card generated in one hospital is compatible with another hospital.*
Ensure that information can be aggregated or combined through the use of compatible formats.	*Product wise monthly sales report can be generated by combining the sales reports of three subsidiaries.*
Ensure that the data presentations are familiar to the users even if the application platform is changed.	*A quotation in the SAP system is sent to the RTW system through a message and displayed in the same presentation format.*
Ensure the media and style of presentation are appropriate for the target group.	*(1) A step by step written instruction list in a document is pitched at the level appropriate for a software engineer.* *(2) A video display is designed appropriately for kids.*
Ensure that the presentation formats are flexible and there is a proper mechanism to accommodate changes easily	*An invoice document may require additional space to include authorization evidence.*
Validation metric: *The extent to which required capabilities and processes have been identified and implemented to maintain appropriate presentation*	

Table 6.33: Pattern 31 (Appropriate presentation)

Pattern 32 (Interpretability)		
Characteristic: *Data should be interpretable*		
Dimension: *Usability and interpretability*	**Data granularity:** *Information object*	**Type:** *Usage*
Verification metric: *(1) The number of tasks failed or underperformed due to the lack of interpretability of data (2) The number of complaints received due to the lack of interpretability of data*		
Implementation form: *Process-based approach* *Implementation of capabilities and processes to maintain interpretability*		
Guidelines for DQ processes		**Examples**

(HIQA 2011; Lyon 2008)	
Standardize the interpretation process by clearly stating the criteria for interpreting results so that an interpretation on one data set is reproducible.	*A 10% drop in production efficiency is a severe decline which needs quick remedial actions.*
Facilitate the process of interpretation based on users' tasks at hand.	*A traffic light system to indicate the efficiency of a production line to the workers, a detailed efficiency report to managing the production, a concise efficiency report for production line supervisors*
Design the structure of information in such a way that further format conversions are not necessary for interpretations.	*A rating scale (poor, good and excellent) is better than (1, 2 and 3) to rate a service level.*
Ensure that information is consistent between units of analysis (organizations, geographical areas, populations of concern etc.) and allows comparisons to be made over time.	*The number of doctors per person is used to compare health facilities between regions.* *The same populations are used over time to analyse epidemic growth over time.*
Use appropriate visualization tools to facilitate interpretation of data through comparisons and contrasts.	*Treemaps, bar charts, line graphs*
Validation metric: *The extent to which required capabilities and processes have been identified and implemented to maintain understandability*	

Table 6.34: Pattern 32 (Interpretability)

Pattern 33 (Information value)		
Characteristic: Quality information provides a business value to the organization.		
Dimension: *Usability and interpretability*	**Data granularity:** *Information object*	**Type:** *Usage*
Verification metric: *(1) The number of tasks failed or underperformed due to the lack of business value delivered by the information (2) The number of complaints received due to the lack of business value delivered by the information*		
Implementation form: *Process-based approach* *Implementation of capabilities and processes to maintain information value*		
Guidelines for DQ processes	**Examples**	

(Loshin 2001; McGilvray 2008)	
Continuously assess the relevance and the usefulness of existing data to the organizational goals (strategic level).	*(1) What is the cost of poor quality customer data to the organization of concern?* *(2) What revenue can be generated from the data?*
Continuously assess the usefulness of information based on the tasks at hand (operational level).	*Can we predict our future market share from the existing market information?*
Monitor and measure if the intended goal of the data presentation and/or interpretation is achieved.	*(1) Employee efficiency data is displayed in a dashboard to motivate employees. The effectiveness of this display can be measured by examining the efficiency gain of each employee.* *(2) Has the given sales forecast for the last three years been reasonably accurate compared to actual sales?*
Validation metric: *The extent to which required capabilities and processes have been identified and implemented to maintain information value*	

Table 6.35: Pattern 33 (Information value)

6.4 Compatibility with Meta Object Facility (MOF)

In Figure 6.8 below we present the levels of data quality requirements modelling in comparison to the three layers of conceptualization as specified in Meta Object facility by Omg (2008). What lays on top is the Meta-Meta model (M3) where we have modelled the basic concepts of a data quality requirement.

In M2: layer we instantiate the M3 model using 33 data quality characteristics and create generic data quality patterns which can be used as a meta-model to model data quality requirements in organizational context. We argue that any data quality requirements of structured data can be modelled using these 33 generic patterns in the M2 layer.

M1 model is an organizational specific DQ requirements model generated using the generic data quality patterns in M2 layer. This is the DQ requirements model for the organization in concern and it is expected to deliver the required knowledge to systematically manage DQ. (This model is equivalent to a business process model used in business process management)

M0 is the actual realisation of rules and processes to manage DQ in an organization. (This is equivalent to the actual business processes established in the organization)

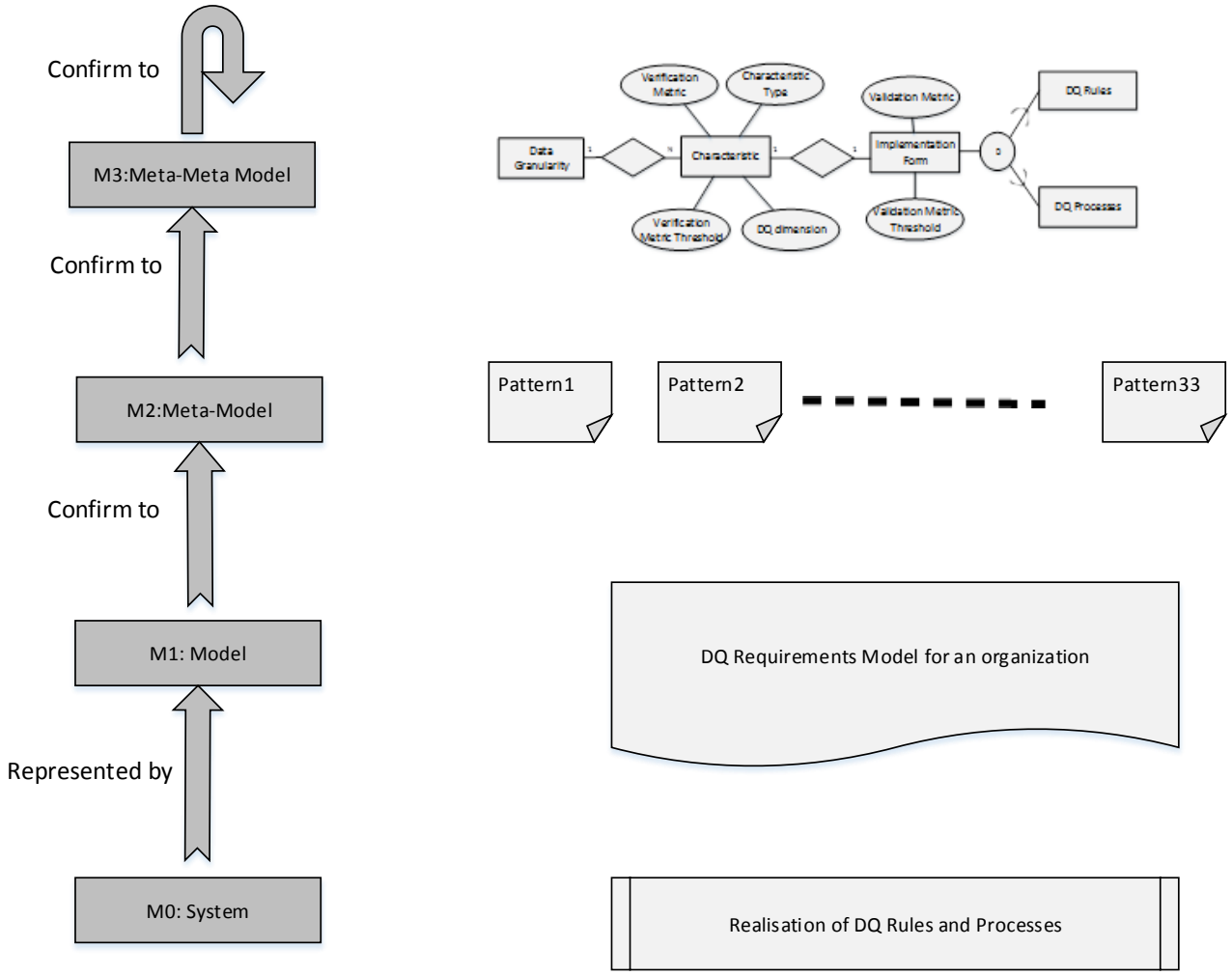


Figure 6.8: Comparison of DQ requirements modelling with MOF

6.5 Summary

In this chapter, we developed thirty-three patterns of data quality requirements. These patterns were presented using a generic terminology as suggested by Rolland et al. (1998) so that they can be adapted to contexts depending on the need. This repository of patterns can be considered as an M2: meta-model for data quality requirements as explained in Figure 6.8 above.

In Section 2.3.1 we pointed out that every meta-model is designed to cater for its purpose and it contains constructs used in the meta-model which has embedded knowledge to fulfil this purpose. In this case, each pattern can be used to model DQ requirements and such a pattern consists of the

required knowledge to model those data quality requirements in an actionable form. Therefore an organization can create their own data quality requirements model (M1: model) by further instantiating the 33 patterns based on their context.

In Chapter 7 we conduct a descriptive validation of the 33 patterns to check the completeness of the repository while in Chapter 8 we will introduce a methodology to use the patterns.

Chapter 7

DESCRIPTIVE EVALUATION OF DATA QUALITY PATTERNS

7.1 Overview

The objective of this chapter is to present the descriptive validation performed for the pattern repository. The descriptive validation provides evidence on whether the DQ patterns are capable of representing any DQ requirement of structured data. In other words, we want to check the completeness (sufficiency) of the pattern repository in representing DQ requirements. Further, this descriptive validation revealed the relationships between patterns. We present these relationships accordingly as a by-product of this validation.

7.2 Introduction

Following Hevner et al. (2004), we recognize the need to evaluate artefacts, and thus planned for two evaluation processes: viz. (1) Descriptive evaluation (2) Empirical evaluation as explained in Chapter 3. In this chapter, we present the descriptive evaluation that we conducted to evaluate the pattern repository. Our aim is to evaluate the coverage of the pattern repository in representing data quality requirements and identify requirements that cannot be modelled by any of the 33 patterns.

According to Hevner et al. (2004), descriptive evaluation can be performed in two aspects:

(1) Informed Argument: "Use information from the knowledge base (e.g., relevant research) to build a convincing argument for the artefact's utility"

(2) Scenarios: "Construct detailed scenarios around the artefact to demonstrate its utility"

In this validation, we use information from the literature to build a convincing argument for the artefact's utility in representing the DQ requirements. As pointed out in Chapter 2, DQ problems and DQ requirements are two sides of the same coin where the DQ characteristic is a central concept in defining requirements and problems. Therefore a DQ pattern can be viewed from a DQ problem perspective as well as a DQ requirement perspective. Thus in this validation, we consider the DQ problem perspective by referring to the literature on DQ problems and the DQ requirements perspective by referring to published repositories of data quality requirements.

7.3 Approach

We use 8 sources from the literature on DQ problems of structured data (Eppler 2006; Garvin 1988; Ge and Helfert 2013; Kim et al. 2003; Lee et al. 2009; Lesca and Lesca 1995; Oliveira et al. 2005; Strong et al. 1997). In this selection we considered the following criteria:

1) A representation of DQ problem literature during the last three decades (covering a wider scope of the lifetime of the DQ domain)
2) The diversity of DQ problems (by avoiding similar classifications)
3) The number of citations to evaluate the credibility of the work (>100 as a guideline)

These sources contained 213 data quality problems. The authors have identified these DQ problems through analysing the data and information quality problems in organizations through empirical studies. By referring to these DQ problem categories, we were motivated to assess if these DQ problems have a corresponding pattern in our pattern repository. In other words, we checked if the existing patterns were sufficient to address the problems discussed in the literature.

From the DQ requirement perspective, we used three real-world rule repositories from three different domains (considering the diversity of context). Namely health data, customer data, and defence data from the Canadian Institute for Health Information-CIHI (Richards and White 2013), IBM (Smith 2012) and the US Department of Defence (DOD 2014) respectively.

- The CIHI repository defines DQ requirements in the form of guidelines to implement in the respective data generation and usage processes.
- The IBM repository focused on DQ rules implemented at the application program level as validation rules to prevent the occurrence of bad data.
- The DOD repository is a general business rules repository that contains rules for data capturing and data processing.

The three sources contained 197 data quality requirements in total.

This evaluation was performed using the interpretive philosophy (Klein and Myers 1999) as also used earlier by Walsham (1993) in interpreting information systems. In the evaluation, we produced an understanding of how the artefact represents the context (first objective) and, in turn, how the context can be influenced to reshape the artefact (second objective).

In the evaluation process, the following steps were followed with regards to each DQ problem or rule:

1) Record the statement that states the DQ problem or requirement as described in the source
2) If the statement contains overloaded DQ problems or requirements split them into atomic problems and rules
3) Identify the relevant DQ Pattern for each problem/requirement by comparing it with the definition pf DQ characteristic of the pattern

For example consider the following DQ problem

"Limited usefulness of information due to an overload of information caused by a lack of cleansing or maintenance activities"

The problem can be split into three atomic problems that can be mapped with three patterns

1. The usefulness of information is limited (Usefulness and relevance : Pattern 29)
2. There is too much information (Data volume : Pattern 4)
3. There is a lack of regular cleansing or periodic maintenance (Data maintainability: Pattern 6)

This analysis was performed by two researchers independently and the independent ratings were evaluated using Cohen's Kappa, with a result of 89.6%, indicating high confidence of inter-rater agreement (Carletta 1996). Coding disagreements were then discussed between the two researchers until consensus was reached and necessary changes were made to the pattern repositories. In the following section, we present the outcome of the validation.

7.4 Results of the descriptive evaluation

A part of the analysis is presented as examples in Table 7.1: Mapping of DQ requirements and problems to patterns

, while the summary of the descriptive evaluation is presented in Figure 7.1. What follows, we provide a discussion of the most frequent and least frequent pattern mappings as well as the improvements to the patterns we made in this evaluation phase.

DQ requirement	Pattern
Edit reports for users are easy to understand.	P30: Understandability - The data is understandable.
Major reports are released on schedule.	P9: Data punctuality - *Data should be available at the time of its intended use.*
Geographical data is collected using the Standard Geographical Classification (SGC).	P26: Value consistency - *Data values are consistent and do not provide conflicting or heterogeneous instances.*

Documentation on historical changes to the data holding exists and is easily accessible.	P22: Traceability - *The lineage of the data is verifiable.*
The “accounting classification reference number approval mount” can be a negative number.	P16: Business rules compliance - *Data should comply with business rules.*
The acquisition original asset recorded cost amount must be obtained from the acquiring document.	P13: Accuracy to reference source - *Data should agree with an identified source.*
Each acre value must be recorded to two decimal places.	P17: Meta-data compliance - *Data should comply with its metadata.*
Each person's compensation eligibility must be determined in accordance with the Annuities Based on Retired or Retainer Pay law.	P18: Standards and regulatory compliance - *All data processing activities should comply with the policies, procedures, standards, industry benchmark practices and all regulatory requirements that the organization is bound by.*
If there is not an advancement of an appropriation for budget authority, then the Advanced Flag value must be 'X'.	P16: Business rules compliance - *Data should comply with business rules.*
DQ Problem	Pattern
Data relationships are missing.	P28: Referential integrity - *Data relationships are represented through referential integrity rules.*

Changing data needs: As information consumers' tasks and the organization environment (such as new markets, new legal requirements, and new trends) change, the information that is relevant and useful changes.	P29: Usefulness and relevance - *The data is useful and relevant for the task at hand.*
Entries are obsolete or outdated.	P12: Data freshness - *Data which is subjected to changes over the time should be fresh and up-to-date with respect to its intended use.*
There are inconsistent data formats or naming conventions.	P27: Format consistency - *Data formats are used consistently.*
There is missing data where the null-not-allowed constraint should be enforced.	P1: Completeness of mandatory attributes - *The attributes which are mandatory for a complete representation of a real world entity must contain values and cannot be null.*
Special characters are used (space, no space, dash, parenthesis, in a social security number or phone number).	P17: Meta-data compliance - *Data should comply with its metadata.*
Wrong derived-field data (due to an error in functions for computing data in a derived field).	P16: Business rules compliance - *Data should comply with business rules.*
Unauthorised manipulation of stored data (deletion, modification).	P10: Data access control - *The access to the data should be controlled to ensure it is*

	secure against damage or unauthorised access.
There is ambiguous data, due to the use of abbreviation (Dr. for doctor or drive).	P25: Semantic consistency - *Data is semantically consistent.*
There are wordy reports that have no logical flow.	P32: Interpretability - *Data should be interpretable*.

Table 7.1: Mapping of DQ requirements and problems to patterns

As per the analysis, *Pattern 17* recorded the highest number of representations (45) in the studied sources. *Pattern17* relates to the characteristic "Metadata compliance" within the "Validity" dimension. The validation process uncovered 33 mappings to rules and 12 mappings to DQ problems. Thus, *Pattern 17* has the highest number of mappings to the rule repositories, indicating that those organizations have focused on ensuring the implementation of DQ requirements related to meta-data.

The second highest number of mappings (38) in the validation process was recorded for *Pattern 13* This pattern relates to the characteristic "Accuracy to reference source" within the "Accuracy" dimension. This characteristic emphasizes on data capturing. We take this high frequency of 38 (22 from the rules repositories and 16 from DQ problems) as an indicator that failure to implement this pattern is likely to cause significant DQ problems in organizations.

The third highest number of mappings in the validation process was recorded for *Pattern 16* (35 occurrences). This pattern relates to "Business rules compliance" within the "Validity" dimension. This pattern implies that a rule management approach should be taken to organize and manage business rules related to data values (rules which determine the values of data elements and which get triggered depending on data values).

As per this analysis, *Pattern 17, Pattern 13 and Pattern 16* appear to have a more prominent stake in organizational data quality management efforts, as reflected by the frequency of mappings to DQ problems and real-world DQ rule repositories considered in this analysis.

Further, *Pattern 18, Pattern 19, Pattern 25, Pattern 26 and Pattern 28* also recorded a relatively higher number of occurrences thus implying the respective characteristics of "Standards and regulatory compliance", "Statistical validity", "Semantic consistency", "Value consistency", "Referential integrity" have a greater role in achieving high-quality data in organizations.

Further, *Pattern 2* (completeness of optional attributes) and Pattern *24* (Non-redundancy) have received the lowest frequencies of occurrence. However, it should be noted that since these patterns were originally created using data quality characteristics which are well grounded in the literature, they cannot be discarded from the repository even though they have received a low frequency of occurrence in this validation.

It should be noted that all the patterns have at least one occurrence in DQ problems or DQ rules and in turn all the DQ problems and rules had at least one corresponding pattern. Hence this provides an authentic validation on the rigor of the methodology adopted in synthesising the patterns.

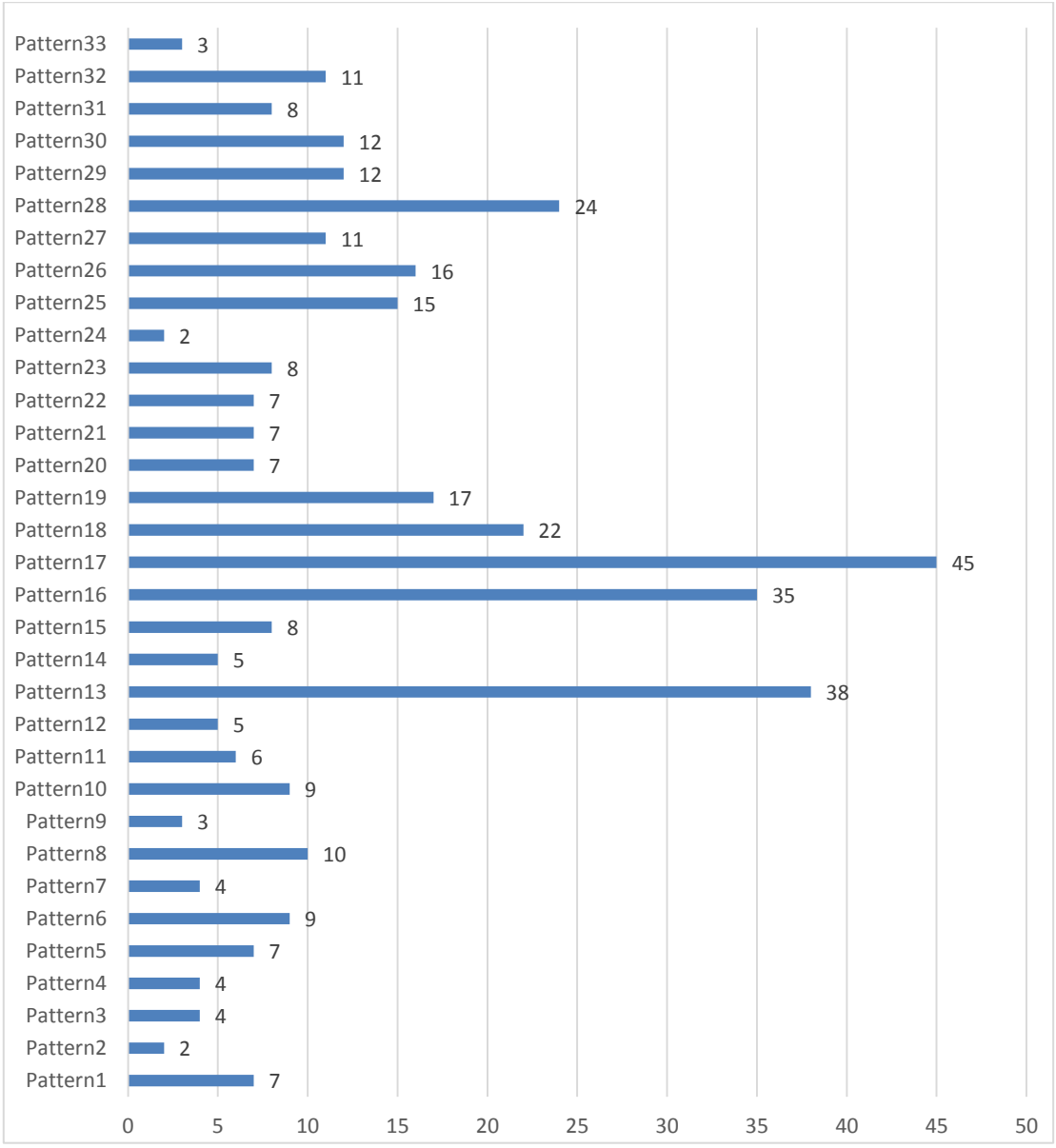


Figure 7.1: Summary of the descriptive evaluation

7.4.1 Refinements to the patterns introduced through the evaluation

The validation process, using the DQ problems and requirements, provided valuable insights to improve the patterns. The validation of the DQ requirements, in particular, allowed us to improve associated implementation guidelines of many patterns. Specifically, the CIHI repository (Richards and White 2013) emphasizes various activities that should be undertaken to maintain the quality of data capture and transfer which helped us to improve the implementation guidelines of *Pattern 13*. For instance, standardization of data capturing procedures, the establishment of data capturing points and data transferring frequencies, and practices for encouraging data providers were all included in the pattern guidelines based on the insight provided by CIHI guidelines. The DOD repository (DOD 2014) gave insights into practices relating to standards and regulatory compliance, which helped us to improve the guidelines in Pattern 18. In summary, 17 patterns were amended to improve descriptions relating to implementation guidelines.

7.5 Relationships between patterns

During the evaluation process, we found that definitions of some DQ problems included more than one problem bundled into it. Since the problems seem to have influenced each other, authors have overloaded the definition by bundling all the problems together. For example, in Figure 7.2: Overloaded problems, the definition of the overloaded problem consists of three atomic problems. The main problem that maps to the Pattern X is influenced by two related problems problem-1 and problem-2 that can be mapped to pattern Y and Pattern Z respectively. This situation has made the author bundle all three problems and presents them as a single problem (Problem A). Therefore in this analysis, we observed that there are relationships between DQ problems, and thus in turn between DQ patterns, analysis of which can be insightful in using the DQ patterns in the real world.

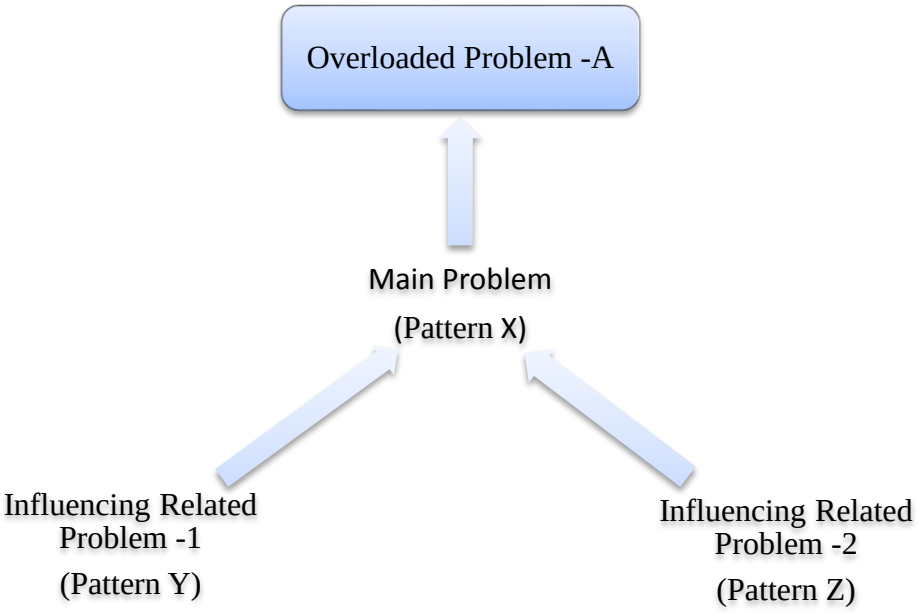


Figure 7.2: Overloaded problems

It should be noted that the central concept behind a DQ pattern is a DQ characteristic. DQ characteristics were defined in Chapter 4 using thematic analysis where each characteristic is a unique theme. The relationships between the themes were not considered in this analysis as it is out of the scope of this thesis. Future research relevant to this is discussed in section 10.5.

However considering the importance of the relationships between DQ patterns in DQ requirements analysis (Chapter 8), we present some of the relationships between the DQ patterns as our observations. For example, in Figure 7.2 since problem-1 and problem-2 seem to have a causality relationship with the main problem we suggest that the Pattern Y and Pattern Z influences Pattern X. In order to avoid the ongoing debate on the causality relationships in qualitative data (Maxwell 2004), we use the term "influence" to express the nature of the relationships observed between patterns.

In the following subsection we present such relationships that we observed during the evaluation.

7.5.1 Accuracy to reality

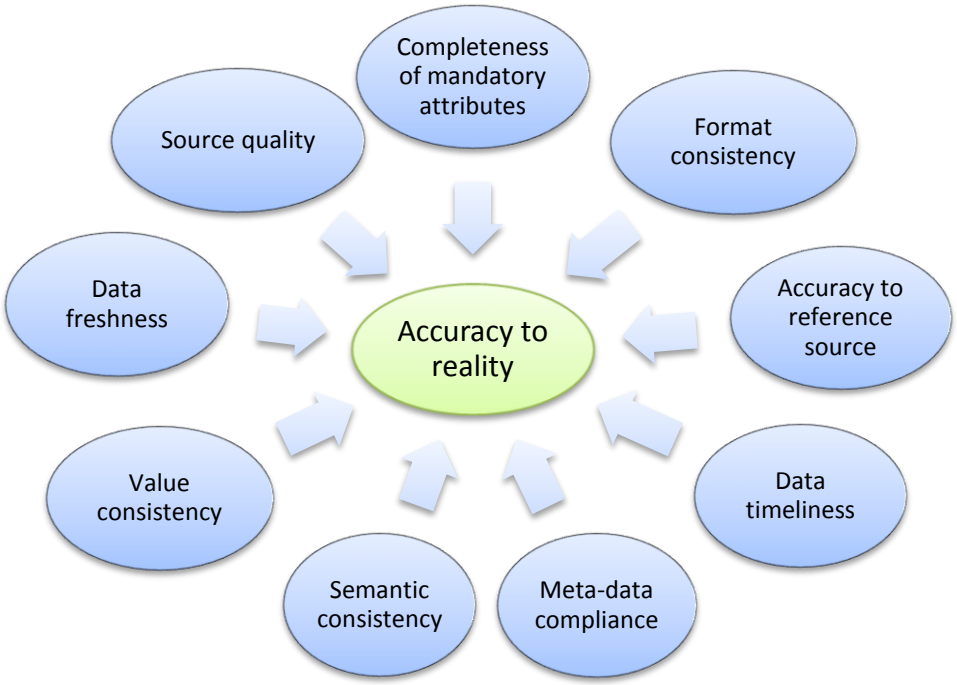


Figure 7.3: Characteristics that influence accuracy to reality

The analysis of following DQ problems (Table 7.2,Table 7.3, Table 7.4) provides evidence for the above relationships.

Overloaded problem	
"Multiple sources of the same information produce different values and lead to confusion, less credibility, and acceptance" *(Strong et al. 1997)*	
Main problem	Relevant pattern
Different values exist for the same real-world entity which leads to confusion (in identifying the entity accurately).	Accuracy to reality (Pattern 14): *Data should truly reflect the real world.*
Influencing related problem	***Relevant pattern***

Multiple sources lead to less credibility (some sources may not be reliable and credible).	Source quality (Pattern 20): *Data used is from trusted and credible sources*.

Table 7.2: Accuracy to reality (Strong et al., 1997)

The problem mainly refers to confusions created for users due to the existence of different values for the same entity derived from different sources. In this situation, users cannot determine which value is correct in the real world.

Overloaded problem	
"Multiple data sources: Multiple sources of the same information produce different values for the information. This can include values that were accurate at a given point in time" *(Lee et al. 2009)*	
Main problem	Relevant pattern
Different values exist for the same information.	Accuracy to reality (Pattern 14): *Data should truly reflect the real world.*
Influencing related problem	Relevant pattern
Multiple sources exist (some sources may not be reliable and credible).	Source quality (Pattern 20): *Data used is from trusted and credible sources.*
Values were accurate at a given point in time (This relates to both data which refers to time and data which is subjected to change over time).	Data timeliness (Pattern 11): *Data which refers to time should be available for use within an acceptable time relative to its time of creation.*
	Data freshness (Pattern 12): *Data which is subjected to changes over the time should be fresh and up-to-date with respect to its intended use.*

Table 7.3: Accuracy to reality (Lee et al., 2009)

Overloaded problem	
"*Garbling (meaningless entries)"* (Eppler 2006)	
Main problem	Relevant pattern
Data is meaningless since there is no corresponding real-world entity that can be identified.	Accuracy to reality (Pattern 14): *Data should truly reflect the real world.*
Influencing related problems	Relevant pattern
Data entry mistakes exist (*meaningless states).*	Accuracy to reference sources (Pattern 13)
Records are not complete with regards to the mandatory attributes (*incomplete representation).*	Completeness of mandatory attributes (Pattern 1): *The attributes which are mandatory for a complete representation of a real world entity must contain values and cannot be null.*
Data do not conform to semantics and hence are meaningless (*ambiguous representation).*	Semantic consistency (Pattern 25): *Data is semantically consistent.*
Data do not conform to formats (*meaningless states, ambiguous representation).*	Format consistency (Pattern 27): Data formats are consistently used.
Data values do not conform to the standard values (*ambiguous representation).*	Value consistency (Pattern 26): Data values are consistent and do not provide conflicting or heterogeneous instances.
Data values do not conform to meta-data (*meaningless states, ambiguous representation).*	Meta-data compliance (Pattern 17): Data should comply with its metadata.
Data values are not up-to-date (*meaningless states, ambiguous representation).*	Data freshness (Pattern 12): Data which is subjected to changes over the time should be fresh and up-to-date with respect to its intended use.

Table 7.4: Accuracy to reality (Eppler, 2006)

Even though the authors have not provided many details about the problem, garbling in the data quality context has been well delineated by Batini and Scannapieco (2006). In general, garbling occurs when there is no corresponding real-world entity. According to these authors, garbling can be

due to *incomplete representation*, *ambiguous representation*, and *meaningless states* when compared with the real world. Hence the probable root causes for garbling can be analysed as in Table 7.4 above.

7.5.2 Interpretability

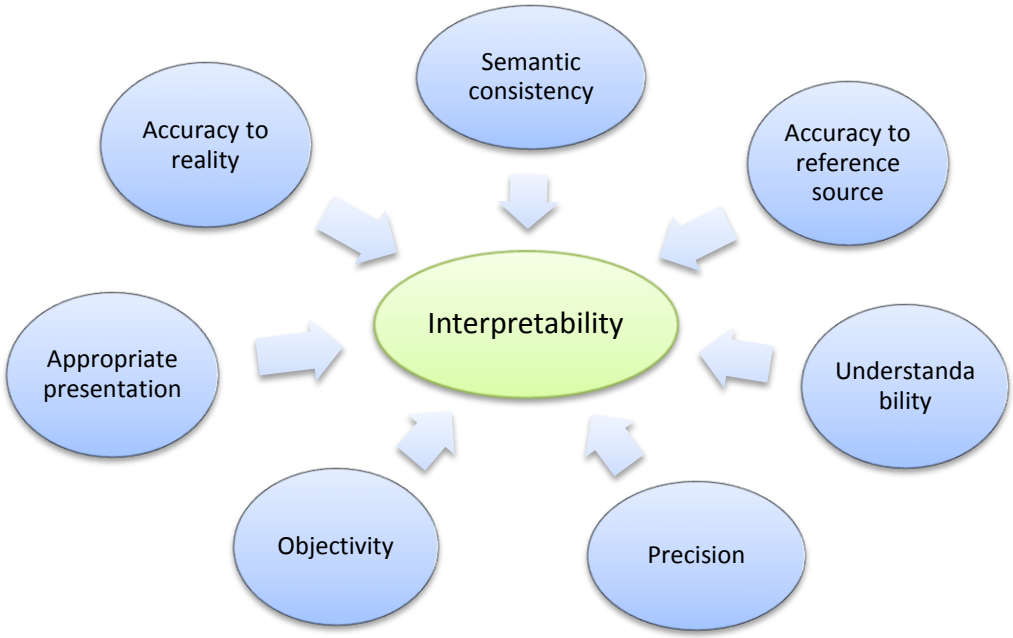


Figure 7.4: Characteristics that influence interpretability

The analysis of following DQ problems (Table 7.5,Table 7.6,Table 7.7) provides evidence for the above relationships

Overloaded problem	
"Information leading to differing or wrong interpretations due to lacking precision or accuracy, the use of abbreviations or jargon, or simply different points of view" *(Lesca and Lesca 1995)*	
Main problem	Relevant pattern
Differing or wrong interpretation of data exist.	Interpretability (Pattern 32)
Influencing related problems	Relevant pattern

Data lacks precision or accuracy - Not accurate as per the reference source - Not accurate as per the reality - Not accurate as per linguistics and granularity.	Accuracy to reference source (Pattern 13)
	Accuracy to reality (Pattern 14)
	Precision (Pattern 15)
Use of abbreviations or jargon (leading to the ambiguity of data).	Understandability (Pattern 30)
	Semantic consistency (Pattern 25)
Different points of view exist.	Objectivity (Pattern 21)

Table 7.5: Interpretability (Lesca and Lesca, 1995)

Overloaded problem	
"Manipulation of decision processes (overloading, confusing, diverting)" (Eppler 2006)	
Main problem	Relevant pattern
There is manipulation of decision processes.	Interpretability (Pattern 32)
Influencing related problem	Relevant pattern
Overloading, confusing and/or diverting exist.	Understandability (Pattern 30)

Table 7.6: Interpretability (Eppler, 2006)

Overloaded problem
*"'**Massaged' information (inappropriate format): massaging is the putting together of data in a manner that applies to a particular problem at hand. The problem that lies in this massaging is that the very same information - when massaged (or represented differently) – may lead to different (and sometimes inadequate) interpretations**" (Garvin 1988)*

Main problem	Relevant pattern
Problems of interpretations exist (different and sometimes inadequate interpretations).	Interpretability (Pattern 32)
Influencing related problem	Relevant pattern
There is an inappropriate format.	Appropriate presentation (Pattern 31)

Table 7.7: Interpretability (Garvin, 1988)

7.5.3 Objectivity

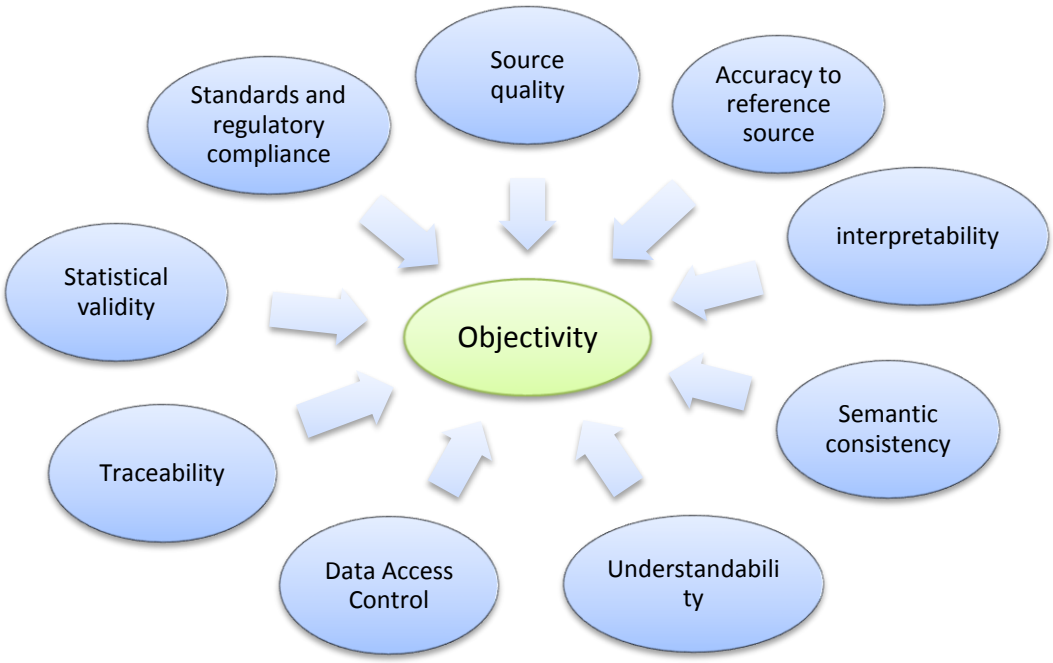


Figure 7.5: Characteristics that influence objectivity

The analysis of following DQ problems (Table 7.8,Table 7.9,Table 7.10,Table 7.11,Table 7.12) provides evidence for the above relationships.

Overloaded problem	
"Biased information: that is to say information that is inaccurate or distorted due to the interests or motives of the source or information transmitter" *(Garvin 1988)*	
Main problem	Relevant pattern
Biased information exists.	Objectivity (Pattern 21)
Influencing related problems	Relevant pattern
There are distortions due to interests or motives of the source.	Source quality (Pattern 20)
There are distortions due to interests or motives of the transmitter (the entity providing information to the system).	Accuracy to reference source (Pattern 13)

Table 7.8: Objectivity (Garvin, 1988)

Overloaded problem	
"Information is produced using subjective judgements, leading to bias. The objectivity of the information decreases and information is difficult to evaluate" *(Strong et al. 1997)* ***"Subjective judgement in data production: Information production in using subjective judgement can result in the production of biased information"*** *(Lee et al. 2009)*	
Main problem	Relevant pattern
Biased information exists.	Objectivity (Pattern21)
Influencing related problem	Relevant pattern
Subjective judgement has been used in producing information	Interpretability (Pattern 32) Accuracy to reference sources (Pattern 13)

- Producing new data/information in the system by interpreting existing data/information - Producing data/information by capturing new data/information.	

Table 7.9: Objectivity (Lee et al., 2009)

Overloaded problem	
"Distortion of information, e.g., when the original message is no longer the same when it is received. Causes for this problem are too many intermediaries, too much specialization, and jargon, or even voluntary distortion (e.g. misinformation) such as modifying, delaying or blocking the information to harm the receiver"*(Lesca and Lesca 1995)*	
Main problem	Relevant pattern
There is a distortion of information.	Objectivity (Pattern 21)
Influencing related problems	Relevant pattern
There are too many intermediaries.	Source quality (Pattern 20)
	Accuracy to reference source (Pattern 13)
There is too much jargon - Not consistently used with standardised meanings - It cannot be understood.	Semantic consistency (Pattern 25)
	Understandability (Pattern 30)
The information is purposefully distorted.	Data access control (Pattern 10)

Table 7.10: Objectivity (Lesca and Lesca, 1995)

Overloaded problem	
"*The information is not reliable or trustworthy, i.e., there is a great risk of errors, and the information's background cannot be checked. Causes for this problem are mistakes in the information production and distribution process, as well as unidentified sources"* *(Eppler 2006)*	
Main problem	Relevant pattern
Information is not reliable or trustworthy.	Objectivity (Pattern 21)
Influencing related problems	Relevant pattern
Background or provenance cannot be checked.	Traceability (Pattern 22)
The source is undefined.	Source Quality (Pattern 20)
There are data capturing errors and mistakes in data distribution.	Accuracy to reference sources (Pattern 13)

Table 7.11: Objectivity (Eppler, 2006)

Overloaded problem	
"The level of bias is not significant" *(Richards and White 2013)*	
Main problem	Relevant pattern
Checking the level of bias of data	Objectivity (Pattern 21)
Influencing related problems	Relevant pattern
Statistical significance	Statistical validity (Pattern 19)
Compliance to data collection procedures and protocols	Accuracy to reference sources (Pattern 13)
Compliance to data processing standards	Standards and regulatory compliance (Pattern18)

Table 7.12: Objectivity (Richards and White, 2013)

7.5.4 Ease of data access

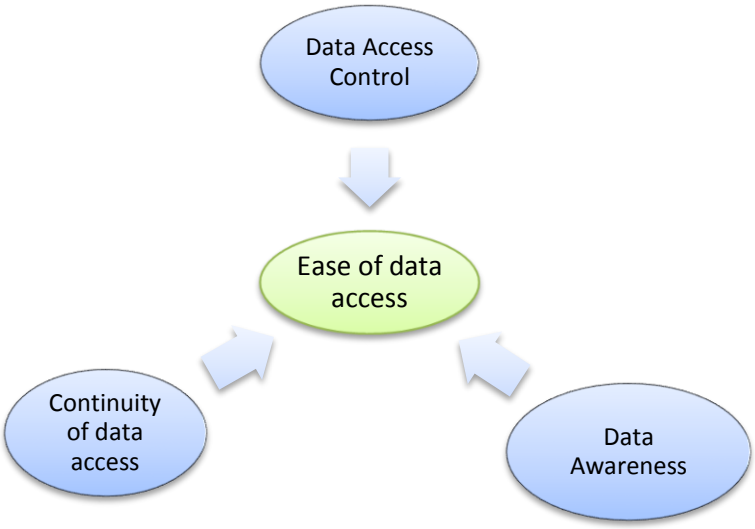


Figure 7.6: Characteristics that influence ease of data access

The analysis of following DQ problems provides evidence (Table 7.13,Table 7.14) for the above relationships.

Overloaded problem	
*"**Security and privacy requirements: Easy access to information may conflict with requirements for security, privacy, and confidentiality**." (Lesca and Lesca 1995; Strong et al. 1997)*	
Main problem	Relevant pattern
Easy access to information is prohibited.	Ease of data access (Pattern 8)
Influencing related problem	Relevant pattern
Security, privacy and confidentiality requirements	Data access control (Pattern 10)

Table 7.13: Ease of data access (Strong et al., 1997, Lesca and Lesca, 1995)

Overloaded problem	
*"**Difficult information navigation and retrieval"** (Eppler 2006)*	
Main problem	Relevant pattern
It is difficult to locate and obtain information.	Ease of data access (Pattern 8)
Influencing Related Problem	Relevant pattern
Difficulty in navigating information	Data awareness (Pattern 7)
Difficulty in information retrieval	Continuity of data access (Pattern 5)

Table 7.14: Ease of data access (Eppler, 2006)

7.5.5 Uniqueness

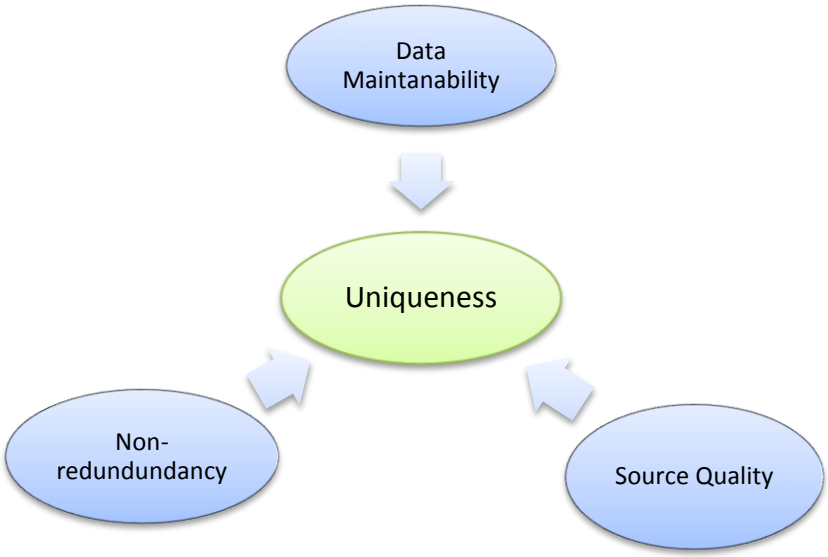


Figure 7.7: Characteristics that influence uniqueness

The analysis of following DQ problems provides evidence (Table 7.15,Table 7.16) for the above relationships.

Overloaded problem	
*"**Duplicates due to multiple data sources**" (Eppler 2006)*	
Main problem	Relevant pattern
Duplicates or multiples exist.	Uniqueness (Pattern 23)
Influencing Related Problem	Relevant pattern
Multiple sources produce duplicates.	Non-redundancy (Pattern 24)
No single version of the truth is identified from multiple sources.	Data maintainability (Pattern 6)
It is hard to know which source(s) provides accurate data to create a single version of the truth.	Source quality (Pattern 20)

Table 7.15: Uniqueness (Eppler, 2006)

Overloaded problem	
*"**Approximate duplicate tuples: The tuple Customer (10, 'Smith Barney', 'Flowers Street, 123', 502899106) in DS1 is an approximate duplicate of the tuple Customer (27, 'Smith B.', 'Flowers St., 123', 502899106) in DS2**" (Oliveira et al. 2005)* *"**Inconsistent duplicate tuples: The tuple Customer (10, 'Smith Barney', 'Flowers Street, 123', 502899106) in DS1 is an inconsistent duplicate of the tuple Customer(27, 'Smith Barney', 'Sun Street, 321', 502899106) in DS2**" (Oliveira et al. 2005)*	
Main problem	Relevant pattern
Duplicates or multiples exist.	Uniqueness (Pattern 23)
Influencing Related Problem	Relevant pattern

Sourcing from both DS1 and DS2 produces duplicates.	Non-redundancy (Pattern 24)
No single version of the truth exists due to multiple sources.	Data maintainability (Pattern 6)
It is hard to know which source(s) provides accurate data to create a single version of the truth.	Source quality (Pattern 20)

Table 7.16: Uniqueness (Oliveira et al., 2005)

7.5.6 Usefulness and relevance

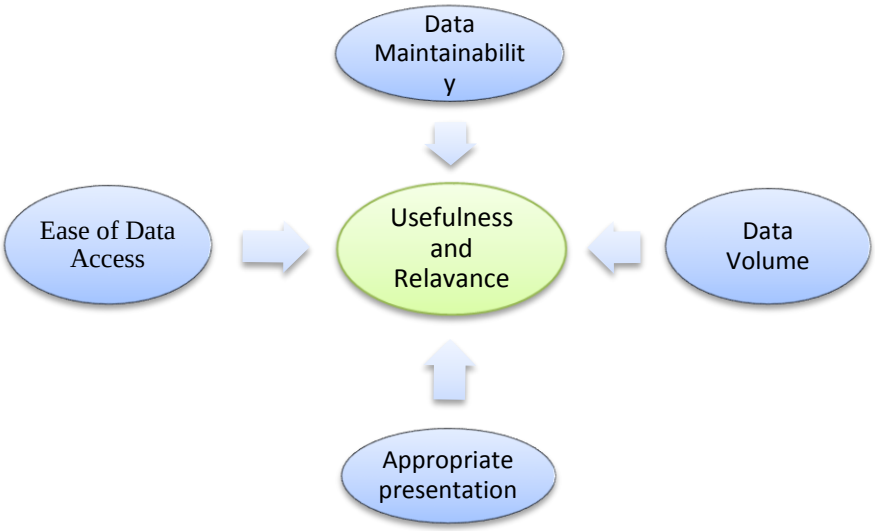


Figure 7.8: Characteristics that influence usefulness and relevance

The analysis of following DQ problems provides evidence for the above relationships.

Overloaded problem	
*"**Limited usefulness of information due to an overload of information caused by a lack of cleansing or maintenance activities**" (Lesca and Lesca 1995)*	
Main problem	Relevant pattern

The usefulness of information is limited.	Usefulness and relevance (Pattern 29)
Influencing Related Problem	Relevant pattern
There is too much information.	Data volume (Pattern 4)
There is a lack of regular cleansing or periodic maintenance.	Data maintainability (Pattern 6)

Table 7.17: Usefulness and relevance (Lesca and Lesca, 1995)

Overloaded problem	
*"**An inadequate presentation format that leads to expensive conversion tasks. The main problem lies in the fact that information is not presented in an order, format or style that allows for a direct use, hence conversion is necessary. Causes for this problem are insufficient dialogue between information producers and consumers, constant time pressure, and a lack of adaptation of information to usage needs or styles**" (Lesca and Lesca 1995)*	
Main problem	Relevant pattern
There is a lack of adaptation of information to usage needs (insufficient dialogue between information producers and consumers).	Usefulness and relevance (Pattern 29)
Influencing Related Problem	Relevant pattern
Information is not presented in an order, format or style that allows for a direct use and hence expensive conversion tasks are needed.	Appropriate presentation (Pattern 31)
Direct use of information is not possible.	Ease of data access (Pattern 8)

Table 7.18: Usefulness and relevance (Lesca and Lesca, 1995)

7.5.7 Information value

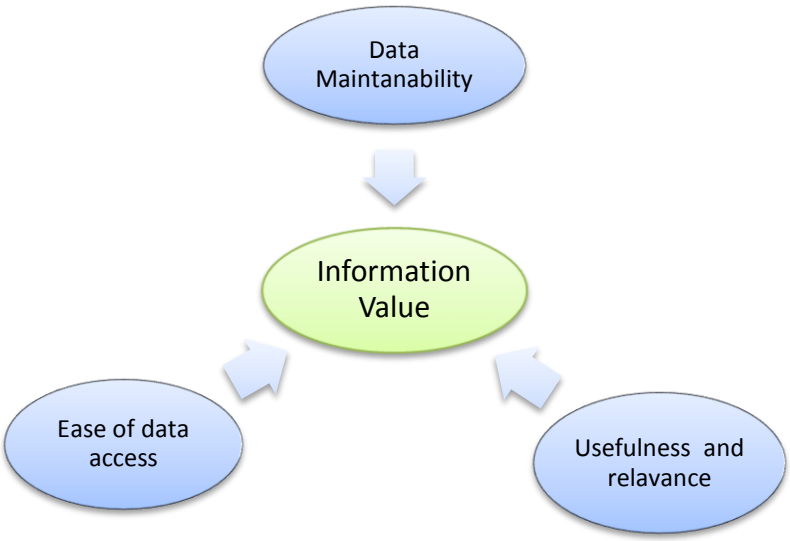


Figure 7.9: Characteristics that influence information value

The analysis of following DQ problems provides evidence (Table 7.19, Table 7.20) for the above relationships.

Overloaded problem	
*"**Incompleteness of information that can lead to inadequate decisions. The main causes for this problem are the fragmentation of work and the resulting specialization that leads to fragmentation of information**"* (Lesca and Lesca 1995)	
Main problem	Relevant pattern
Inadequate decisions are being made.	Information value (Pattern 33)
Influencing Related Problem	Relevant pattern
Information is incomplete.	Usefulness and relevance (Pattern 29)
Specialization leads to fragmentation of information.	Understandability (Pattern 30)

Table 7.19: Usefulness and relevance (Lesca and Lesca, 1995)

Overloaded problem	
*"**The information is not accessible. It is lost over time because of unclear responsibilities or technological changes. This can demotivate staff and lead them to wrong decisions**"* (Lesca and Lesca 1995)	
Main problem	Relevant pattern
The wrong decisions are being made.	Information value (Pattern 33)
Influencing Related Problem	Relevant pattern
There is a lack of access to information.	Ease of data access (Pattern 8)
Information is lost over time because of unclear responsibilities.	Data maintainability (Pattern 6)

Table 7.20: Usefulness and relevance (Lesca and Lesca, 1995)

7.5.8 Understandability

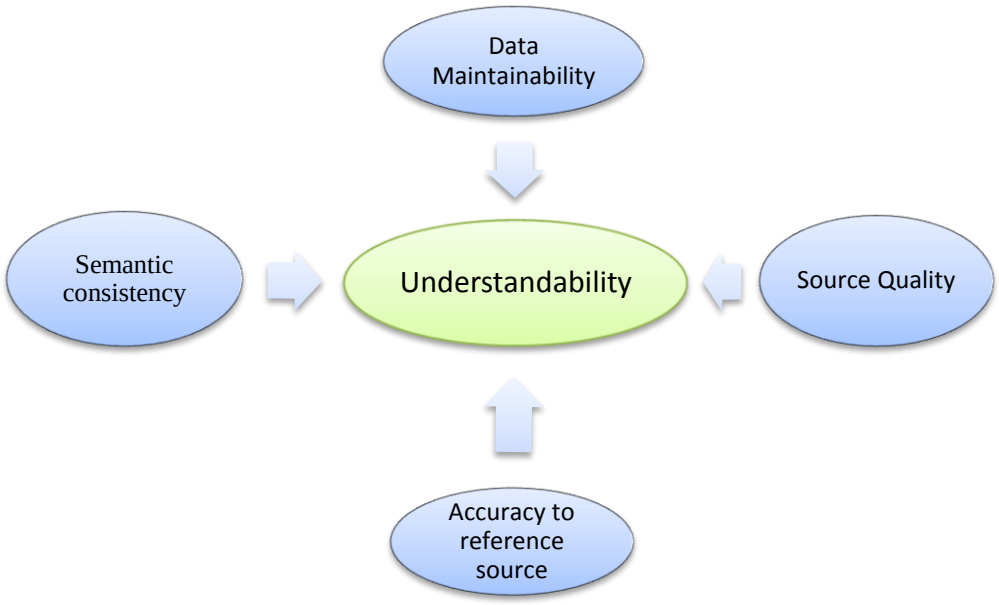


Figure 7.10: Characteristics that influence understandability

The analysis of following DQ problems provides evidence (Table 7.21,Table 7.22, Table 7.23) for the above relationships.

Overloaded problem	
*"**Coded data across disciplines: Coded data from different functions and disciplines is difficult to decipher and understood. Also, codes may conflict**"(Lee et al. 2009)*	
Main problem	Relevant pattern
Data which is coded is difficult to decipher and understand.	Understandability (Pattern 30)
Influencing Related Problem	Relevant pattern
Codes from different functions may conflict.	Semantic consistency (Pattern 25)

Table 7.21: Understandability (Lee et al., 2009)

Overloaded problem	
*"**Abbreviations are unclear (ste for suite, hwy for highway)"** (Kim et al., 2003)*	
Main problem	Relevant pattern
Abbreviations lead to confusion.	Understandability (Pattern 30)
Probable root causes	Relevant pattern
Abbreviations have not been standardized.	Semantic consistency (Pattern 25)

Table 7.22: Understandability (Kim et al, 2003)

Overloaded problem	
*"**Inconsistency of information leads to confusion. Causes for inconsistencies or contradictory statements are a lack of co-ordination between information authors and distributors, unclear responsibilities, or the use of multiple, inconsistent, information sources**" (Lesca and Lesca 1995)*	
Main problem	Relevant pattern
Information is inconsistent or contradictory.	Understandability (Pattern 30)
Influencing Related Problem	Relevant pattern
There is a lack of co-ordination between information authors and distributors.	Accuracy to reference source (Pattern 13)
Responsibilities are unclear.	Data maintainability (Pattern 6)
Multiple and inconsistent information sources are used.	Source quality (Pattern 20)

Table 7.23: Understandability (Lesca and Lesca, 1995)

7.5.9 Data freshness

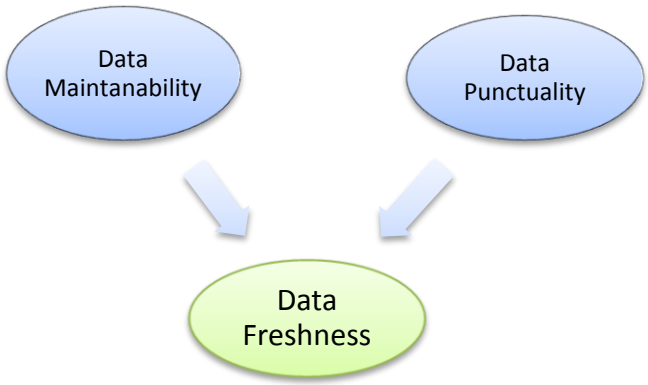


Figure 7.11: Characteristics that influence data freshness

The analysis of following DQ problem provides evidence (Table 7.24) for the above relationships.

Overloaded problem	
*"**Outdated information that is no longer current due to its tardy delivery or a failure to update it (is being used)."** (Garvin 1988)*	
Main problem	Relevant pattern
Outdated information is used.	Data freshness (Pattern12)
Influencing Related Problem	Relevant pattern
The delivery of information is late.	Data punctuality (Pattern 9)
The information is not updated regularly.	Data maintainability (Pattern 6)

Table 7.24: Understandability (Garvin, 1988)

7.5.10 Punctuality

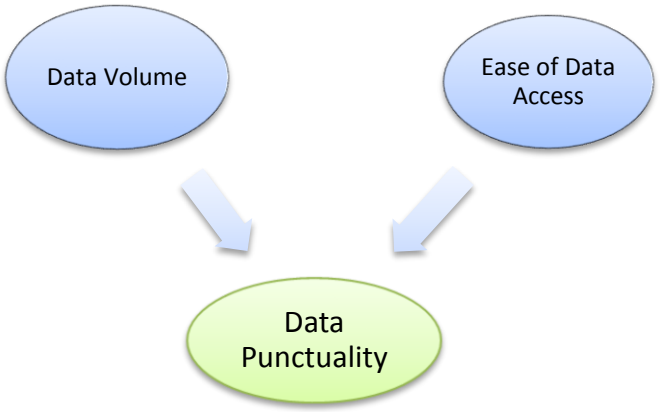


Figure 7.12: Characteristics that influence data punctuality

The analysis of following DQ problems provides evidence (Table 7.25) for the above relationships.

Overloaded problem	
*"**The large volume of stored information makes it difficult to access needed information in a reasonable time**"* (Lee et al. 2009)	
Main problem	Relevant pattern
It is hard to access information in a reasonable time.	Data punctuality (Pattern 9)
Influencing Related Problem	Relevant pattern
The information is stored in the large volume.	Data volume (Pattern 4)
There is difficulty in accessing it (speed or gating issues).	Ease of data access (Pattern 8)

Table 7.25: Punctuality (Lee et al., 2009)

7.5.11 Accuracy to reference sources

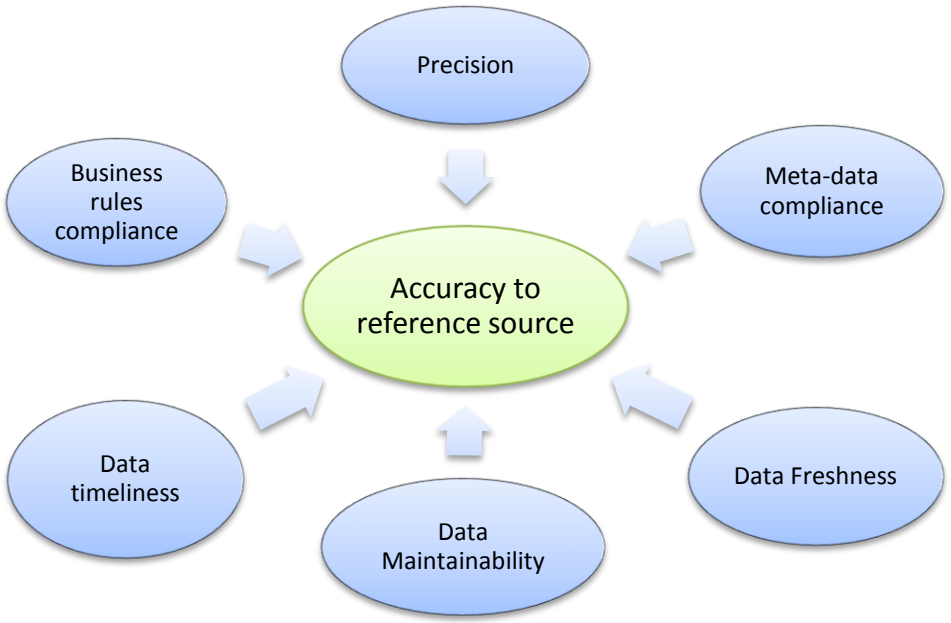


Figure 7.13: Characteristics that influence accuracy to reality

The analysis of following DQ problems (Table 7.26,Table 7.27,Table 7.28,Table 7.29) provides evidence for the above relationships.

Overloaded problem	
"There are spelling errors" *(Eppler 2006)* *"**Misspelling exists (e.g., principle instead of principal, effect instead of affect**)"(Kim et al. 2003)*	
Main problem	Relevant pattern
There are typos and word mismatches (spelling errors).	Accuracy to reference sources (Pattern 13)
Influencing Related Problem	Relevant pattern
Validations for grammar and vocabulary are not being conducted.	Precision (Pattern 15)

Table 7.26: Accuracy to reference sources (Kim et al., 2003)

Overloaded problem	
*"**Entry into the wrong fields**"(Ge and Helfert 2013)* *"**Incorrect data entries because of a lack of validation**" (Eppler 2006)*	
Main problem	Relevant pattern
There are data capturing problems (entry into the wrong fields)	Accuracy to reference sources (Pattern 13)
Influencing Related Problem	Relevant pattern
There are no validation rules at database level	Meta-data compliance (Pattern 17)

There are no validation rules at application program level	Business rules compliance (Pattern 16)

Table 7.27: Accuracy to reference sources (Ge and Helfert, 2013; Eppler, 2006)

Overloaded problem	
*"**Misplaced data is saved in the wrong database**"(Eppler 2006)*	
Main problem	Relevant pattern
There are data capturing problems (entry into the wrong database).	Accuracy to reference sources (Pattern 13)
Influencing Related Problem	Relevant pattern
No responsibility is taken for the data.	Data maintainability (Pattern 17)

Table 7.28: Accuracy to reference sources (Eppler, 2006)

Overloaded problem	
*"**Incorrect values: Let u(t,a) be the correct and updated value that the attribute of a tuple t was supposed to have. There is an incorrect value in attribute a $\in R(A)$ if: $\exists t \in r : v(t,a) \in Dom(a) \wedge v(t,a) \neq u(t,a)$"** (Oliveira et al. 2005)*	
Main problem	Relevant pattern
Incorrect values are ascribed.	Accuracy to reference sources (Pattern 13)
Influencing Related Problem	Relevant pattern
Values are not being updated.	Data maintainability (Pattern 6)

	Data freshness (Pattern12)
	Data timeliness (Pattern11)

Table 7.29: Accuracy to reference sources (Oliveira et al., 2005)

7.5.12 Continuity of data access

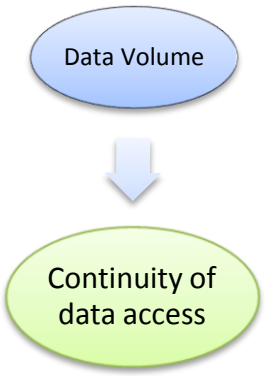


Figure 7.14: Characteristics that influence Continuity of data access

The analysis of following DQ problem (Table 7.30) provides evidence for the above relationships.

Overloaded problem	
*"**Too much information: Large volumes of stored information make it difficult to access information in a reasonable time"**(Strong et al. 1997)*	
Main problem	Relevant pattern
difficult to access information in a reasonable time (Speed)	Continuity of data access (Pattern3)
Influencing Related Problem	Relevant pattern
Too much information	Data volume (Pattern6)

Table 7.30: Continuity of data access (Strong et al., 1997)

7.5.13 Appropriate presentation

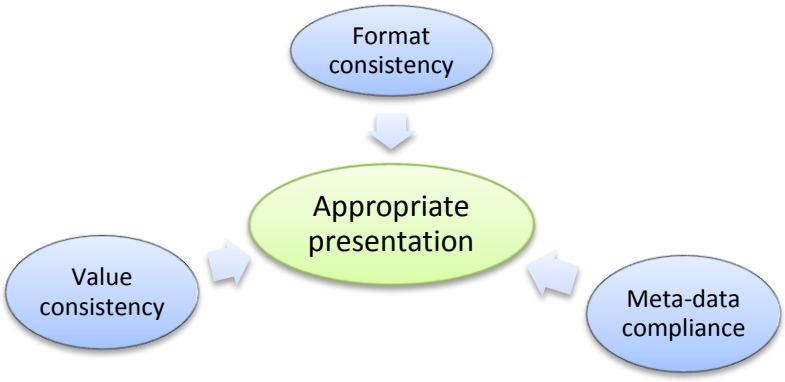


Figure 7.15: Characteristics that influence appropriate presentation

The analysis of following DQ problems (Table 7.31) provides evidence for the above relationships.

Overloaded problem	
*"**Distributed systems: Distributed, heterogeneous systems lead to inconsistent definitions, formats, and values. Information can no longer be easily aggregated or combined, due to the format differences and incompatibilities**" (Strong et al. 1997)*	
Main problem	Relevant pattern
Information cannot be aggregated into a required presentation format for use.	Appropriate presentation (Pattern 31)
Influencing Related Problem	Relevant pattern
Heterogeneous systems lead to inconsistent formats	Format consistency (Pattern27)
Heterogeneous systems leading to inconsistent values	Value consistency (Pattern26)
Inconsistent definitions	Meta-data compliance (Pattern17)

Table 7.31: Appropriate presentation (Strong et al., 1997)

7.6 Summary and conclusion

In this validation, we used 213 data quality problems from eight credible sources and 197 data quality rules from three real-world rule repositories. The validation revealed that there is a corresponding pattern for each of the problems and rules considered in this validation. In fact, some problems were overloaded and we split such problems into atomic problems in performing this validation. Based on the results of the validation, it was evident that the repository of the patterns is a complete representation of the data quality requirements in information systems.

Further, in this validation, we identified 13 relationships between patterns which we consider as an important finding. In the literature on patterns, authors argue that relationships between patterns enable the pattern to use them successfully to design complex solutions (Noble 1998; Rolland et al. 1998). Hence we believe that the above identified relationships can be useful in modelling data quality requirements. The summary of the relationships isss shown in Table 7.32 below.

	P1: Comp. of mandatory att.	P2: Comp. of optional att.	P3: Comp. of records	P4: Data volume	P5: Continuity of data access	P6: Data maintainability	P7: Data awareness	P8: Ease of data access	P9: Data punctuality	P10: Data access control	P11: Data timeliness	P12: Data freshness	P13: Acc. to reference sources	P14: Acc. to reality	P15: Precision	P16: Business rules compliance	P17: Meta-data compliance	P18: Stan. & reg. compliance	P19: Statistical validity	P20: Source quality	P21: Objectivity	P22: Traceability	P23: Uniqueness	P24: Non-Redundancy	P25: Semantic consistency	P26: Value consistency	P27: Format consistency	P28: Referential integrity	P29: Usefulness & Relevance	P30: Understandability	P31: Appropriate presentation	P32: Interpretability	P33: Information value
P1														+																			
P2																																	
P3																																	
P4					+				+																				+				
P5								+																									
P6												+	+										+						+	+			+
P7								+																									
P8									+																				+				+
P9												+																					
P10								+													+												
P11													+																				
P12													+	+																			
P13														+							+									+		+	
P14																																+	
P15													+																			+	
P16													+																				
P17													+	+																	+		

P18																				+											
P19																				+											
P20														+						+		+							+		
P21																														+	
P22																				+											
P23																															
P24																						+									
P25														+						+									+		+	
P26														+																+	
P27														+																+	
P28																															
P29																															+
P30																				+										+	
P31																												+		+	
P32																				+											
P33																															

Table 7.32 Summary of the relationships between DQ Pattern

In the next chapter, we discuss a methodology for using the patterns and the relationships to model data quality requirements.

Chapter-8

GOAL ORIENTED DATA QUALITY REQUIREMENTS MODELING

8.1 Overview

In this chapter, we present a methodology for using the 33 data quality patterns to model data quality requirements in an organization.

In chapter-6 we presented a set of generic data quality patterns we developed to facilitate DQ requirements modelling. Each pattern provides the required knowledge on a particular generic data quality requirement. The question remains as to how these generic DQ requirements patterns can be used to model specific data quality requirements in an organization which constitutes the aim of this chapter.

8.2 DQ requirements analysis

Elicitation of data quality requirements is a crucial aspect of data quality management. In literature not many attempts can be found on DQ requirements elicitation as pointed out in the literature review in chapter-2. Loshin (2001) has suggested an approach to defining data quality requirements using the use-case method (Jacobson 1992), where the business analysts can derive DQ requirements from the use cases related to the following main aspects of an information system,

1) *Model generation/Relational inputs*
2) *Invariants, boundary conditions, constraints*
3) *Quality issues*
4) *Report generation and queries*
5) *Performance*

Use cases provide an understanding the actors in a system, the information chain that produces data, and the impact of low-quality data. Then these impacts are translated into data quality terms (dimensions/metrics) and thereby elicit data quality requirements for particular data elements.

Redman (1997) describes defining data quality requirements using a step by step process which he termed as the quality functional deployment (QFD). The idea of this step by step process is to identify and translate the voice of the customer (data consumer) into a technical specification of data quality requirements which can be further incorporated into the process that creates data. The steps are,

1) *Understand what customers want*
2) *Develop a single set of consistent user requirements*

3) *Translate user requirements into data quality requirements (requirements in technical language)*
4) *Map data quality requirements into individual performance requirements (individual level metrics to measure data quality*
5) *Establish performance specifications for processes (Process-based metrics to measure data quality)*

Redman suggests focus groups and interviews to gather customer requirements and he uses requirement matrix as a tool to sequentially refine requirements in the above steps.

A similar approach has been suggested by Sebastian-Coleman (2012) where she considers data quality requirements as specific kind of business requirements. Hence a systematic review of business requirements with SMEs (subject matter experts) is used to elicit DQ requirements. In this process, business requirements, data model, data standards, data profiling results and data quality characteristics are used as the basis for recognising data quality requirements and measurement criteria for each requirement.

It should be noted that the first two methodologies have focused mostly on data consumers' needs in identifying DQ requirements, whereas, in the third methodology, the author has focussed on the business requirements in identifying DQ requirements. Therefore this methodology has taken a step forward in combining business requirements with data quality requirements so that DQ becomes an integral part of business requirements. As mentioned in the literature review, current approach in DQ management focuses on the enterprise view of data where information is treated as an organizational asset to achieve organizational goals (Pierce et al. 2013) . Therefore we posit that DQ requirements should steam from organizational goals and stay focused towards bringing some value to the organization and data users.

Therefore in light of the above notion, in the following section, we propose a methodology to use the DQ patterns to facilitate systematic identification of DQ requirements with regards to organizational goals.

8.3 Goal oriented DQ requirements engineering

In requirements engineering for software development, Goal-oriented approaches have been proposed by many researchers to overcome the weaknesses in traditional requirements engineering approaches (Albers 1998; Endsley et al. 2003; Robinson and Elofson 2004). Goals capture the intentionality behind software requirements and therefore researchers argue that goals are a useful

abstraction to represent stakeholders' needs and expectations providing an intuitive way to elicit and analyse requirements (Lapouchnian 2005; Yu and Mylopoulos 1998). In literature goal oriented approaches such as i* framework (Yu 2011), Tropos (Bresciani et al. 2004), GDTA (Endsley et al. 2003) GDIA (Prasanna et al. 2009; Yang et al.) and KAOS (Van Lamsweerde 2001) have been proposed for requirements analysis considering goals as the rationale for requirements.

i* Framework (distributed intentionality) is focussed on reasoning the organizational environment through two interrelated models SD (strategic dependency) and SR (strategic rationale). The SD model is focused on the actors in the environment and the dependency between actors is considered as the intentionality is achieving something, which is in turn characterised as a goal. The SR model is focussed on modelling the goals, tasks, actors and the resources of actors.

TROPOS (Bresciani et al. 2004) adopts the fundamental concepts in i* framework. The methodology starts with early requirements analysis in which the stakeholders and their intentions are identified as social actors and goals respectively. Then a goal analysis is performed where each actor has three alternatives, either to accept and fulfil it as a responsibility, or to refine into sub-goals (using AND/OR decomposition), or delegate the goal to an existing actor or new actors. The process end when all goals have been assigned to actors. Therefore it does not just focus on users' data needs, but on how data can be used within decision making to achieve goals.

GDTA (Goal-Directed Task Analysis) by Endsley et al. (2003) is used to elicit information needs for complex decision making in environments where high cognitive skills are used in decision making. It uses structured interviews and observations of users performing their tasks and detailed analysis of the documentation on users' tasks to identify the goals of the users. Then the methodology suggests identifying the decisions taken in achieving each goal. For each decision, the information requirements are identified.

GDIA (Goal-directed information analysis) by Prasanna et al. (2009), is focussed on finding information requirements in emergency responding environments. Due to the vast diversity exist among the tasks performed in emergency responding situations the methodology starts with reasoning every task and identify the goals achieved by the tasks. Then the information requirements are assessed based on the goals. The methodology contains sequentially performed steps starting from context analysis, scenario development (to identify tasks), physical task identification, defining of goals and sub-goals of the tasks, goal validation, and information requirement identification.

KAOS (keep all objectives satisfied) is another goal oriented approach (Lamsweerde 2009; Van Lamsweerde 2001) where the system requirements are generated by evaluating different models that

support one another. In KAOS the following four models are developed incrementally during requirements elicitation process,

(1) The goal model

(2) The object model

(3) The agent responsibility model and

(4) The operation model.

As per KAOS, a goal is a non-operational requirement to be achieved by the system. In other words, it is a high-level abstraction of a system requirement. The goals are defined using patterns (achieve, cease, maintain, avoid and optimise) which are reusable abstractions of goals. Therefore these patterns initially provide grounds to recognize the goals in the organizational environment. The goals are linked to objects, where the objects are the "things" of interest in the organizational environment whose instances are handled through the system with state transitions. Agents are either human beings or automated components that are responsible for achieving the goals by performing the operations. Operations are the activities defined to fulfil goals through interacting with objects. Therefore operations are the lowest level abstraction of requirements that are used in specifying an information system.

KAOS suggests that the four models are strongly related to each other and therefore the focus on each model facilitates a complete analysis of organizational requirements. Though the process of requirement elicitation starts with the identification of goals, KAOS do not specify that the models should be sequentially developed, but they have to be developed simultaneously such that each model helps to improve the others. For example, the objects in the object model may intern emphasize on new goals which were not identified in the initial goal model and therefore the goal model is further improved by the object model. Further, the agents in the agent model may provide an insight into new objects interacting with the agents and hence new objects and new goals may result in respective models. Therefore this incremental analysis through the related models leads to a realization of rich requirements models ensuring that every operational requirement identified has some relevance to an organizational goal and they are not just user expectations and needs.

KAOS is independent of any specific requirement gathering techniques (interviews, focus groups, surveys etc.) but it emphasizes on three important aspects of requirement gathering.

1) Use of requirement patterns to make sure the requirements are gathered effectively and efficiently.

2) Provide validations and verifications for each requirement.
3) Facilitate both top-down and bottom-up analysis in requirements elicitation resulting a complete set of requirements.

We observe that, in the context of data quality, the above four models can be recognised and evaluated accordingly. Data quality goal is a non-operational requirement related to organizational data. For example such as achieving an accuracy of customer data to strengthen the customer relationship. The data objects related to the goal are therefore customer data. The agents can be identified as the responsible staff members for customer data. The activities that are necessary to improve the quality of contact data can be identified as the operations. Therefore the same four models can be developed for DQ context.

Traditionally the requirements are gathered by means of open interviews, but KAOS supports the notion that a more efficient way to gather requirements is to conduct less open interviews by reusing requirements patterns (Objectiver 2007). KAOS supports in progressively modelling generic patterns of requirements which can be used on new cases to guide the identification of requirements (Lamsweerde 2009) .

Since every DQ pattern represents a generic DQ requirement, we observe that a DQ pattern provides the necessary basics in developing each model in KAOS methodology. In other words, the constructs of a DQ pattern represents the concepts in each model and hence a pattern is a single representation of the four models. Thus by following the KAOS methodology we can instantiate each relevant pattern and develop a DQ requirements model for the organization.

In Table 8.1 we have presented the summary of adaptation of KAOS for DQ context.

KAOS Model	Adaptation for DQ	Corresponding construct of a DQ pattern
The goal model	DQ goals which are non-operational requirements identified referring to organizational goals.	DQ dimension which is a high-level abstraction of a DQ requirement (eg: Accuracy of customer contact data) DQ characteristic which is a low-level abstraction of a DQ requirement (eg: Accuracy to reality of customer address)

Object model	Data objects associated with a goal.	Data granularity (data elements, records or information objects)
Agent responsibility model	Data Stewardship (Act of responsibility of managing the quality of the data object(s)	Verification metric and threshold
Operational model	Implementaion requirements to meet the DQ goal	Implementation form - Process-based approach - Rule based approach - Validation metric and threshold

Table 8.1:Adaptation of KAOS for DQ context

We observe several analogies within KAOS model and DQ pattern constructs:

The DQ characteristics can be considered as high level DQ goals since they specify what needs to be achieved to maintain quality of data.

Data granularity specifies the object at which the DQ characteristic is applicable and hence provides a link between the DQ goal and the data object associated with the goal.

Agent responsibility is identical to the data stewardship where the responsibility of managing DQ is established.

Verification metric can be considered as a mechanism that ensures the DQ goal has been achieved.

Implementation form can be considered as the operational model that ensures the ultimate achievement of the goal.

Based on the above observations we identified KAOS as a suitable methodology for data quality requirements modelling using DQ patterns. In the following sections, we explain how we can use KAOS systematically in modelling DQ requirements.

8.4 KAOS for DQ requirements modelling (KAOS4DQ).

The fundamental of KAOS is the systematic elaborations of high-level abstraction of requirements (Goals) towards a low-level abstraction of requirements (operations) through the development of four models. Therefore it ensures that all requirements will be identified without missing any, and also

non-value adding requirements will not be identified. Focussing on the four respective model, we propose a two-phase approach to using KAOS in DQ requirements modelling as follows.

8.4.1 Phase-1: Top down analysis (Analysis of DQ context)

In this phase, the main aim is to recognize the DQ context of the organization. For this, a survey is performed using the high-level executives who are responsible for the DQ management function (DQ manager, DQ team leader, line managers etc.) of the organization.

1) The participants are asked to focus on a particular data set in the organization.
2) They are provided with a survey instrument to facilitate their thought process to identify critical DQ problems in the organization.
3) The participants provide a rating for each selected DQ requirement (characteristic) considering its current status in the organization as per their judgement. (how likely the requirement is met at present)

The survey tool used here is based on DQ patterns where the thirty-three DQ patterns (with their definitions) are listed under eight main DQ dimensions. The purpose here is to use the terminology to articulate the DQ problems relevant to an area of data in concern. Owing to the fact that, not meeting a DQ requirement is a DQ problem, we ask the participants to focus on the DQ problems pertaining to each characteristic (requirement). In order to facilitate their thought process, we provide the negative form of each DQ characteristic. For example, the attribute completeness of mandatory attributes (The attributes which are necessary for a complete representation of a real world entity must contain values and cannot be null), takes its inverse form as The attributes which are necessary for a complete representation of a real world entity *contains null values*.

Once a problem is identified, they are asked to, provide their perceptual judgement about the occurrence of DQ problems and the severity of the impact of the DQ problem. A sample question is presented in Table 8.2 while the full survey instrument is given in Appendix-B.

Characteristic	Problem frequency			Severity of impact		
	Low	Medium	High	Low	Medium	High
The attributes which are necessary for a complete representation of a	☐	☐	☐	☐	☐	☐

real world entity must contain values and cannot be null.						

Table 8.2: A sample survey question

Therefore through this survey, we use the power of DQ patterns to formulate critical DQ problems relevant to the organizational dataset in concern.

The outcome of the survey provides a list of the most important DQ characteristics applicable for the data set. In assessing the importance we consider both the problem frequency and the severity of the impact of each characteristic which provides a measure of how likely the characteristic is prone to DQ problems under current circumstance. Therefore the identified DQ characteristics are the most critical DQ problems in the data set.

We used the identified DQ characteristics as input to the next phase and further elaborate them into more realistic DQ problems

8.4.2 Phase-2: Bottom-up analysis (Analysis of DQ problems)

In this phase, we analyse the real DQ problems in the organization by interviewing the stakeholders of the data set in concern. Data quality problems can be identified by either examining the physical data model or by interviewing the end users. In both cases, the identified DQ characteristics in phase-1 are used as input to the investigation.

8.4.2.1 DQ Problem identification through end users.

In this case, the stakeholders (data users, data providers, DQ team members etc.) of the data set are interviewed to identify the existing DQ problems and the potential DQ problems in the organizational dataset in concern. We propose group interviews as a problem elicitation method since it enables the exchange of ideas between participants and pave the way for a rich discussion (Rosemann and Vessey 2008).

Yang et al. (2014) have pointed out that unstructured interviews have limitations in identifying realistic problem scenarios and hence efforts have to be made to structure the interviews with end users as much as possible. Therefore we suggest structuring the interview using DQ characteristics identified in phase-1. Also, it is important that the group interview should not limit to those characteristics. Prior to the interview participants can be provided with the required knowledge about the DQ patterns (DQ characteristics in particular) and during the interview, the definition of each DQ characteristic should be presented to the participants with examples before asking questions about the

DQ problems caused by it. Further, we can use the knowledge of the relationship between patterns to (observed in Chapyer-7) to move from one characteristic to another ensuring a wide coverage of the related context.

The following example scenario is presented to illustrate how the patterns are used in structuring the interview questions. Here we consider a scenario regarding perinatal data in a hospital. Let's assume that we have found the following 3 DQ characteristics as critical based on in phase-1 context analysis.

1) Ease of data access
2) Business rules violations
3) Completeness of mandatory attributes

A group interview is conducted with the perinatal data users

Question: What are the problems faced with regards to "ease of data access"?

Answer: Perinatal data has to be further processed and transformed to make them suitable for the federal government reporting requirements. This task is a tedious, time-consuming, manual task.

Note: As per the relationships between the DQ characteristics identified in Table 7.32 in Chapter-7, we know that *ease of data access* influence data *punctuality*. Hence the interviewers can ask a question based on *data punctuality* as follows.

Question: Are there any data punctuality issues because of the above problem?

Answer: Yes, we have to submit the report by the 7th day of each month and often we fail to submit the report on time.

Note: Further *ease of data access* is influenced *by data access control*, *continuity of data access* and *data awareness.*

Question: Are there any problems regarding continuity of data access?

Answer: Yes, downloading patient data is a very slow process because of the volumes and sometimes we get timeout errors.

Question: Any problems due to access controls or data awareness with regards to perinatal data?

Answer: Perinatal data exists in various locations in the system and sometimes when system updates happen same data can be recorded in different locations. In such cases, we are unaware of the best possible location to pick perinatal data.

With regards to access control, we have required access to all data.

Question: Any problems regarding business rules violations?

Answer: In patient data, we find men having babies, therefore, they end up in perinatal data which is wrong.

Question: Any problems regarding completeness of mandatory attributes

Answer: sometimes in appointment records date referred is missing

From the above example scenario it is apparent that DQ problems can be elicited through a well-conducted interview(s) and using patterns as a guidance to find the problems more effectively and efficiently. The interview process should continue until a saturation point is reached where there are no more new DQ problems encountered.

8.4.2.2 DQ Problems identified through physical data model

In addition to the problems found through the stakeholders, some of the DQ problems related to declarative DQ characteristics can be found in the physical data model itself. In this case, the actual data instances are examined for quality problems which are known as data profiling. Data profiling is a well-developed technique whereby one examines the data available in a database or a file and collect statistics and information about that data to understand its quality (Sadiq 2013). As per Lindsey (2008), this technique was first introduced by Evoke software (2016) in the late 90s and later adapted by many commercial software vendors and currently data profiling remain as commercial tools level without much exposure to the concepts and techniques behind it. Hence in order to perform data profiling, we use a generic methodology developed in academia by Zhang et al. (2014). The authors have developed a methodology for identifying problems from physical data model using data quality patterns developed by (Jayawardene et al. 2013a). Therefore we recommend to use this methodology to identify data quality problems in existing databases and the declarative DQ characteristics found in phase-1 can be used as input to this process.

8.4.2.3 DQ goal identification

Once the DQ problems are identified sufficiently, a high-level DQ requirement can be defined to address each problem using the corresponding DQ pattern. For example, the problems identified above with regards to perinatal data can be translated into high-level DQ requirements by using the corresponding pattern definition itself.

DQ problem	Corresponding Pattern	High-level DQ requirement (DQ goal)
Perinatal data has to be further processed and transformed to make them suitable for the federal government reporting requirements which are a tedious, time-consuming, manual task	P8: Ease of data access	**Perinatal data** should be easily accessible in a form that is suitable for the reporting requirements of the federal government
Fail to submit the perinatal data report on time	P9: Data punctuality	**Perinatal data report** should be available by the 7th day of each month.
Downloading perinatal data is a very slow process because of high data volumes and complex SQL statements. Thus often result in a timeout errors.	P5: Continuity of Data Access:	The technology infrastructure should not prohibit the speed and continuity of access to **perinatal data.**
Report preparation staff is unaware of the best possible location to pick perinatal data.	P7: Data awareness	Data users should be aware of all available **perinatal data** and its location
In patient data we find men having babies, therefore, they end up in perinatal data which is wrong.	P16: Business rules compliance	**Patient data** must comply with the business rules If **Gender** = Male then **babies** should be null
sometimes in appointment records date referred is missing	P1: completeness of mandatory attributes	**The date referred** which is mandatory for a complete representation of an **appointment** must contain values and cannot be null.

Table 8.3: DQ Goal Identification

It should be noted that DQ goals need not be realised only through the existing DQ problems. Van Lamsweerde (2001) pointed out that the goals may correspond with future requirements. Therefore, the future DQ requirements should also be considered to ensure that the goals are complete. For this, the stakeholders of data can be asked about future business initiatives and DQ patterns can be used directly to identify DQ requirements with regards to future business initiatives.

As per Drucker (1995) goals have to be specific and measurable. Therefore in DQ context, a DQ pattern represents a generic DQ goal since each DQ goal is defined from a DQ characteristic that relates to a validation metrics and a verification metric. Therefore the high-level DQ requirements identified based on patterns can be seen as DQ goals.

8.4.2.4 DQ Object identification

As per Van Lamsweerde (2001), the purpose of the object model is to identify the objects attributes and relationships hidden in a goal specification. Hence in the context of DQ, we adapt this step to specify the data associated with the DQ goals identified in the goal model. The pattern corresponding to each goal specifies a data granularity level which can be either Data element (E), Data record (R) or information object (IO). Hence at this stage, it is required to specify the data objects using semantics. For example in the above-identified goals, the associated data objects are Perinatal data report (IO), Patient record (R), Appointment record (R), Gender (E), Babies (E) and Date referred (E)

As per Van Lamsweerde (2001) object model can be used to identify new goals by referring to the related objects or decomposing the objects into smaller components. Therefore at this stage, the identified data objects can be used in turn to identify new goals. It should be noted that information objects can be further decomposed into data records and data records can be decomposed into data elements. Therefore at this step, it is required to decompose the identified data associated with a goal into lower level data granularity and identify all the data structures until atomic data units are reached. Then focussing on these data further questions can be asked and new goals can be identified. For example, perinatal data report is an aggregation of patient records and appointment records (Figure 8.1). Therefore questions can be asked focussing on patient records and appointment records.

Question: Do you have any problems with appointment records when you use them for perinatal data report?

Answer: the outcome of an appointment is sometimes left blank whereas it is a mandatory attribute when perinatal data is concerned.

Question: Any problems found in patient records when they are used in the perinatal data report.

Answer: We find duplicate records for the same patient

Above two answers leads to two more goals based on pattern-1 (completeness of mandatory attributes) and pattern-23 (Uniqueness)

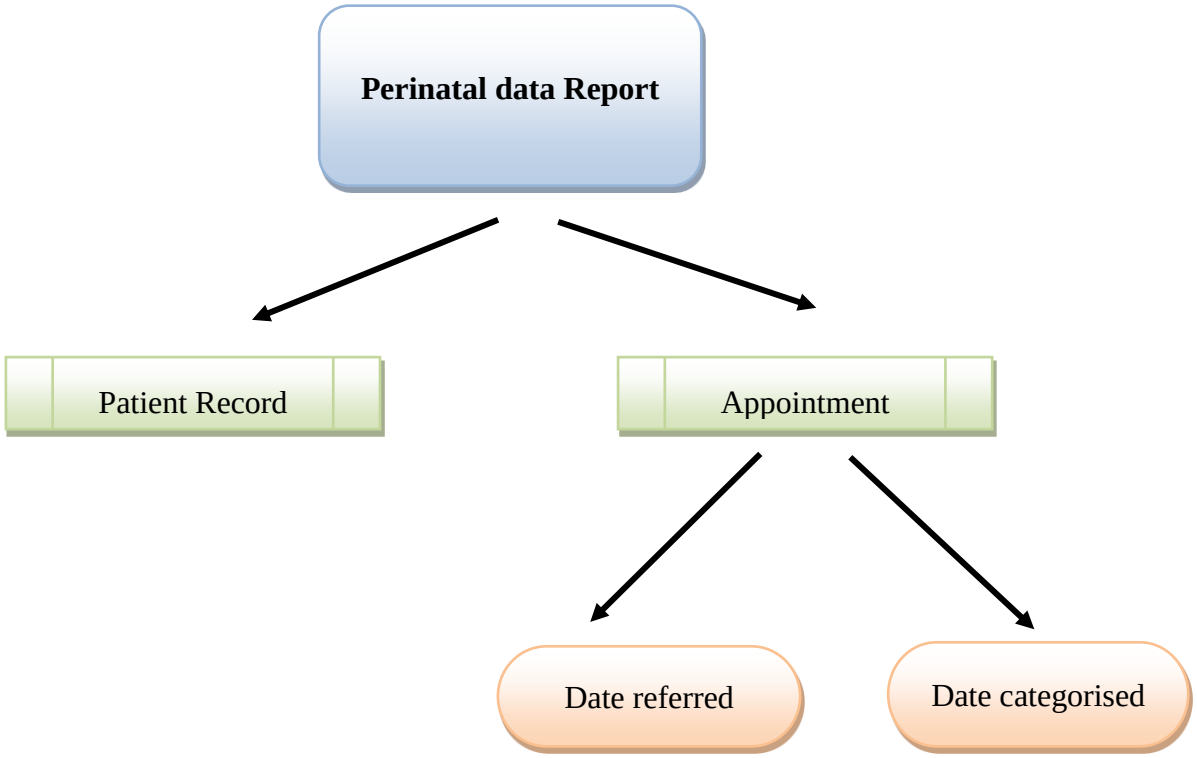


Figure 8.1: Decomposing data objects

Further questions can be asked by referring to the individual data elements in appointment record (date referred, date categorised) and patient record and new goals can be iteratively identified.

Hence when we decompose the information objects into records and elements, new data quality goals can be identified since DQ problems in lower level data may arrive in the discussion gradually.

The interviewing process should continue until a saturation point is reached in terms of DQ requirements and data objects at a given point in time. Under practical circumstances we suggest this to be an ongoing process since new DQ problems emerge throughout the time giving birth to new DQ requirements.

8.4.2.5 DQ responsibility identification

Once DQ requirements and relevant DQ objects identified satisfactorily, then it is required to move ahead with considering responsibility perspective. Therefore for each DQ requirement, three important tasks will be conducted.

I. Assigning data stewards

In data quality management, the person responsible for the quality of data is termed as *data steward* (Batini and Scannapieco 2006; English 2009; Loshin 2001; Redman 1997). Data stewards hold the ultimate responsibility for all processes policies and procedures in place for managing the quality of a data object. Therefore at this stage, all important data objects attached to DQ goals will be assigned with stewards. In practical situations, the stewardship of data may be shared across different teams and individuals (Loshin 2006) who are involved in data creation and data manipulation in addition to the end users of that data. Therefore multiple personals /functional units can be assigned as stewards considering the impact/authority they hold on the data object.

For the above example scenario, separate stewards can be assigned to the objects, perinatal data report, patient record and appointment record. Then stewards each steward is responsible for all the DQ requirements attached to that object.

II. Validating DQ requirements

Validation of goals and objectives has been recognised as an important aspect in KAOS as well as other goal oriented methodologies. (Prasanna et al. 2009; Yang et al.). Due to the subjectivity brought into the process with the involvement of human beings in eliciting and defining the DQ goals, it is essential to validate the identified goals at this stage. Therefore assignment of data stewards can be considered as a convenient point to validate the DQ goals.

Prasanna et al. (2009) suggest that it is necessary to conduct brainstorming sessions with the stakeholders of each requirement and revisit the context of the requirements. Hence we suggest that the DQ goals and objects have to be validated through group interviews with the participation of both data stewards and the real data users along with other parties who provided information to identify the goals and objects. During the group interview the goals and objects can be validated and further new DQ goals and objects may emerge.

III. Verification metric

Each DQ pattern provides guidelines for defining a verification metric. Verification metric is a measure whether the goal has been achieved or not at any given point in time. Since the metrics highly depends on the context, it is vital to define them referring to the context. Therefore during the same brainstorming sessions used to validate the goals can be used to define verification metrics for each goal. Each DQ pattern provides a generic form of verification metrics which can be considered as a guideline to develop a context based metric (Table 8.4). Therefore the stewards and data users

can effectively discuss the most appropriate and feasible verification metric for the DQ requirements in concern and an agreed threshold for the metric value.

Therefore using this verification metric, at any point in time, it is possible to check if the required DQ goal has been achieved or not.

DQ requirement	**Perinatal data report** should be available by the 7th day of each month.
Pattern	P9: Data punctuality
Generic verification metric(s)	The number of tasks failed or underperformed due to lack of data punctuality The number of complaints received due to lack of data punctuality
Verification metric	The number of complaints received **from federal government** due to lack of punctuality of the **perinatal data report**
DQ requirement	**The date referred** which is mandatory for a complete representation of an **appointment** must contain values and cannot be null.
Pattern	P1:Completeness of mandatory attributes
Generic verification metric(s)	The number of null values reported in a mandatory attribute per thousand records/ per month
Verification metric	The number of null values reported in **date referred** attribute per hundred **appointment records**/per month

Table 8.4: Definition of verification metrics

8.4.2.6 DQ implementation form

As per KAOS, the operation model describes all the required functions performed by agents to achieve a goal. Similarly in DQ once all DQ goals and objects are finalised, and responsibilities assigned, in this phase the DQ team will evaluate the implementation form to meet the goals. The implementation form of a DQ requirement can take either *process based approach* or *rule-based*

approach depending on the type of the DQ characteristic of the pattern. Each DQ pattern provides guidelines for implementing the DQ requirement. Therefore the guidelines can be used to

The process-based approach is focussed on the implementation of capabilities and processes to meet the DQ goals derived from DQ characteristics of type *usage*, while the rule-based approach is focussed on implementing rules to meet the DQ goals derived from characteristics of type *declarative*. Therefore, depending on the characteristic type, each pattern provides guidelines to consider in designing DQ processes or DQ rules and hence they can be used in designing the required rules or processes. Hence the high-level DQ requirements are finally translated to the operational requirements at this stage in terms of DQ rules and DQ processes. In designing the DQ rules and processes, the DQ team should consult the relevant SMEs (subject matter experts) to design DQ processes and rules. Therefore this will be a collaborative task.

Further in order to make sure whether a goal can be achieved, every pattern recommends a validation metric. Generally, a validation metric measures the extent to which the recommended operational requirements have been implemented at any given point in time. Since it may take a while to implement all the operational requirements it is necessary to measure the implementation progress to monitor its progress. Therefore it provides an assurance on the achievement of the goal in future. Table 8.5 shows an example operational DQ requirements and validation metrics for two selected DQ requirements.

DQ requirement	**Perinatal data report** should be available by the 7th day of each month.
Pattern	P9: Data punctuality
Implementation form	Process-based approach (implement required capabilities and processes to improve punctuality)
Relevant Implementation guideline(s) for the context	Identify the bottlenecks in the information delivery process and provide resources to remove them
DQ Process (Operational requirements)	Each division of the hospital should download the appointment data during the last day of the month and submit to the data analyst division.

	Dedicate an individual staff member to aggregate appointment data (since appointment data is the most time-consuming aggregation in preparation of perinatal data report)
Generic validation metric	*The extent to which required capabilities and processes have been identified and implemented to maintain data punctuality*
Validation metric	How many divisions submit appointment data on time
DQ requirement	**The date referred** which is mandatory for a complete representation of an **appointment** must contain values and cannot be null.
Pattern	P1:Completeness of mandatory attributes
Implementation form	Rule based approach (Implement rules to prevent or detect null values)
Relevant implementation guideline(s) for the context	Specify which attributes are required to maintain a meaningful representation of an entity and create validation rules.
DQ Rules (operational requirements)	Implement a front-end validation rule to prevent null values being entered into date referred
Generic validation metric given by the pattern	The extent to which required rules have been identified and implemented to maintain the mandatory attribute in concern
Validation metric	The extent to which the validation rule for date referred has been implemented

Table 8.5: Definition of validation metric

Therefore the operational DQ requirements are the real requirements that should be implemented successfully to achieve DQ goals. In other words, they are the key towards achieving high-quality data.

8.5 Utility of DQ patterns in modelling DQ requirement

As discussed in the above steps, identification of DQ requirements is a systematic process starting from DQ problems and finally realising DQ operational requirements. This process ensures that every operational DQ requirement has a reason and are value adding for the quality of data. Formalization provided by DQ patterns were used in every step to improve the effectiveness and efficiency of requirements elicitation. In defining DQ problems the patterns provided generic DQ problems through the negative form of its respective characteristic which was used to elicit specific DQ problems in the context. In defining DQ goals (high-level requirements) the patterns provide generic DQ goals through its characteristics which could be used to define specific DQ goals in the context. In assigning DQ responsibilities, the patterns provide a generic form of guidelines to define specific verification metrics in the context. Finally in identifying operational DQ requirements, the patterns provide a generic implementation form with guidelines to design specific DQ operations in the context along with a validation metric to measure the progress of the implementation.

Therefore the generic DQ pattern template is instantiated incrementally at every step starting from high-level DQ requirement to the operational DQ requirement systematically during this process. Thus a DQ pattern can be considered as a useful construct to represent DQ requirements for an organization. We provide two such example instances for illustration purposes in Table 8.6 and Table 8.7

Requirement number: 2	**Base pattern : P9 (Data Punctuality)**
DQ goal: Perinatal data report should be available by the 7th day of each month.	
Data object: *Perinatal data report*	
Verification metric: The number of complaints received from federal government due to lack of punctuality of the perinatal data report	
Implementation form: Process-based approach (implement required capabilities and processes to improve punctuality of perinatal data report) • Each division of the hospital should download the appointment data during the last day of the month and submit to the data analyst division. • Dedicate an individual staff member to aggregate appointment data	

Validation metric: How many divisions submit appointment data on time

Table 8.6: Instance of P9 (Data punctuality)

Requirement number: 6	**Base pattern : P1(Completeness of mandatory attributes)**
DQ goal: Date referred which is mandatory for a complete representation of an appointment must contain values and cannot be null.	
Data object: date referred	
Verification metric: The number of null values reported in date referred attribute per hundred appointment records/per month	
Implementation form: Rule-based approach (Implement rules to prevent or detect null values) • Implement a front-end validation rule to prevent null values being entered into date referred	
Validation metric: The extent to which the validation rule for date referred has been implemented	

Table 8.7: Instance of P1 (Completeness of mandatory attributes)

KAOS has proposed a graphical notation to represent four models while in this thesis we do not provide a notation since our objective is not to provide a new notation. However, it should be noted that a modelling notation can be developed to represent generic DQ patterns using the meta-model of the DQ requirement developed in chapter-5. The meta-model for a DQ requirement has been developed using data granularity level as a construct so that an instance of a DQ pattern can be connected to an existing database tool like data object or data dictionary. Therefore a modelling grammar can be developed in future using sufficient technical interfaces to develop instances of DQ patterns (DQ requirements). Thus DQ requirement scripts can be developed and can be connected with database tools so that the DQ requirements can be viewed as a part of database tools like data dictionaries and data catalogues.

8.6 Chapter summary

In this chapter, we examined the KAOS methodology used in requirements engineering in software development. KAOS is a goal-oriented methodology and we adapted this methodology to model DQ

requirements. The methodology is focused on four types of models viz. Goal model, object model, agent responsibility model and operational model where a high-level DQ requirements identified referring to a DQ problems can be elaborated into a low-level DQ operational requirement. DQ operational requirements are the actual requirement that should be implemented to achieve a particular DQ goal. It should be noted that these requirements were derived as a result of a systematic process of evaluating the four models specified by KAOS. DQ goals and objects are identified in phase-2 and it should iterate until a saturation point is needed in terms of DQ problems with regards to a particular data set. The responsibility assignment is considered as a validation point for DQ goals and objects. Then finally the DQ team design the operational DQ requirements to be implemented to achieve the DQ goals.

Chapter 9

APPLICABILITY OF DQ PATTERNS

9.1 Overview

The main objective of this chapter is to evaluate how the pattern based approach is applicable to model real world data quality requirements. As explained in chapter-3, Rosemann and Vessey (2008) has proposed three parameters (importance, accessibility, suitability) to consider in evaluating whether an artefact is applicable in research. Based on these parameters we check the following perspectives about the DQ patterns to evaluate their applicability.

4. How useful DQ patterns are in analysing and modelling DQ requirements (importance)
5. How well the data quality users connect with the artefact (accessibility)
6. How accurately the DQ requirements were elicited and modelled (suitability)

In applicability checks, Rosemann and Vessey (2008) suggest to conduct group interviews using industry practitioners and present the artefact for their feedback. The feedback is collected in response to the answers to the questions designed to evaluate the above three parameters. But in this applicability check, we take a step forward and practitioners' feedback is collected by using the artefact in a practical setting. Thus we use the DQ patterns in two organizations using the methodology KAOS4DQ and analyse and model the DQ requirements relevant to a particular data set. The practitioners' feedback on the artefact is collected with sufficient trail of evidence from the organizational context so that it is not just their perception about the artefact's applicability, but a comprehensive analysis on how the artefact could be used in organizational context. Such an evaluation provides a more realistic measure on the applicability of the artefact. Therefore, we conducted two applicability checks in two organizations with evidence of DQ management practice.

9.2 Protocol for the applicability check

This study was designed in such way that it follows the steps of the KAOS4DQ explained in Chapter-8. The full protocol is given in Appendix-A while we explain the outline of the protocol as follows.

Phase-1: Top-down analysis (Analysis of DQ context).

In KAOS4DQ, the main aim of phase-1 is to recognize the DQ context of the organization. A survey tool (see Appendix-B) is used with high-level executives to recognize the most relevant DQ patterns applicable to the organization and they are used as input to phase 2 so that a more focused interview can be performed. We used the DQ manager in each organization as the participant of this phase

considering his broad exposure to the organizational DQ problems. In answering the survey he was asked focus on a particular dataset in the organization so that we could limit the focus of this study.

The survey tool (questionnaire) used in phase-1, was pilot tested with a well-experienced data quality practitioner attached to International Association of Data and Information Quality (IAIDQ) who has served as a chief data quality officer in an Australian organization operating in oil and gas manufacturing. Then we also used members of a research group within the university who use research data and the questionnaire was used to elicit DQ problems in their data sets. After the pilot testing, necessary modifications were done to the survey tools. The definition of the DQ characteristics in this survey was changed to its negative form to facilitate DQ problem elicitation. The participants seemed to have connected well with the generic form of DQ problems with this change. Further, the educational materials were made more clear and understandable for practitioners by eliminating academic terminologies etc.

Phase-2: Bottom-up analysis (Analysis of DQ problems).

It should be noted that in KAOS4DQ, phase-2 is an iterative process with multiple group interviews until a saturation point is reached in terms of DQ problems and thereby DQ requirements. But in this applicability check, as per Rosemann and Vessey (2008), we conducted phase-2 in a single group interview and identified a set of DQ requirements to demonstrate the applicability of the patterns in terms of importance and accessibility i.e

1. How useful DQ patterns are in analysing and modelling DQ requirements (importance)
2. How well the data quality users connect with the artefact (accessibility)

During the first two phases we received the participants' feedback about importance and accessibility of DQ patterns in the form of direct feedback (participants' comments) as well as indirect feedback (through the successful compliance with the use of patterns to identify the problems), so that we could evaluate the applicability of the patterns in terms of importance and accessibility with sufficient trail of evidence. After phase-2 then we conducted phase-3 which is a validation of the DQ requirements elicited and modelled in phase-2. The aim of phase-3 is to check the suitability of the patterns.

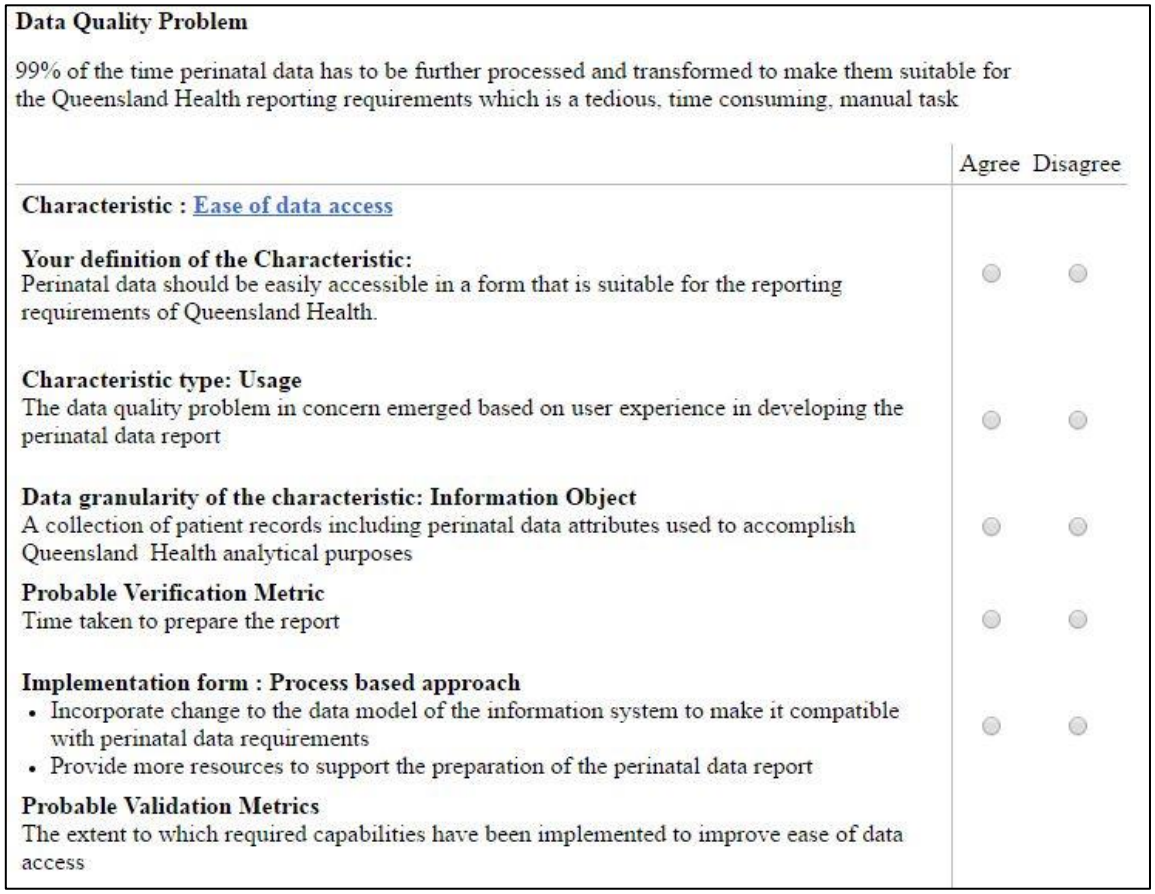
Phase-3: Validation of findings

In the third phase, we presented the identified DQ requirements to the responsible DQ professionals in the organization and checked if they agree with them. The main purpose of this validation is to check the suitability of the patterns i.e.

3. How accurately the DQ requirements were elicited and modelled (suitability)

The main information that we collect in this phase is users' agreement/disagreement about our data analysis. Since the surveys provide grounds to collect data more precisely through providing answer choices like agree/disagree, we selected a survey as a suitable methodology for this validation. For this validation, we selected the DQ manager of each organization as the participant considering his overall understanding about DQ in the organization. An online survey tool was used in this phase where the DQ requirements identified in phase-2 were presented and the participants had to either agree or disagree with each element of the requirement (DQ goal, implementation form, validation metric, and verification metric). In the case of any disagreements, they were asked to provide reasons accordingly.

In answering this survey the DQ manager was asked to get assistance from the team members (who participated in phase-2) if needed. Further in this survey, for each DQ requirement presented, the participants were asked to provide any similar DQ requirements where the same DQ pattern can be applicable. This enabled us to assure that the participant has understood the original requirement with sufficient depth and breath. A sample question of the survey is given in Figure 9.1.



Data Quality Problem

99% of the time perinatal data has to be further processed and transformed to make them suitable for the Queensland Health reporting requirements which is a tedious, time consuming, manual task

	Agree	Disagree
Characteristic : Ease of data access		
Your definition of the Characteristic: Perinatal data should be easily accessible in a form that is suitable for the reporting requirements of Queensland Health.	○	○
Characteristic type: Usage The data quality problem in concern emerged based on user experience in developing the perinatal data report	○	○
Data granularity of the characteristic: Information Object A collection of patient records including perinatal data attributes used to accomplish Queensland Health analytical purposes	○	○
Probable Verification Metric Time taken to prepare the report	○	○
Implementation form : Process based approach • Incorporate change to the data model of the information system to make it compatible with perinatal data requirements • Provide more resources to support the preparation of the perinatal data report	○	○
Probable Validation Metrics The extent to which required capabilities have been implemented to improve ease of data access		

Figure 9.1: Phase -3 Sample survey question

9.3 Introduction to the subject groups

Selecting suitable organizations for the study is a crucial task since this is a qualitative evaluation and thus information rich cases are required for in-depth analysis (Patton 2005). Coyne (1997) describes that all sampling done in qualitative research are "purposeful sampling". Marshall (1996) refers to purposeful sampling as "judgement sample" where he describes as the most common and most intellectual strategy used in academia based on the researchers' practical knowledge of the research area. In this case, the researcher actively selects the most productive sample to answer the research question considering the special expertise of the subjects based on real world evidence. Based on the above argument we used purposeful sampling technique to select the participant organizations with the following selection criteria to ensure information rich cases.

1. Participant organizations should have a dedicated DQ management team
2. Affiliations to professional bodies for DQ
3. Special recognitions/awards achieved for credible DQ initiatives

Based on this criteria we selected the best two organizations at our proximity to conduct the evaluation.

ABC hospital: A renowned hospital in Australia which has its reputation for patient care and research over more than hundred years. The hospital's information infrastructure is well equipped with a backbone ERP solution catered for the healthcare sector, supported by a CRM solution and several best of breed applications for specific tasks like medical imaging, laboratory operations, and medical research. The hospital has recognised its data as a valuable component in their organizational strategy and dedicated a separate team to manage data quality. The team has been in existence more than a decade and consist of experienced data quality professionals. The hospital has granted the corporate membership of IAIDQ (International association for information and data quality) and won the IQ excellence award in 2014.

XYZ Insurance: One of the largest insurance companies in Australia which operates in four other countries employing more than 15000 people worldwide. The company's information systems infrastructure consists of a top tier ERP system backed by state of the art solutions for CRM and BI functions. The company uses many loosely coupled distributed systems to handle its operational activities providing more flexibility to the operational staff. Further being in a very competitive market, the company has provided many online services to its customers. The company depends highly on its data and hence they have established a fully-fledged data quality team that operates 24

hours a day, seven days a week. The company is a corporate member of IAIDQ (International association for information and data quality) and won the IQ excellence award for 2015.

9.4 Applicability checks for ABC hospital

9.4.1 Phase-1 results

In the first phase of the study, the manager of the data quality team of ABC hospital answered the surveys focussing on a critical DQ scenario of their organization. He considered the perinatal data as their DQ management scenario. Depending on the manager's responses most relevant data quality patterns for the organization with regards to perinatal data was identified as follows (Figure 9.2). The data analysis process is explained in Appendix B.

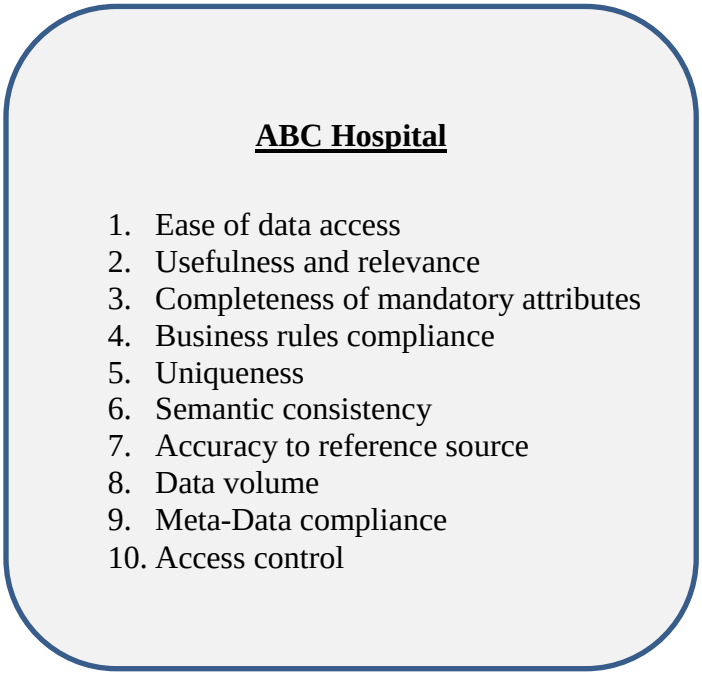


Figure 9.2: Critical DQ characteristics for ABC hospital

9.4.2 Phase-2 results

There were three participants in the group interview which included the DQ manager, a DQ analyst who is responsible for perinatal data reporting and a database technician who is assigned with the role of fixing DQ bugs. The three participants were well qualified in terms of the criteria given in the protocol for participant selection (Appendix-A).

The group interview process started with an introduction to the 33 DQ characteristics in general and focussing specifically on relevant DQ characteristics found from phase-1. The participants had

studied the materials on the full list of DQ characteristics that we had provided few days prior to the interview and they seemed to have comfortable with it. Once the introduction was over the interview process started and went on for one hour and 10 minutes.

During the interview, the participants were focused on perinatal data, which they considered as a challenging aspect in their DQ management portfolio. Perinatal data reporting is a federal government requirement where every hospital has to periodically send a report in a prescribed format. The federal government uses perinatal data collected from all hospitals for various research purposes around birth complications, medication and other treatments, ethnicity, lifestyle and a wide range of other purposes. Perinatal data reporting has almost no clinical value and the hospital captured this data solely for reporting purposes. Hence, it had been an ongoing crisis between the staff to get this data entered into the system by the clinical staff at various stages of patient management process through several systems/databases leading to DQ issues.

During the group interview, we managed to focus on six characteristics out of the top 10 listed in Figure 9.2 and the discussion around them revealed eleven DQ problems and their context information. The problems were analysed based on the protocol and DQ requirements were modelled using patterns. A summary of the findings is as follows.

DQ problem-1: 99% of the time perinatal data has to be further processed and transformed to make them suitable for the reporting requirements of the federal government. Therefore, it is a tedious and time-consuming, manual task.

The participants revealed the problem with reference to the first pattern brought to the discussion P8: Ease of data access and the problem seemed to have been well positioned with the pattern. During the interview, the participants revealed that the required perinatal data contained 142 attributes that exist in 32 different data structures, and the report requires them to aggregate the full details quarterly. Due to the complexity of the requirement, the system is still unable to produce this report with a single button click.

The DQ requirement pertaining to this problem was modelled as shown in Figure 9.3.

DQ problem-2: Perinatal data report has to be sent to the federal government quarterly. Sometimes the hospital fails to deliver this file on time due to bottlenecks and delays in data processing.

The problem emerged naturally followed by problem one when the moderator introduced P9: Data punctuality as this pattern is influenced by P8: Ease of data access (see Table 7.32). The participants

considered this to be a very critical issue since it creates a non-compliance situation with government regulations.

Figure 9.4 shows the DQ requirement pertaining to this issue.

DQ problem-3: Failing to comply with changing requirements of federal government department with regards to perinatal data reporting

The problem emerged with reference to P29**:** Usefulness and relevance when the moderator selected the pattern based on Figure 9.2. The participants revealed that this had been an ongoing issue since the federal government needs do change often due to government policy changes and new initiatives. For example, recently the government focus had been on domestic violence and hence information regarding domestic violence related to perinatal data had to be provided.

Figure 9.5 shows the requirement elicit with regards to this problem.

DQ problem-4: The system generated report for perinatal data did not conform to the format prescribed by the federal government department. The validation system rejected the report without even processing it due to the format issue.

The problem emerged with the discussion around problem-3 when one of the researchers moderated the discussion with the introduction of P31**:** Appropriate Presentation (as per Table 7.32 in chapter-7, *appropriate presentation* influence *usefulness and relevance* and hence the moderation was done).

The DQ requirement was modelled in Figure 9.6

DQ problem-5: An appointment does not have an *outcome* and therefore data elements of patient records are not complete in terms of the required values. Hence, the Perinatal data report receives validation errors

The problem emerged with the introduction of P1: Completeness of mandatory attributes. The discussion. The discussion revealed that data elements in clinical systems cannot be made mandatory since it may impede patient care. Therefore, the attribute outcome, in fact, is an optional attribute which paved the way for occurrences of invalid null values.

Hence, this requirement was modelled using P2: Completeness of optional attributes as shown in Figure 9.7.

DQ problem-6: The Federal government department has some validation rules. Sometimes perinatal data do not comply with these business rules. For example, babies are included for patients whose gender is male and such patients ended in perinatal data report giving validation errors.

Having asked about problems related to P16: Business rules compliance the participants revealed this problem. Further, the participants revealed that business rules violations occur for other attributes of patient record such as date of birth having dates more than 100 years ago which were not real.

The requirement for problem-6 was modelled as in Figure 9.8 in the representation of all the other similar problems discussed.

The researchers focussed on the data objects *patient record* and *appointment record,* which was exposed during the discussion, and the problem-7, problem-8, problem-9, problem-10 were revealed as a result of the moderation.

DQ Problem-7: Internally used conventions for the fields "Date registered", "Date arrived", "date referred" and "dates categorised" in patient record are not matching with federal government definitions. Hence, wrong date fields have been used to calculate the patient wait times (difference between "date referred" and "date categorised")

The participants revealed this as a critical problem since the two dates in concern is used by the federal government to calculate an important statistic average patient wait time for the hospital. Due to the semantic conflict mentioned above the average patient wait time has been overstated affecting the hospital's reputation.

Figure 9.9 shows the requirement for the above problem

DQ Problem-8: It is quite common practice (and also legal) to have duplicate patient records but they have to be resolved after the patient care aspect has been fully covered. Yet a few unresolved duplicates are found.

The participants revealed that they have used state of the art technology to identify duplicates and resolve them quite satisfactorily during the last year only 32 duplicates found for 6000 records.

Figure 9.10 shows the relevant requirement for the problem.

It should be noted that the above two problems (8 and 9) are not directly related to the perinatal data reporting. But it elicited as a result of the discussion around the data object *patient record*. Similarly, the below two problems were revealed as a result of the proliferation of ideas around the *appointment record* and *patient record* respectively.

DQ problem-9: Sometimes the patient that shows up is not the same as is meant to be on the Medicare card. As a result, the clinical details recorded in the appointment record are not pertaining to the actual card holder)

The respective DQ requirement is modelled in Figure 9.11

DQ problem-10: Patient Name, phone number and address in the M system is not current. Patient name/address has changed since the last visit, and hence the issue. Figure 9.12

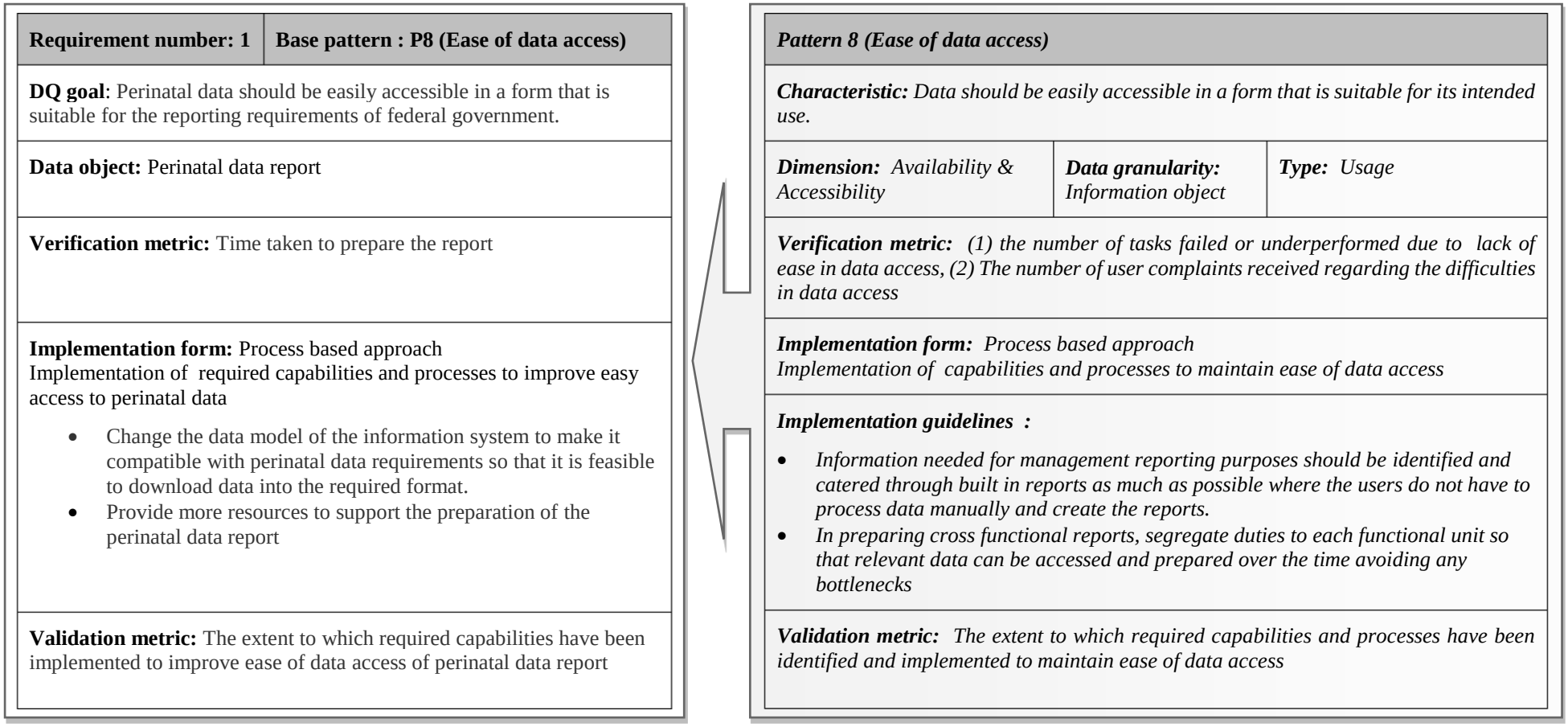


Figure 9.3: Requirement number 1 - ABC hospital

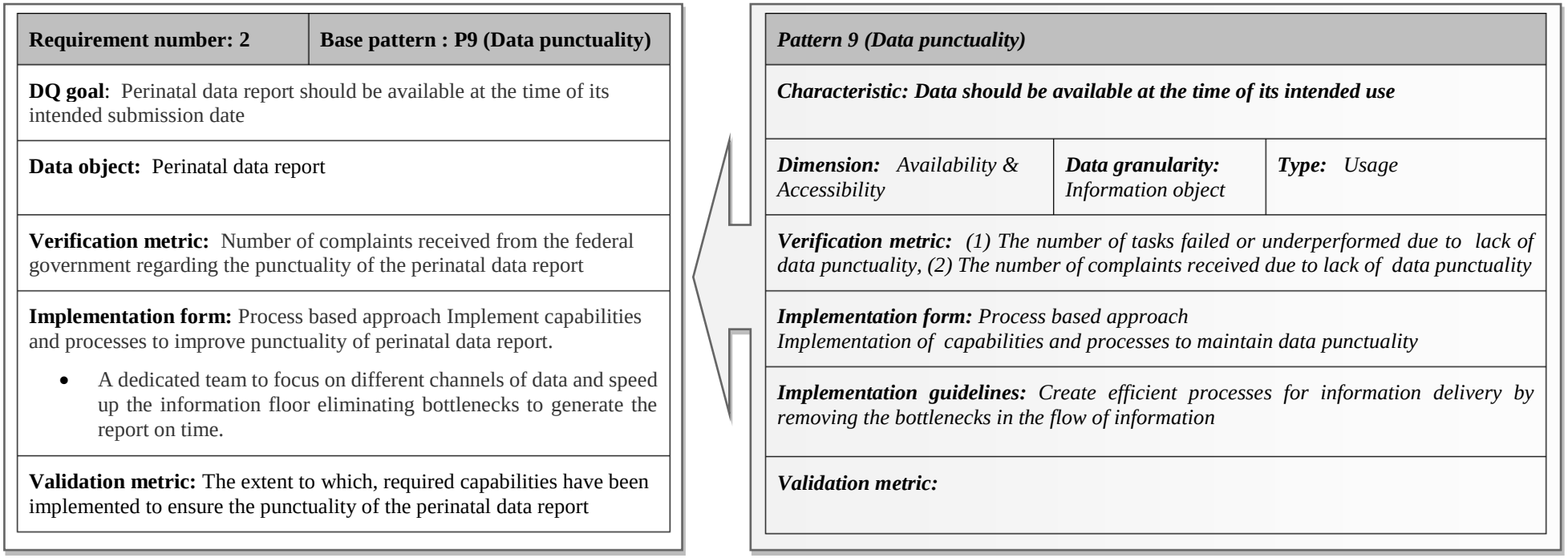


Figure 9.4: Requirement number 2 - ABC hospital

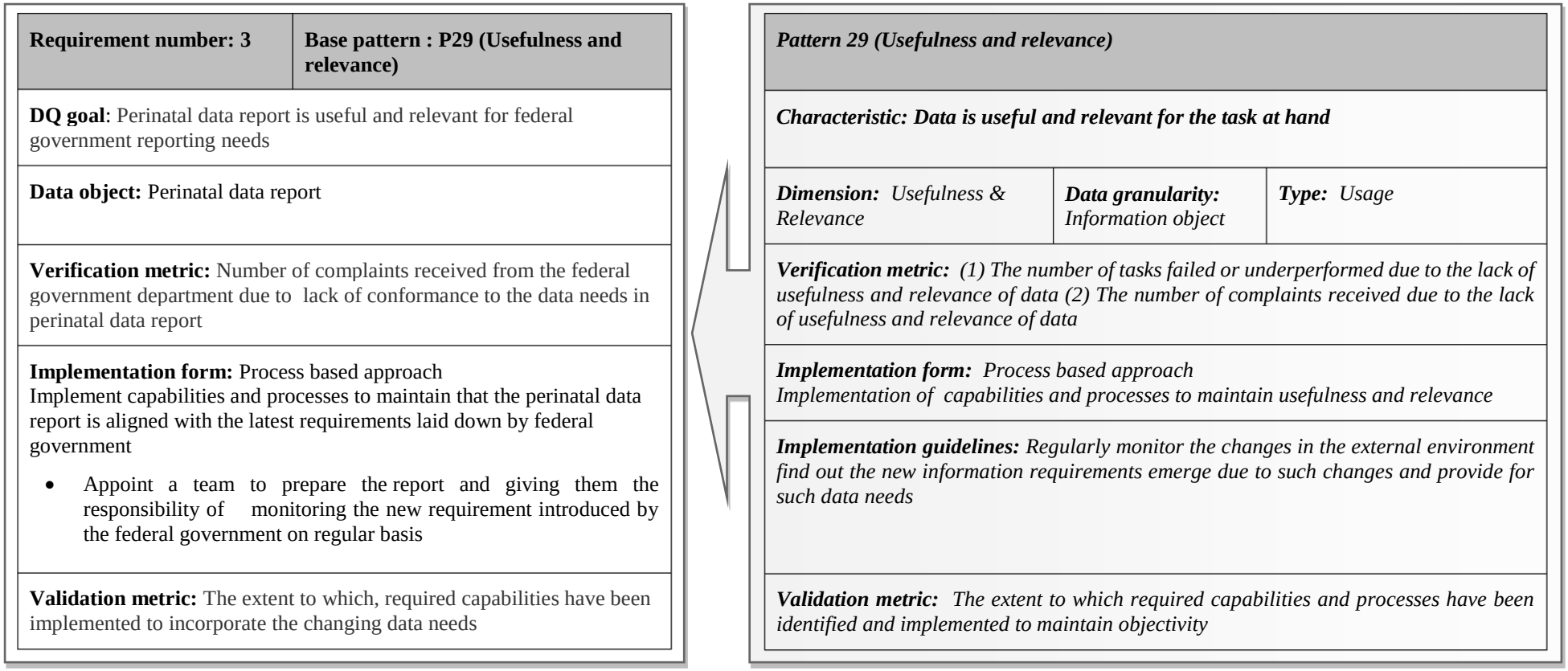


Figure 9.5: Requirement number 3 - ABC hospital

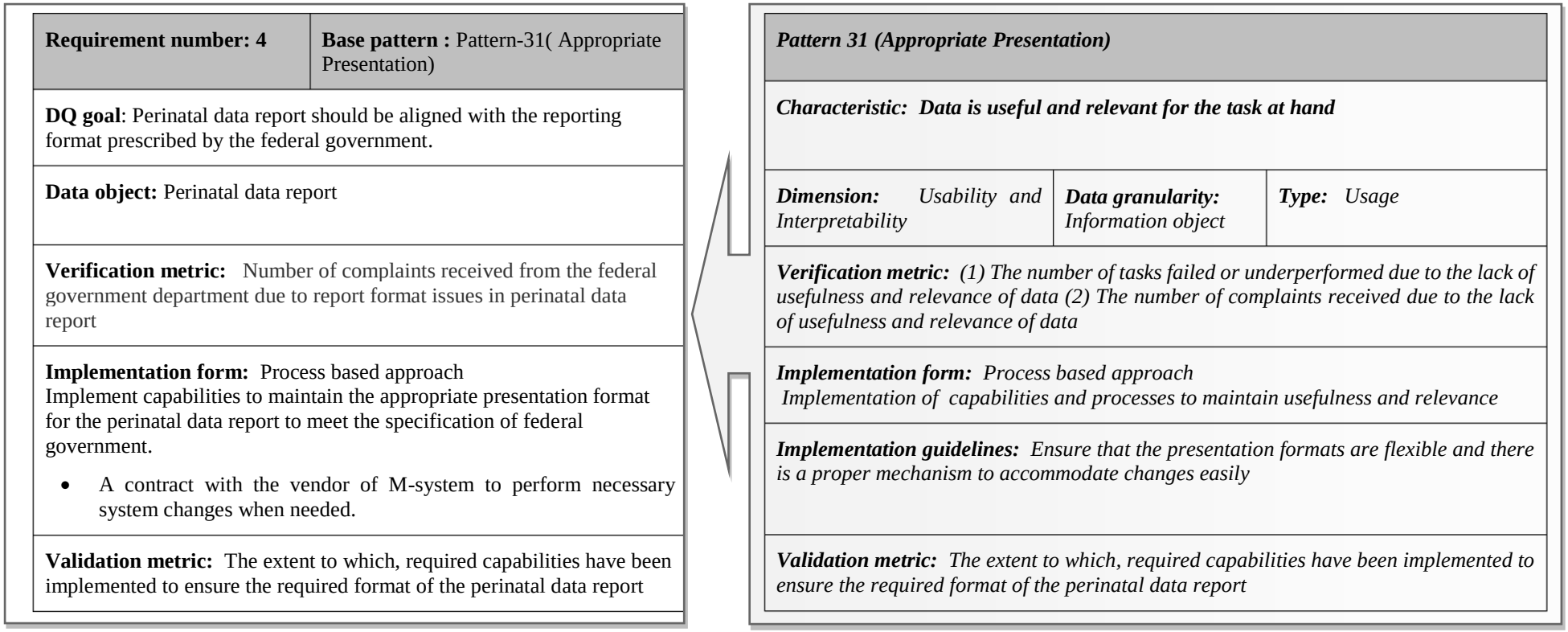


Figure 9.6: Requirement number 4 - ABC hospital

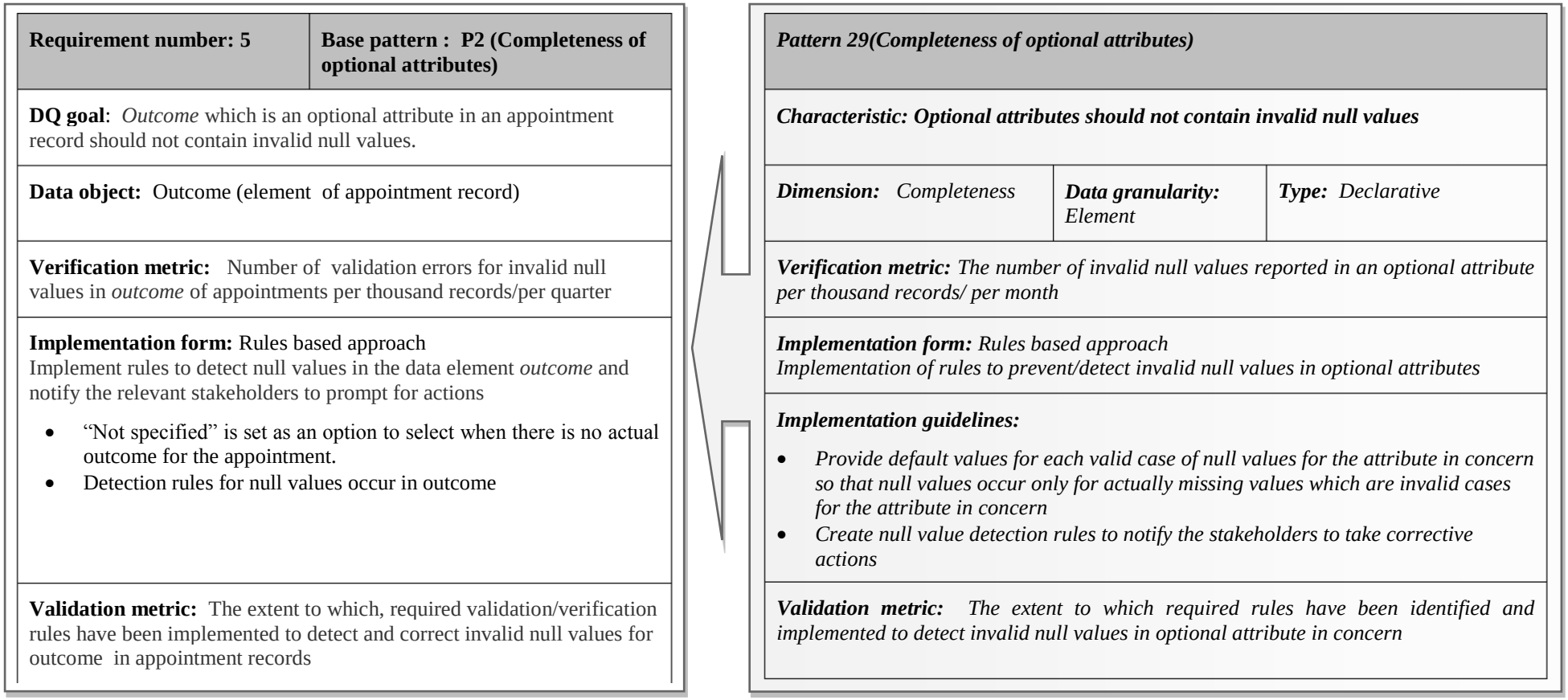


Figure 9.7: Requirement number 5 - ABC hospital

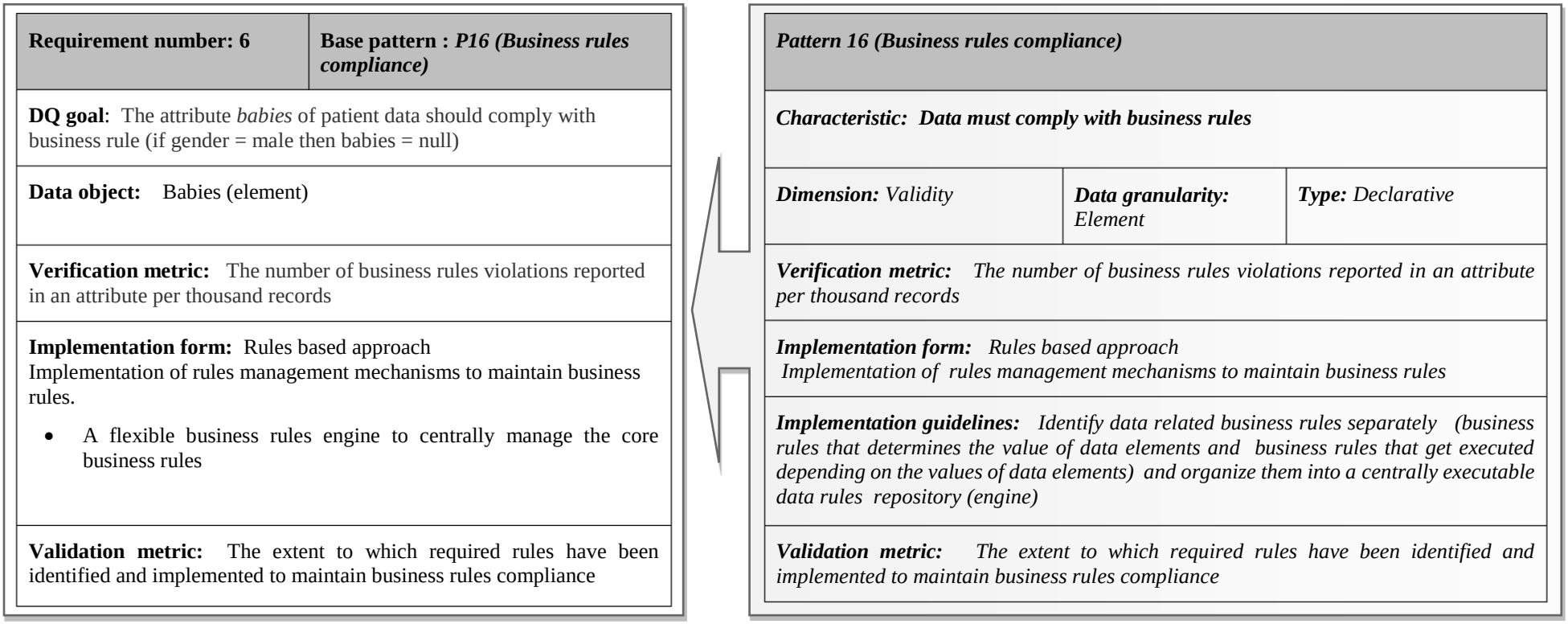


Figure 9.8: Requirement number 6 - ABC hospital

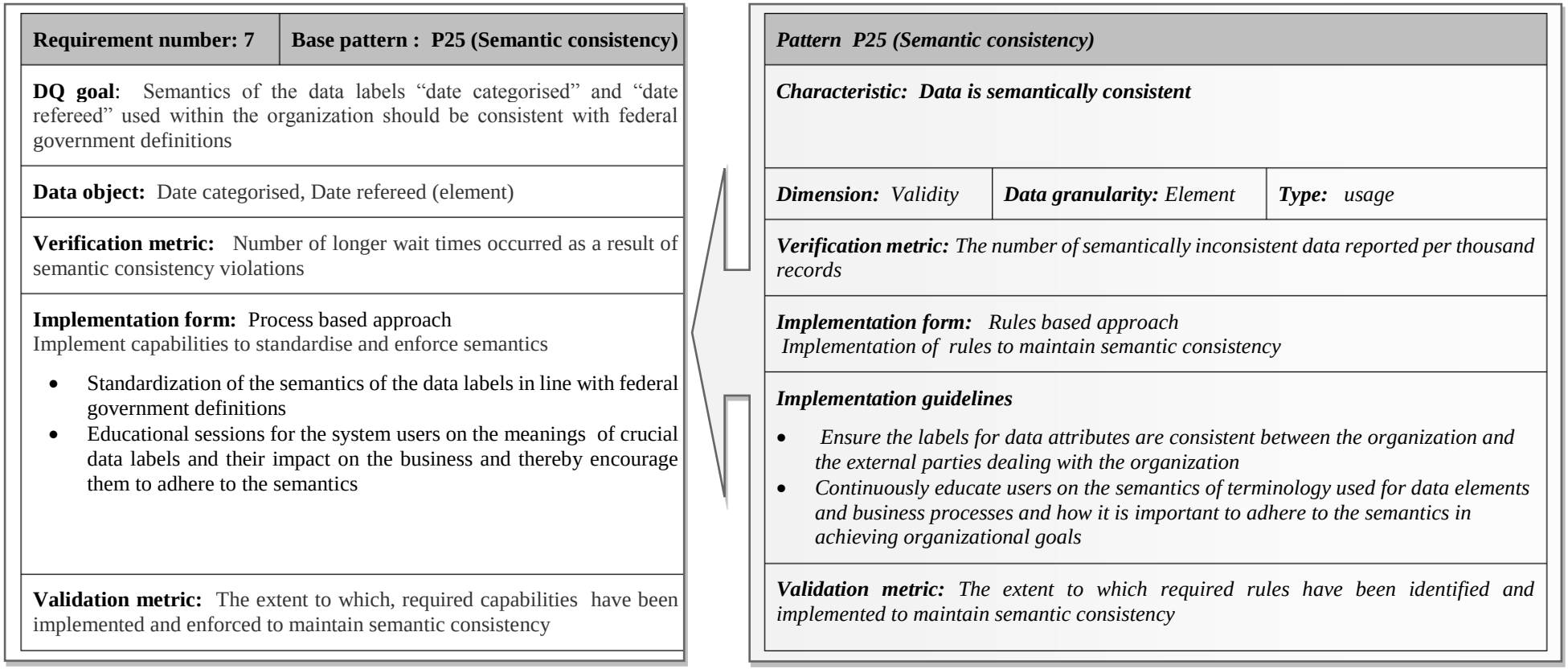


Figure 9.9: Requirement number 7 - ABC hospital

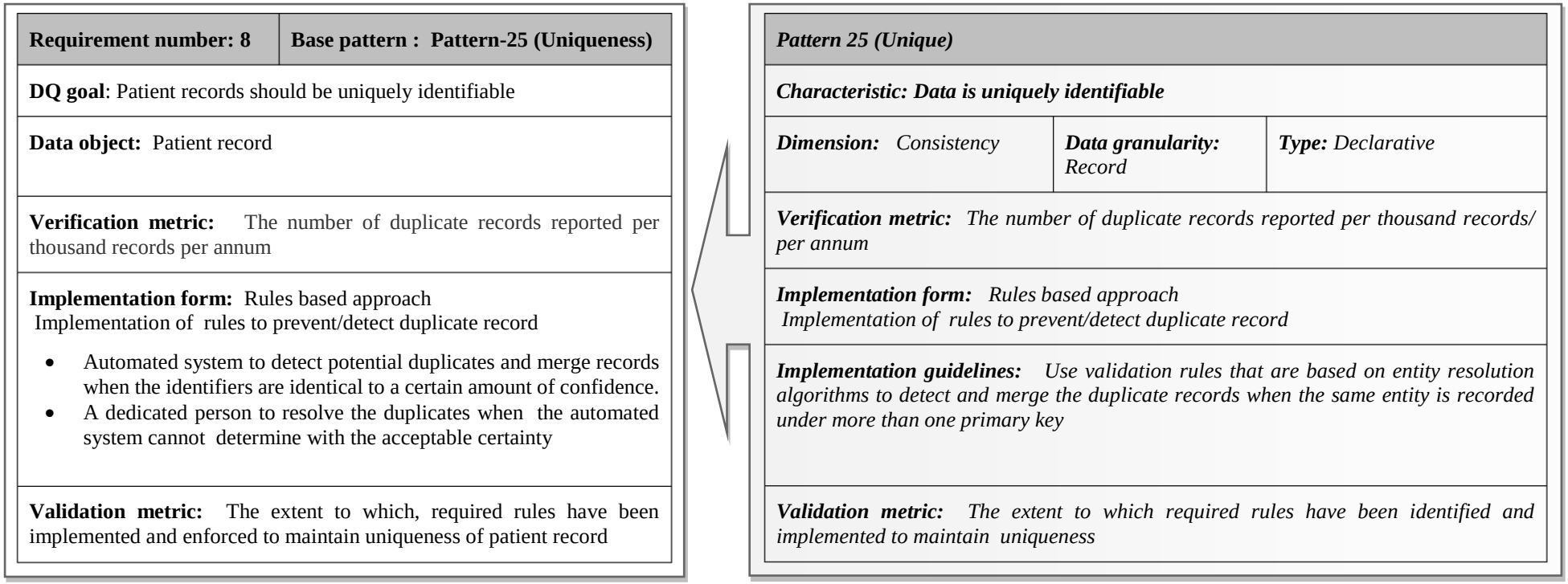


Figure 9.10: Requirement number 8 - ABC hospital

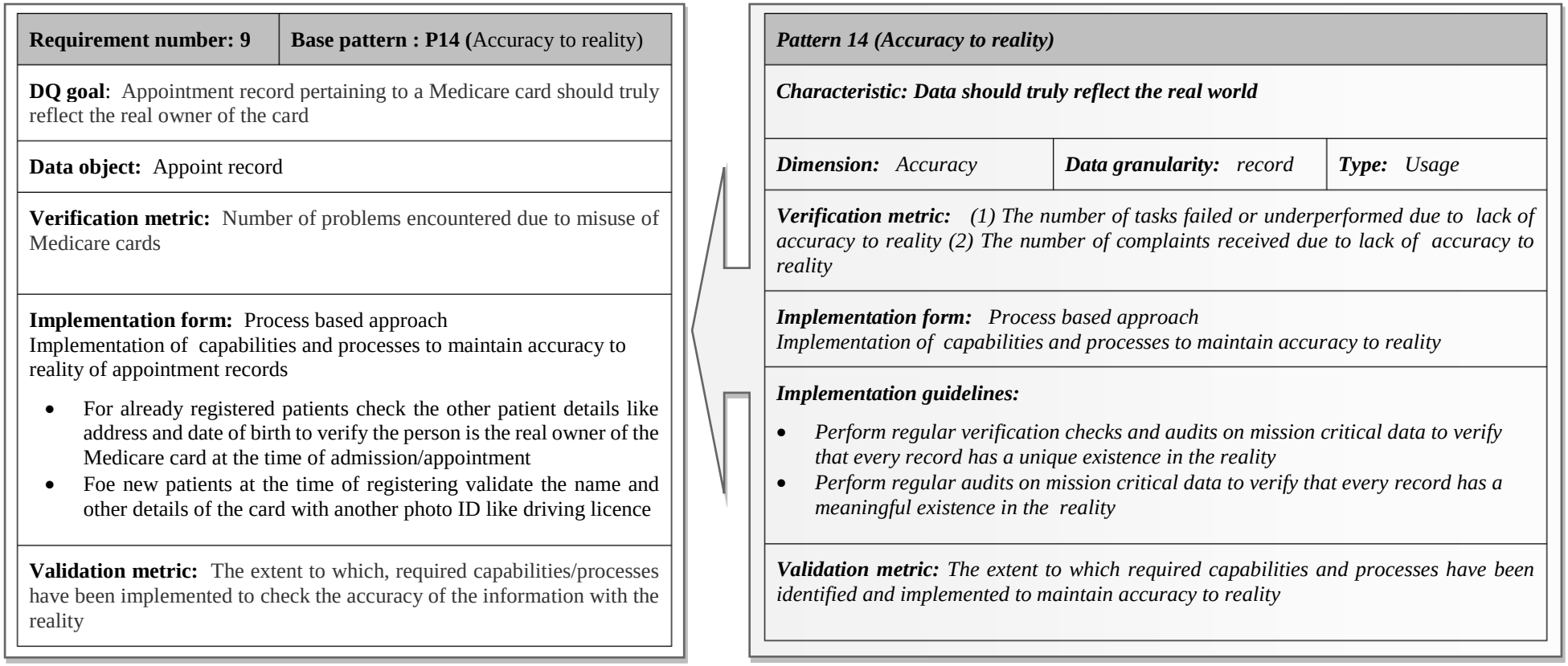


Figure 9.11: Requirement number 9 - ABC hospital

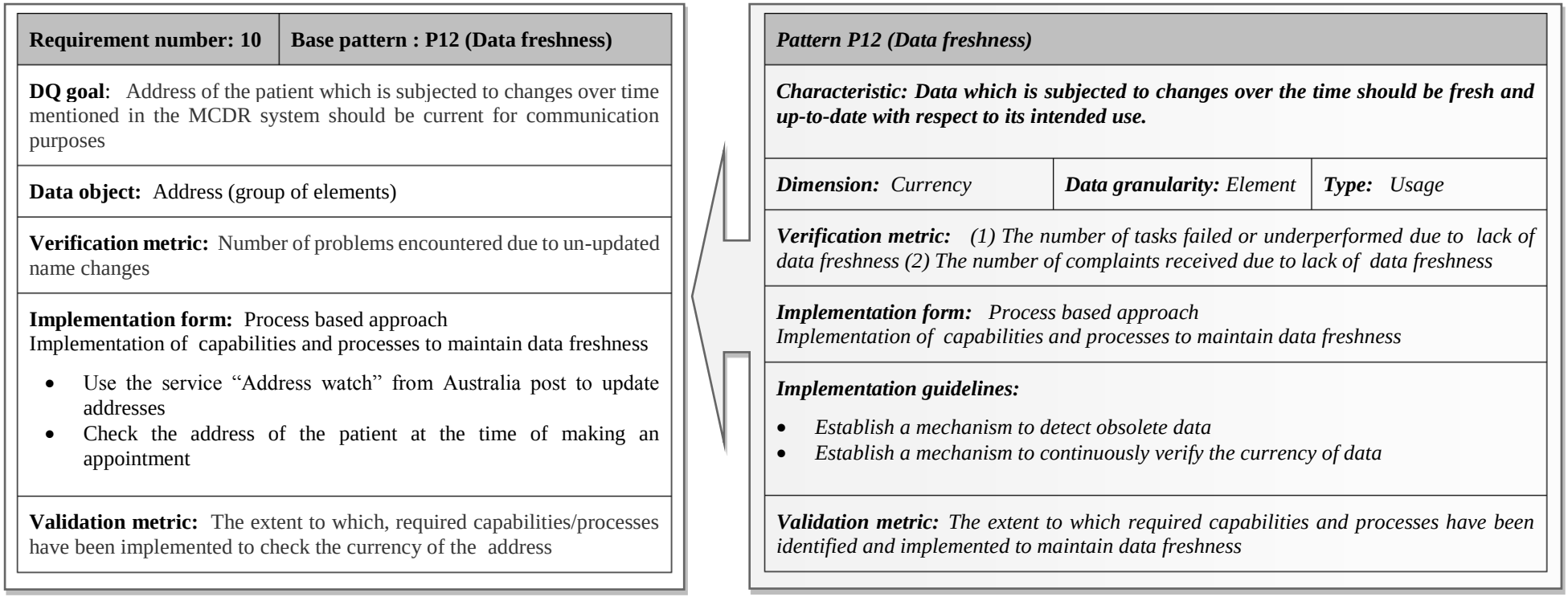


Figure 9.12: Requirement number 10 - ABC hospital

9.4.3 Phase-3 results

In this phase, the purpose was to validate the accuracy of the DQ requirements identified in the above phase. In this applicability study, we selected the DQ manager as the participant to validate the DQ requirements considering his broad understanding about the DQ in the organization. He was asked to consult the two team members who participated in phase-2 if required. An online survey was conducted where the participant could provide his consent about each DQ requirement elicit in the interview. All DQ requirements identified were presented in this survey as questions and the DQ manager was asked to provide his agreement\disagreement to each component of a DQ requirement. For each disagreement, he was asked to provide reasons. Further, he was asked to provide examples for similar DQ scenarios which can be modelled using the same pattern. The summary of data analysis is presented in Table 9.1:

DQ problem	**Applicable pattern**	**Applicable similar DQ scenarios in the organization**	**Validation status**
Problem-1	P8: Ease of data access	Some data requested by BI teams are difficult to be retrieved and prepared due to the complexity of data structures	Agree
Problem-2	P9: Data punctuality	It is difficult to meet the deadlines for BI related data due to the bottlenecks in complex data preparation tasks	Agree
Problem-3	P29: Usefulness and relevance	Cross-system linked data sets from various clinical research projects have changing data needs which are difficult to satisfy.	Agree
Problem-4	P31:Appropriate Presentation	Data from patient admission system is not appropriately presented for annual federal government reporting and often receive validation errors	Agree
Problem-5	P2: Completeness of optional attributes	No mandatory rules have been set in any clinical systems and receive validation errors when used in external reporting	Agree
Problem-6	P16: Business rules compliance	Business rules are violated in outpatient data since it is very difficult to enforce rules in data entry.	Agree
Problem-7	P25:Semantic consistency	Some standardisations of semantics in outpatient data has been done as per federal government. But still ambiguities exists among the users and at times they	Agree

		take their own decisions considering the conventional meaning	
Problem-8	P25: Uniqueness	Maintaining a unit record number (URN) for a patient across all systems including the GP data	Agree
Problem-9	P14: Accuracy to reality	Patients lie about date of birth to hide their age	Agree
Problem-10	P12: Data freshness	Some clinical data with regards to a patient should be current (eg: investigation data)	Agree

Table 9.1: Validation of DQ requirements

The DQ manager of ABC hospital provided his consent for all the DQ requirements that we had identified. Further in response to the question about similar DQ requirements where the same DQ pattern could be used, he managed to provide similar scenarios to each problem. In some cases, he has explicitly mentioned about the data objects (problem-8, 9, 10) whereas in others he has generally referred to the data objects which are a limitation. But it is clear that the responses are in line with the respective DQ patterns. Therefore, these responses can be considered as an additional proof about the DQ managers understanding about the DQ patterns. Thus we can conclude that he has agreed to the given requirements with legitimately.

9.5 Applicability check for XYZ insurance

9.5.1 Phase-1 results

In phase-1, the DQ manager of XYZ insurance answered the surveys focussing on a critical DQ scenario of their organization. She considered customer data as critical in their DQ management portfolio. Depending on the manager's responses most relevant data quality characteristics for the organization was identified as follows (Figure 9.13).

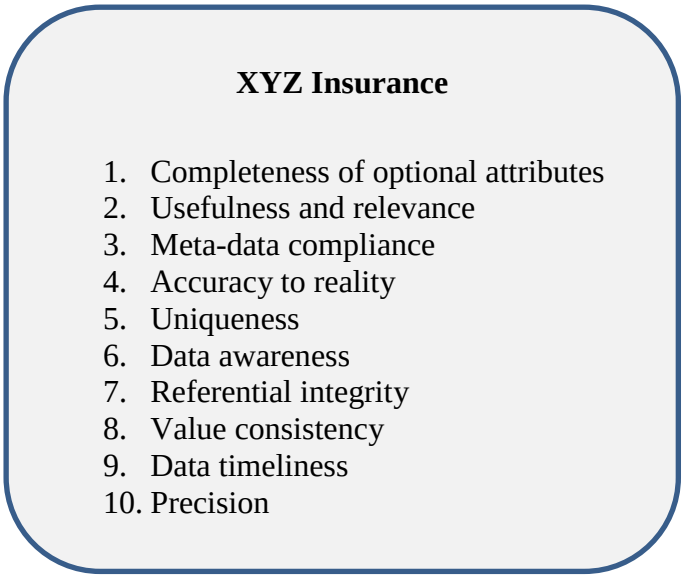


Figure 9.13: Top ten DQ Patterns for XYZ insurance.

9.6 Phase-2 results

In the group interview, as the first step, we introduced the DQ patterns identified as important for the organization based on the results of the first interview. Further, we presented other characteristics in general. The participants had already gone through the materials that we sent them prior to the interview day, and they were very familiar with the concept of DQ dimensions. This introduction revealed that the organization uses their own data quality dimensions for which they had provided their own definitions eg: "*our definition for completeness is no attempt to capture any meaningful information*". However, in this case, the terminology used in our eight main DQ dimensions provided a base to connect their terminology with DQ dimensions with our classification of 33 data quality characteristics. In order to avoid any conflicts in the terminology, we provided an extra explanation to each characteristic when we use them to elicit data quality problems. During the interview, we elicited 9 DQ problems and one future DQ requirement as follows.

DQ problem-1: Having null values for customer phone number which leads to communication problems with customers

This problem emerged at the very beginning of the interview when brought the pattern P2: completeness of optional attributes into the discussion forum. Further discussion about the problem revealed that the organization has a culture which considers speed over quality and thereby the sales staff pay no attention to entering optional attributes when issuing an insurance policy or a quotation. Customer phone number is considered as optional attributes since only a valid postal address is considered as mandatory in issuing an insurance policy. Therefore null values in the phone have to be detected and corrected afterwards (Figure 9.14)

DQ problem-2: The e-mail addresses captured for different purposes cannot be used for promotional activities. Unauthorised usage of e-mails for promotional purposes has resulted in complaints to the privacy commissioner.

During the discussion on the second pattern in the list, participants revealed that, on many occasions, the email provided for sending the contractual documents is not the same e-mail address that a customer provided to request a quotation (in which case the customer agree to receive emails from the organization). Therefore, the context of e-mails used in promotional activities should be relevant to the context. This requirement is modelled using P29: Usefulness and relevance (Figure 9.15)

The participants revealed the following two problems regarding meta-data compliance which was the third pattern in the list.

DQ problem-3: Having single character first names for customers leading to communication and identification problems (eg: Dear Mr. A)

As per the participant, the business teams do not allow to implement online validation rules for the first name to avoid single character first names being entered into the field since it may impede the core business (issuing quotations). Therefore, detection rules in meta-data are used to model the requirement (Figure 9.16).

DQ problem-4: Data goes to the wrong field when data is transferred between systems. For example, *ford* is entered into the field *VIN number* whereas it should go to the field *vehicle make*.

During the discussion, the participants revealed the following root causes for this problem.

- Brokers using their own tools to capture data and then talk to the main application which would result in data being transferred to the wrong field because of hard coding issues in the front-end and back-end of the tools

- Data migration projects where the tools were not tested properly and wrong data ended in fields
- One channel collect data into a field and another channel present that data in a different field since they know what that data really means and it serves their purpose (Fleet number into VIN number in case of insurance for shipments)

The requirement was modelled using P17: Meta-data compliance (Figure 9.17). The following two problems were revealed in response to uniqueness which was the next pattern considered in the forum.

DQ problem-5: E-mail address is used as a unique identifier of customers (used as the log-in ID for customers in accessing their record) but in reality the same E-mail address is shared by the family members. Hence, identification of the actual customers is problematic and leads to business problems in customer service function.

Even though this problem had the flavour of uniqueness it was related to P14: Accuracy to reality which was focussed on comparing data with the reality. Figure 9.18 shows the requirement.

DQ Problem-6: The same customer is captured again and again from different channels resulting in duplication of records.

The problem was revealed in relation to P23: Uniqueness and the relevant DQ requirement is shown in Figure 9.19.

The next two patterns in the list were introduced to the discussion forum but the participants were unable to connect with problems. Then the moderator selected the pattern P7: Data awareness, which was not in the initial list and the participants responded as follows.

DQ Problem-7: Bought the same motor vehicle data multiple times due to the lack of awareness of different divisions of the organization that the data exist in the organization.

It was revealed that after a new system implementation this problem occurred due to the lack of understanding about data landscape. The requirement is modelled in Figure 9.20.

DQ Problem-8: The same vehicle has been insured three or four times in its lifetime. But some of the vehicles and policy documents are not connected due to the legacy systems.

The problem revealed related to the discussion on P28: Referential integrity. The relevant requirement is shown in Figure 9.21.

DQ Problem-9: Failing to standardize the values for the title field which creates communication and identification problems.

The problem emerged with the discussion on P26: Value consistency. The participants revealed that they used the value “State off” as the title of deceased people whereas some staff members have used different acronyms leading to confusions. Further representation of the transgender customers is too unhappy in this sense due to the usage of many acronyms. Figure 9.22 shows this DQ requirement.

Finally the DQ pattern P11: Data timeliness was taken to the discussion forum and the participants revealed a DQ requirement which will arise in near future pertaining to data timeliness as follows.

DQ requirement: Timeliness of weather data is important to provide a proactive service to customers. The organization has planned to use weather data to identify disasters like floods in advance and raise the insurance claims proactively for the customer and start customer service function even before the disaster so that the customers will be at ease. Therefore, the DQ requirement was modelled using the pattern P11: Data timeliness as in Figure 9.23.

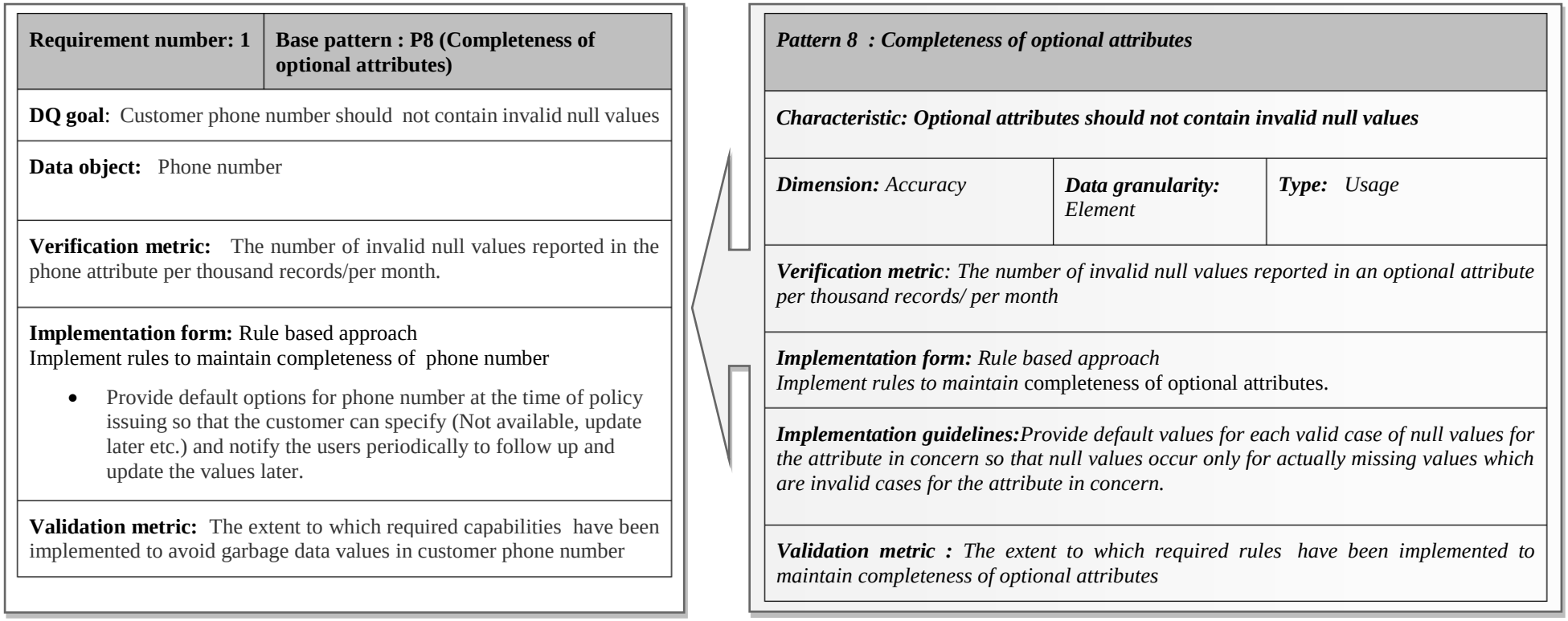


Figure 9.14: Requirement number 1 - XYZ insurance

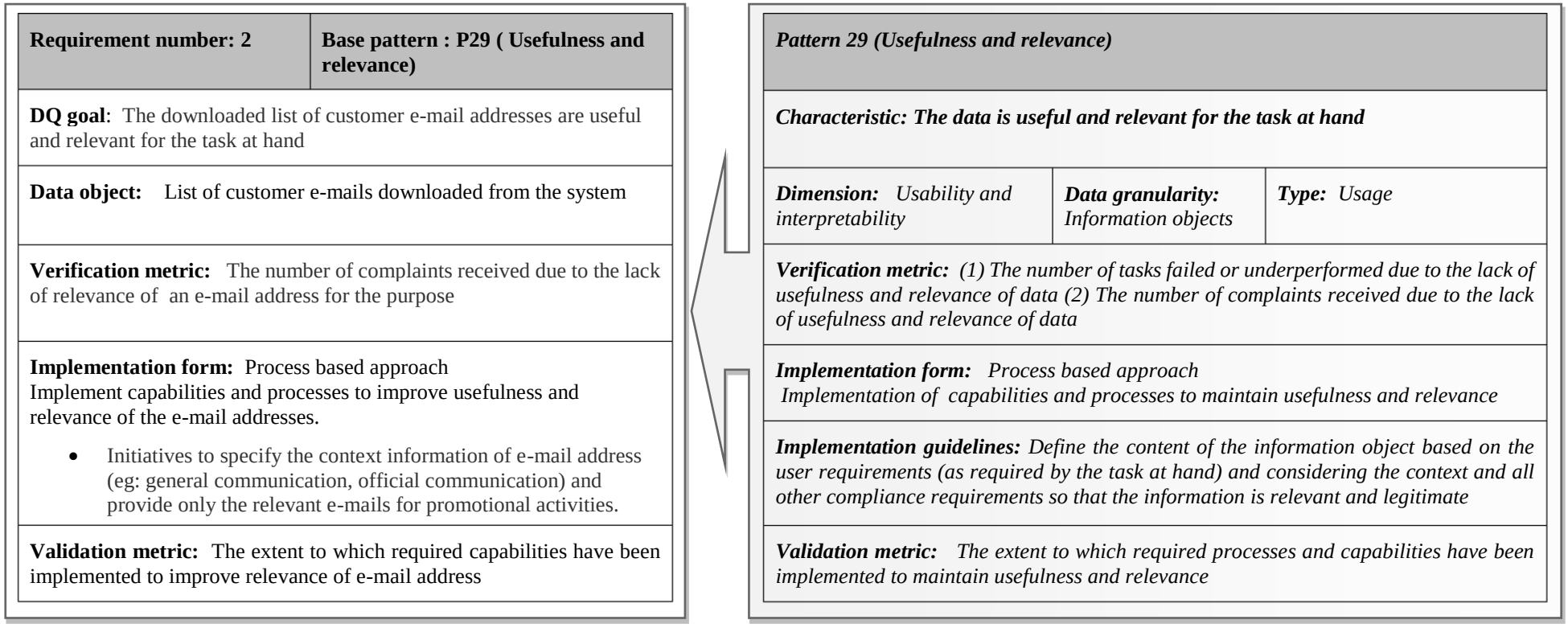


Figure 9.15: Requirement number 2 - XYZ insurance

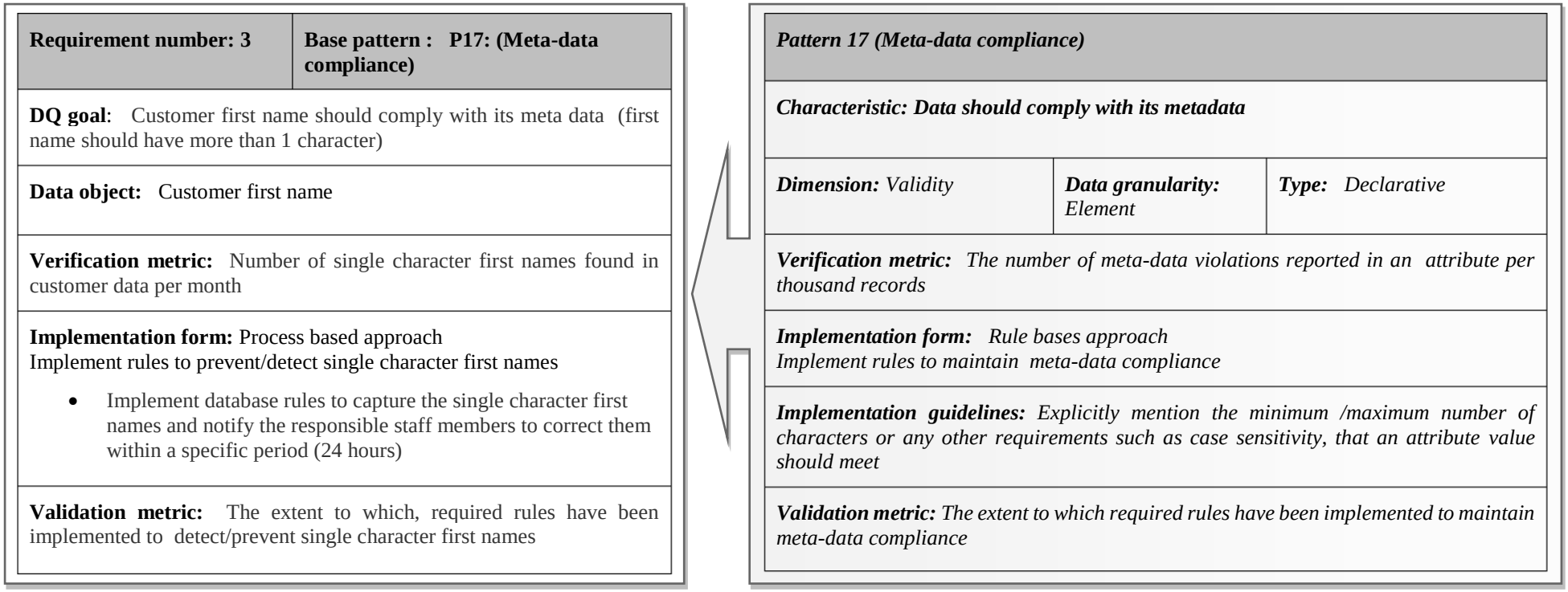


Figure 9.16: Requirement number 3 - XYZ insurance

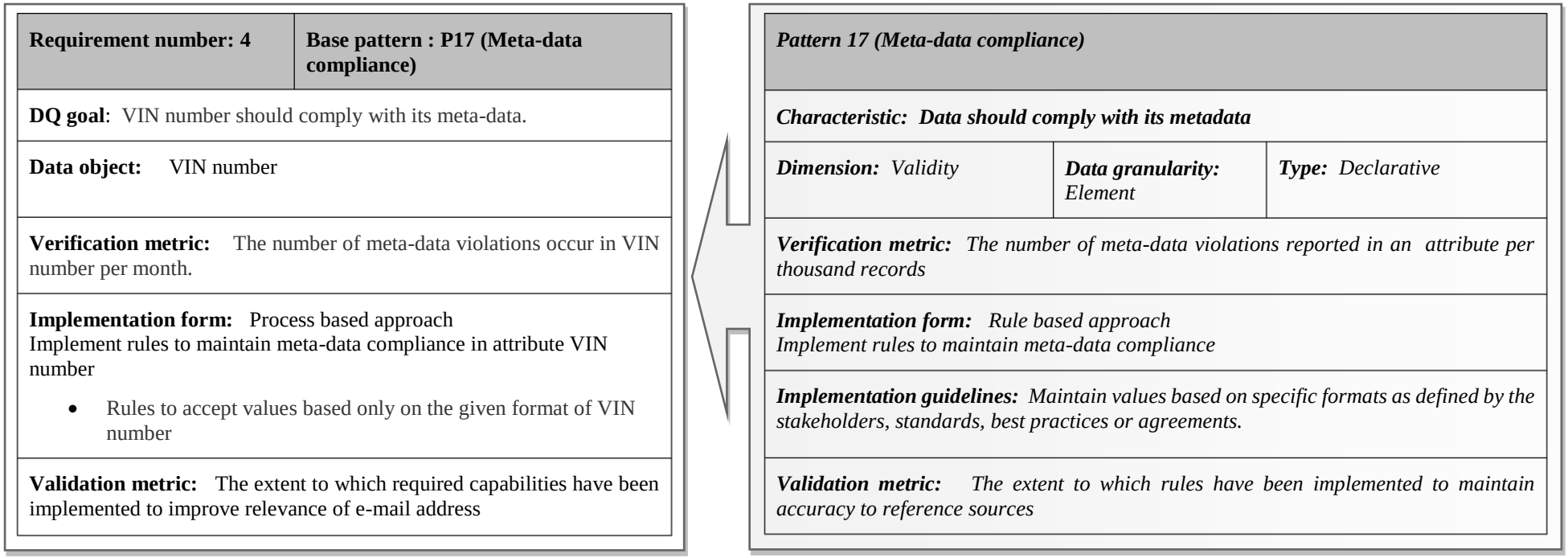


Figure 9.17: Requirement number 4 - XYZ insurance

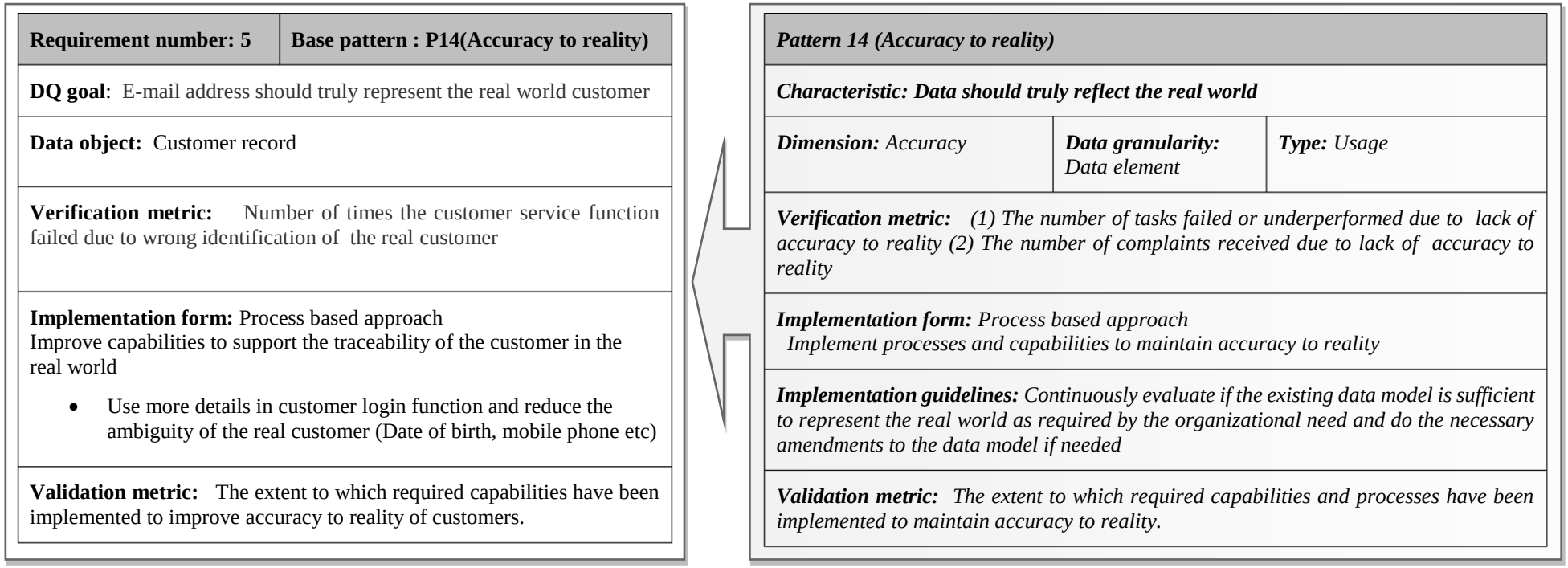


Figure 9.18: Requirement number 5 - XYZ insurance

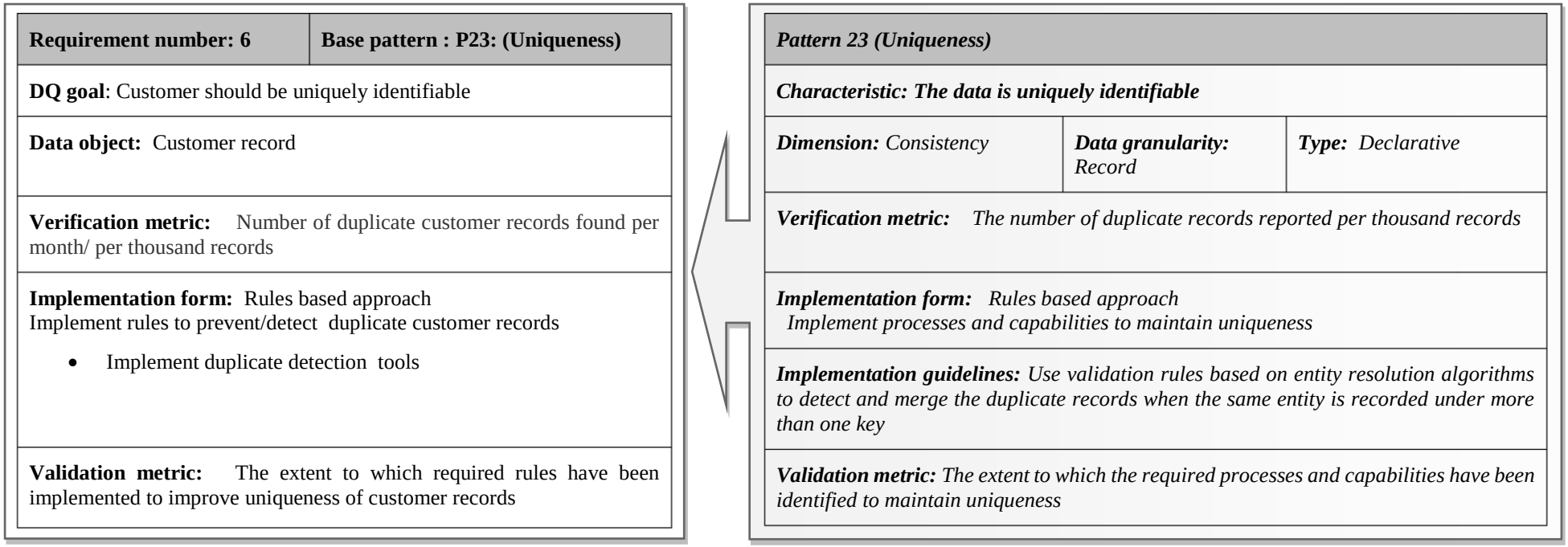


Figure 9.19: Requirement number 6 - XYZ insurance

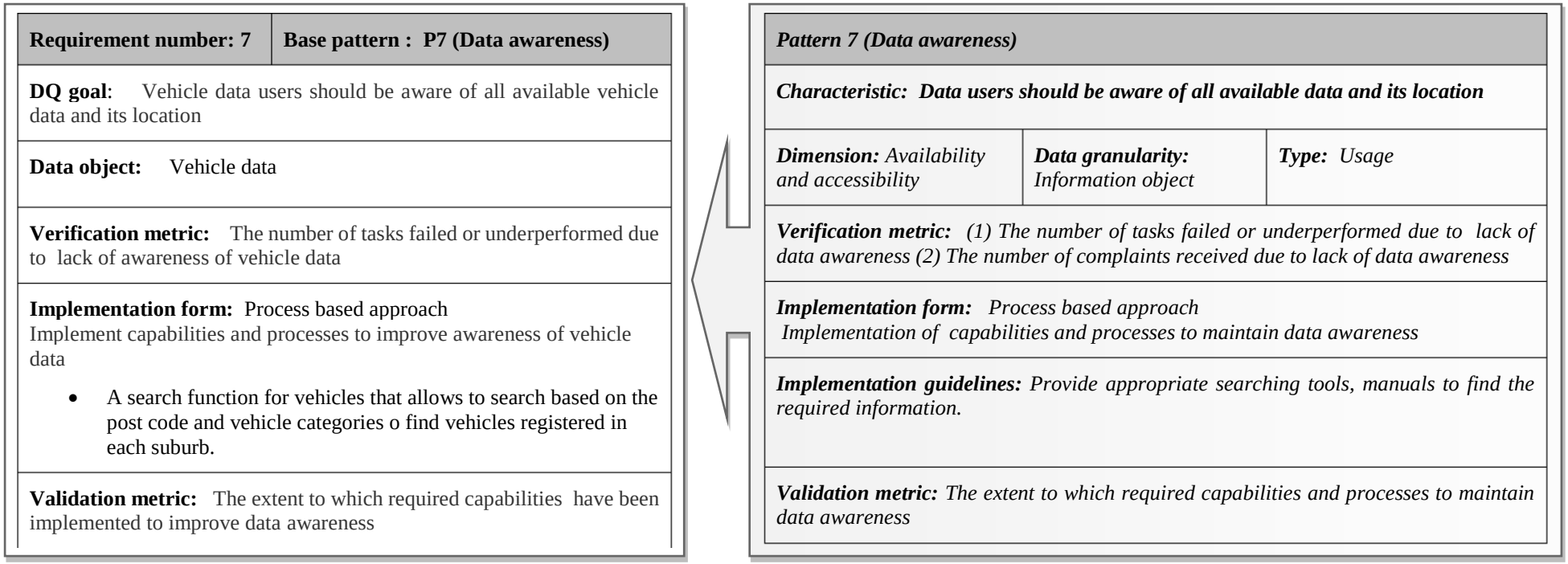


Figure 9.20: Requirement number 7 - XYZ insurance

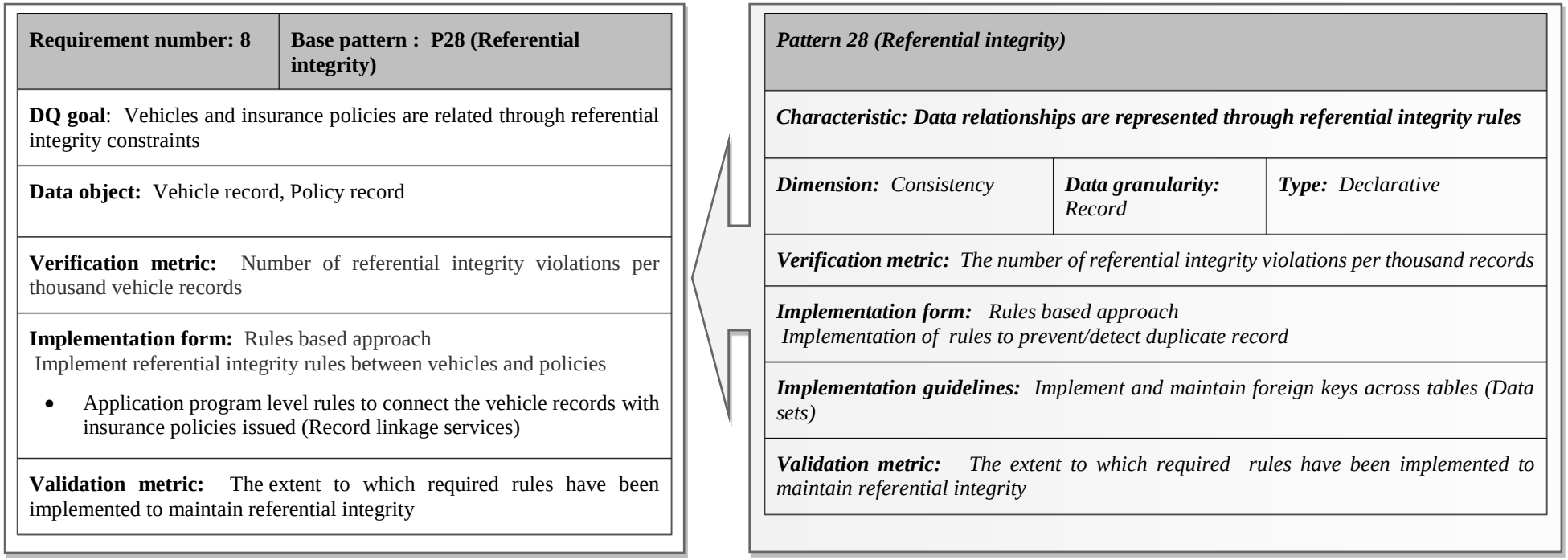


Figure 9.21: Requirement number 9 - XYZ insurance

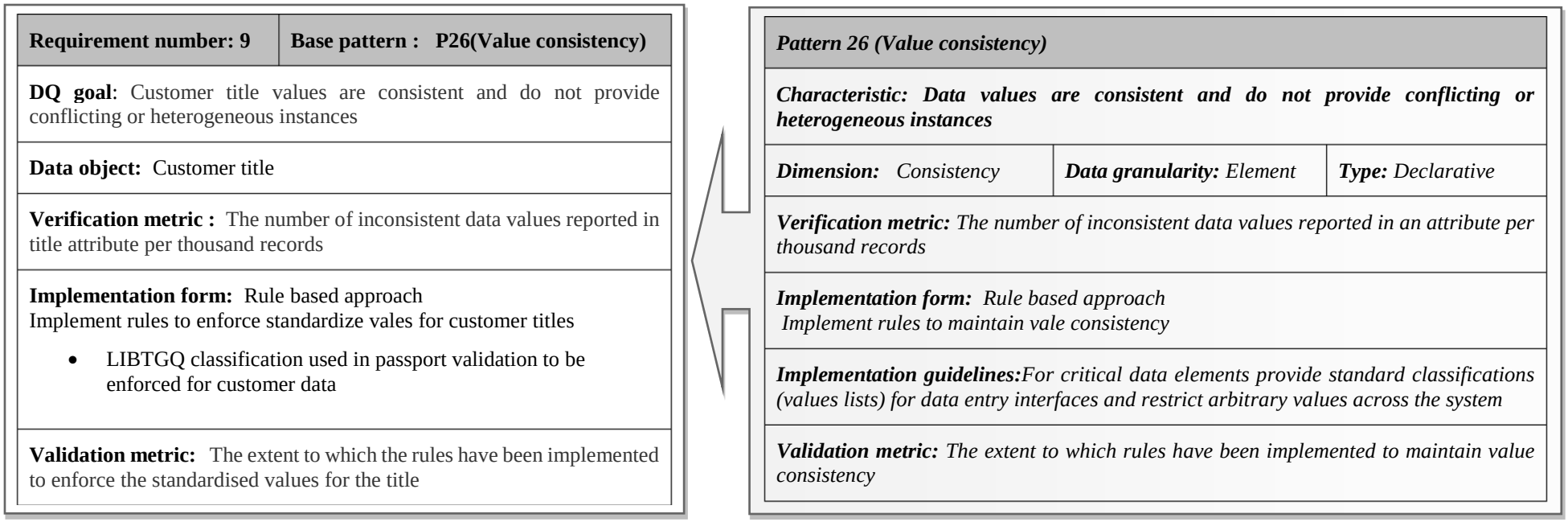


Figure 9.22: Requirement number 9 - XYZ insurance

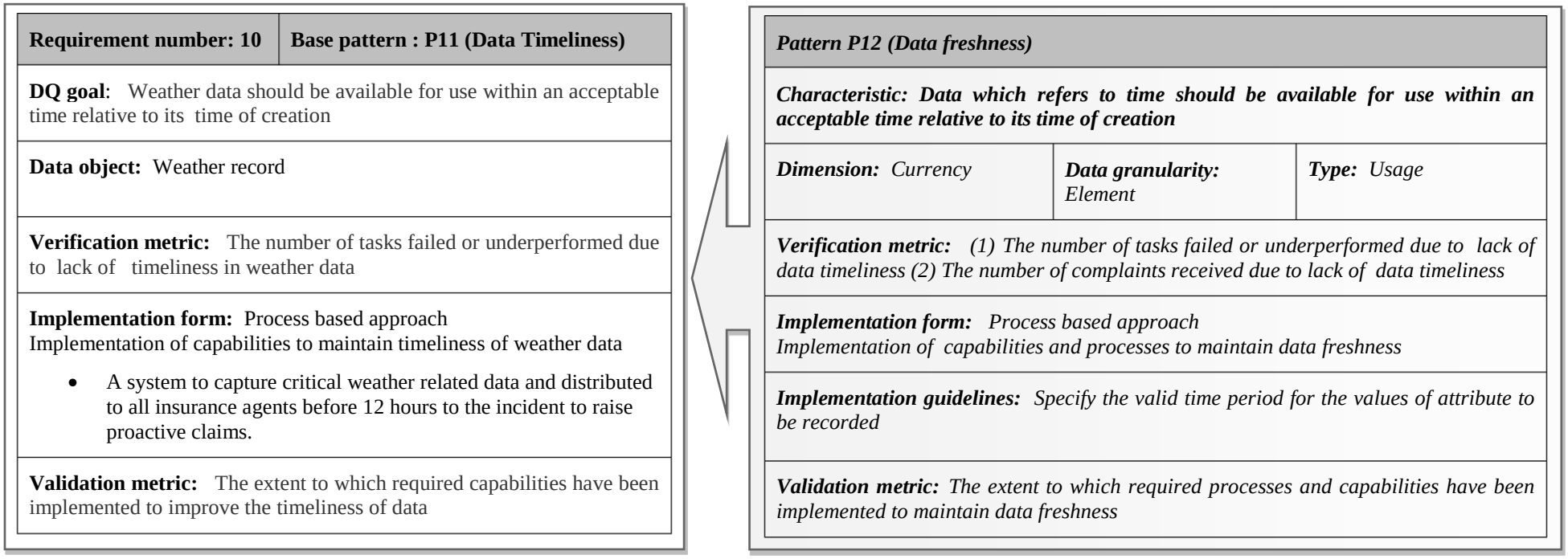


Figure 9.23: Requirement number 10 - XYZ insurance

9.6.1 Phase-3 results

In this phase, the DQ manager answered the online survey with regards to the DQ requirements derived above. The summary of the results is mentioned in Table 9.2 as follows.

DQ problem	Applicable pattern	Applicable similar DQ scenarios in the organization	Validation status
Problem-1	P2 :Completeness of optional attributes	Null VIN numbers in Quotations	Agree
Problem-2	P29: Usefulness and relevance	Use of mobile phone numbers for promotional activities	Agree
Problem-3	P17: Meta-data compliance	Inclusion of state-based area codes	Agree
Problem-4	P13: Meta-data compliance	Entering postcode as street name in address records	Agree
Problem-5	P14: Accuracy to reality		Agree
Problem-6	P23: Uniqueness	Duplicate addresses with slight syntax errors	Agree
Problem-7	P7: Data awareness	Party data of insurance claims (some are in legacy systems)	Agree
Problem-8	P28: Referential integrity	Houses transferred between owners	Agree
Problem-9	P26: Value consistency	Vehicle type classification (already in place)	Agree
Requirement	P11: Data Timeliness	Vehicle location data to guide on hazards	Agree

Table 9.2: Phase-3 XYZ insurance

It should be noted that the DQ manager of XYZ insurance agreed to all ten DQ requirements that we modelled pertaining to the context. Further, she managed to provide similar example scenarios to nine of the ten requirements where the ten DQ patterns in concern could be applicable. It should be noted that she has explicitly given the data object for seven of the examples and therefore it was quite clear that the corresponding patterns can be applied for these seven instances accurately. Thus we can be confident that she has agreed to the requirements with a certain understanding of the patterns pointing to the validity of the responses.

9.7 Discussion

In the above two applicability checks, we used the DQ patterns in KAOSE4DQ methodology intuitively to identify the DQ requirements in the two participant organizations. In this section, we discuss the evidence on the importance, accessibility and the suitability of the DQ patterns for DQ requirements analysis and modelling in our modelling experience in the above two studies.

9.7.1 Importance: How useful DQ patterns are in analysing and modelling DQ requirements?

In these applicability checks, we used the knowledge formalised in DQ patterns throughout the requirement elicitation process (Figure 9.24). The DQ patterns were used to,

(1) Analyse the DQ context in phase 1 where the DQ characteristics provided in patterns were helpful to understand the critical DQ characteristics in the studied organizational data.
(2) Identify the DQ problems in phase 2 where the DQ characteristics provided in patterns were helpful to focus on DQ problems pertaining to the data set in concern.
(3) Model DQ requirements in phase 2 where the DQ patterns were helpful in defining DQ goals, verification metrics, validation metrics and the implementation form which is the real requirement that should be implemented to maintain DQ.

In phase-2, we observed that in ABC hospital, the participants were comparatively less interactive with the interviewers. Therefore, the DQ characteristics taken into the discussion forum provided them with a substance to focus on and thereby raise their voices regarding their own experiences of violation of the DQ characteristics in concern. Therefore, the discussion was driven smoothly without any bottlenecks and necessary information was elicited systematically. In XYZ insurance, the participants seemed to have been more interactive with each other and occasionally the discussion went beyond the focus of DQ (current business initiatives etc.). In such situations, the DQ characteristic helped the moderator to divert the focus back to the discussion by pointing to the DQ characteristics in the top ten list of DQ characteristics. The interviewers made an effort to keep the focus of the interview towards DQ and ultimately the necessary information was elicited satisfactorily. Therefore, in both cases, we could observe that the DQ characteristics helped us to drive the discussion more effectively and efficiently.

Towards the end of each interview, we asked the participants if they can think of any DQ requirement pertaining to any of the 33 DQ patterns. In response, in XYZ insurance, the participants revealed a future DQ requirements relevant to the DQ pattern P11: Data timeliness. Therefore, it is clear that DQ patterns can be used to derive DQ requirements directly as pointed out in section 8.4.2.3.

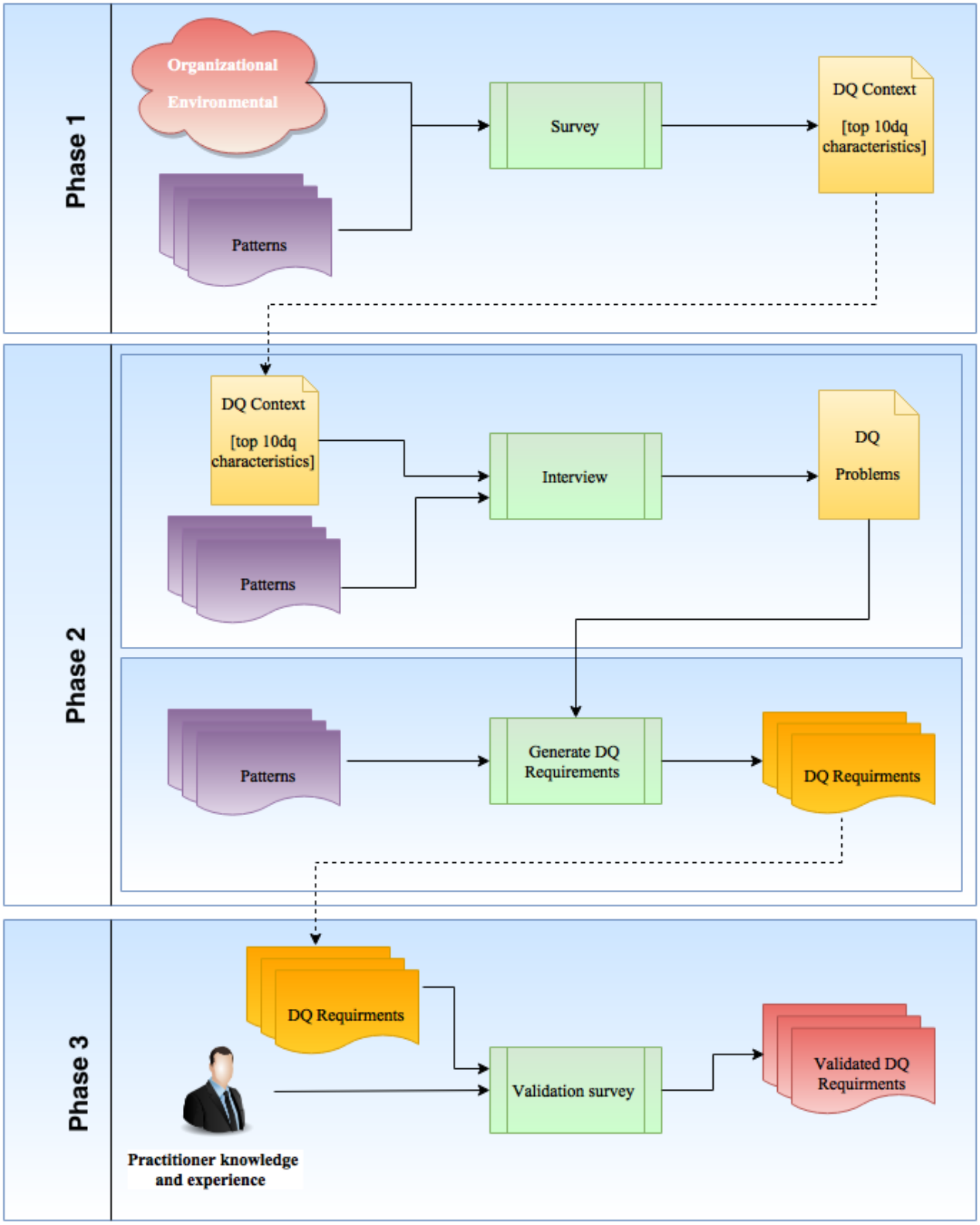


Figure 9.24: Use of DQ patterns in requirements elicitation process

We also observed that the participants were able to relate to DQ problems more comfortably than explicitly talking about the DQ requirements.

In the discussion of ABC hospital, it was observed that the interviewers could use the relationship between DQ patterns (Table 7.32) to elicit more DQ problems. For example, After *P8*: *Ease of data access* the characteristic *P9*: *Data punctuality* was introduced to the discussion and a relevant DQ problem was identified successfully. Further *P31*: *Appropriate presentation* was introduced after *P29*: *Usefulness and relevance* and a relevant DQ problem was again revealed successfully. In both cases, the participants elaborated on the discussion with ease to expose the DQ problems.

Once the DQ problems and related information were collected, the corresponding DQ requirements were modeled by instantiating the corresponding DQ patterns. This was a very straightforward exercise since each pattern provided the required knowledge in defining a DQ goal, a verification metrics, the implementation form and a validation metric by referring to the generic formalization provided in the relevant pattern for all the DQ problems found in phase-2.

During the discussion, having asked about the verification metric for each DQ requirement, in both organizations the participants could not provide a verification metric themselves, but they agreed on the verification metrics that we proposed based on the DQ patterns. Therefore, it was evident that the generic formalization provided by DQ patterns was applicable and assisted in defining the verification metrics effectively.

The participants were asked about the probable solutions for DQ problems, in order to define the implementation form. In both organizations, most of the DQ problems discussed had some kind of a preventive measure taken so far. Being award-winning organizations for DQ, the two organizations had successfully managed the quality of data up to a great extent. Therefore, the participants revealed the DQ processes and rules implemented to resolve most of the DQ problems. The interviewers used the knowledge provided by the implementation guidelines in each pattern to moderate the discussion with regards to the solution, while most of the solution taken so far were in line with the guidelines.

Validation metrics seemed to be a new concept to the participants in both organizations and therefore the responses were limited. Therefore, we proposed validation metrics by ourselves based on the generic definition of the validation metric in each pattern and the participants had no disagreement with them. Although it was somewhat doubtful if the proposed validation metrics were authentically approved by the participants, but further in phase-3, the DQ managers agreed to all the validation

metrics that we had proposed, providing us with some confidence that they can be used to measure the progress of DQ solutions.

Based on the above facts, when considering the overall experience in modelling the DQ requirements in both organizations we can say that the DQ patterns were useful in analysing the DQ requirements and also modelling the DQ requirements. Thus we argue that the patterns played an important role in the process of analysing and modelling of DQ requirements.

9.7.2 Accessibility: How well the data quality users connect with the artefact

We observed that the top-ten DQ characteristics revealed by both managers were credible since in the phase-2 discussion they seemed to have guided the team members to reveal DQ problems by nominating the most appropriate participant to expose the information. Therefore, we can say that both managers had understood the definitions and terminology in answering phase-1 questions. Further in phase-3, both of them demonstrated that they understood the relevant patterns by accurately providing the similar scenarios that can be modelled using the same pattern.

At the beginning of both interviews, we introduced the DQ patterns to the participants and the participants had no questions about the definitions of the characteristics. They were well connected with the definitions and seemed to have a copy of the material we provided prior to the interview as a study guide for the patterns. In the case of XYZ insurance, we observed that the DQ team has their own definitions for DQ dimensions and they used to refer to these definitions during the discussion. For example "Failing to capture any useful data is an accuracy problem". This was quite conflicting with all the three of our patterns under the DQ dimension Accuracy (Pattern 13-Accuracy to reference sources: Data should agree with an identified source; Pattern 14- Accuracy to reality: Data should truly reflect the real world, Pattern-15-Precision: Attribute values should be accurate as per linguistics and granularity). The definition they provided was closer to a completeness problem and it seems the participants were strictly held on to their definitions. In this situation, the eight main DQ dimensions used in our classification was quite helpful to connect with them since the terms used for DQ dimensions were quite familiar to the DQ professionals (eg: completeness, Accuracy, Currency etc.). With the support of this common terminology, we were able to effectively convey our definitions of DQ characteristics by pointing out the closest DQ dimension. Therefore, the floor of discussion was never obstructed by such instances and it was evident that high-level DQ dimensions used in our classification were quite useful in terms of using them in practical circumstances.

In the DQ literature, the practitioners have pointed out that, the concept of DQ dimensions does not reach the stakeholders of data due to terminology conflicts (Johns 2010; Sebastian-Coleman 2012).

But in our case, we were able to use our DQ dimensions and characteristics successfully in an environment with conflicting terminology. It should be noted that we have validated the clarity of the definitions and terms in our classification using a rigorous methodology which was based on cognitive science (see section 5.3.3). Thus we argue that the definitions and the terminology used in our classification are simple and easily understood by the stakeholders of DQ.

9.7.3 Suitability: How accurately the DQ requirement was elicited and modelled

In the phase-3 survey, we observed that both DQ managers have agreed with the DQ requirements that we identified and modelled. This provides direct evidence on the accuracy of the DQ requirement model developed using the DQ patterns. Hence we argue that the patterns were suitable for its intended task since they yielded accurate results.

It should be noted that, within a period of one hour, we were able to elicit ten DQ requirements in each case along with sufficient information to formulate their implementation form and metrics. Phase-3 provided evidence that the requirements identified were accurate from the viewpoint of organizational context. Also, it should be noted that the interviewers had no prior exposure to the organizational contexts and neither were they well-experienced business analysts (or DQ analysts). On these grounds, elicitation of ten full DQ requirements (including their implementation form and metrics) within a period of one hour in each organization, provides evidence on the expressive power of DQ patterns in modelling DQ requirements effectively and efficiently in organizations.

9.7.4 Summary

In this chapter, we reported on an empirical study performed to analyse and model DQ requirements in two organizations based on the DQ patterns we developed. The study was designed based on the notions used in applicability checks (Rosemann and Vessey 2008) and also using KAOS4DQ methodology. The study was conducted in two main phases as described in KAOS4DQ methodology viz. Phase-1: Top-down analysis and Phase-2: Bottom-up analysis. We performed an additional phase which is phase-3 to validate the DQ requirements that we modelled. In both organizations, we successfully performed a DQ requirement analysis and modelling task. Our study provided evidence that the patterns are useful in analysis and modelling of the DQ requirements, the formalization of knowledge in the patterns is understood by the practitioner and user community and finally the patterns provide accurate requirements models for the organization.

.

Chapter 10

CONCLUSION

10.1 Overview

This chapter summarises the major contributions of the study, its limitations, and an overview of potential future work.

10.2 Summary of contribution

The main contribution of this study is the development of a repository of DQ patterns that have demonstrably capable of conducting effective and efficient requirements analysis and modelling - a largely under-studied topic in DQ management. The pattern repository is the result of a systematic process of conceptualising DQ requirements. This study contributes to the body of DQ management by first developing a meta-meta-model to represent a DQ requirement, and second, by developing a comprehensive classification of DQ dimensions by consolidating the existing DQ dimensions defined by both academic and practitioner communities and refactoring them as per the meta-meta-model for a DQ requirement.

Each DQ pattern is an instance of the meta-meta model and in turn serves as a meta-model for specific DQ requirements. Further, to analyse DQ requirements we adapted a requirements engineering methodology. The DQ patterns can be used intuitively in the adapted methodology in analysing and modelling DQ requirements. Through the use of applicability checks, we provide evidence that the DQ patterns can be used to develop DQ requirement models in organizations in an accurate and efficient way. Thus, this research contributes to strengthening DQ requirements analysis and modelling. Below we outline the contributions in detail.

10.2.1 Conceptualization of a DQ requirement

In literature review we revealed that the existing meta-model for DQ requirements are deficient in covering the required domain concepts to represent a DQ requirement. To the best of our knowledge, there is no conceptual formulation for a DQ requirement in the literature, due to the lack of focus paid to DQ requirements in academia. This study conceptualises a DQ requirement by developing a comprehensive meta-meta-model for a DQ requirement referring to the literature on conceptual modelling, as well as global standards IEEE and ISO. In our definition for a DQ requirement, we considered two perspectives: the DQ problem perspective and DQ solution perspective. These two perspectives provide a holistic view of data quality in terms of what to manage and how to manage it. We used five concepts to describe the DQ problem perspective (DQ characteristic, DQ

characteristic type, DQ dimension, verification metric, and verification metric threshold) and three main concepts to represent the DQ solution perspective (implementation form, validation metric, and validation metric threshold). We defined DQ characteristics as the central concept of a DQ requirement that can be used as the starting point of analysing data users' requirements. The two types of metrics are a means of systematically evaluating how successfully the DQ requirement is managed in the organization. Thus, our definition of a DQ requirement is a comprehensive one that provides a holistic view toward understanding DQ problems, developing DQ solutions, and monitoring and maintaining the quality of data systematically.

10.3 Refactoring DQ dimensions

Because a shared understanding is necessary for developing conceptual models, the lack of consensus about the key concept DQ dimension has been a barrier in the journey towards a repository of DQ patterns. In studying the existing classifications of DQ dimensions we noticed that the lack of focus in defining these dimensions was the main issue. Thus, with the plethora of available definitions in the literature, the question arises: what are the boundaries of defining a DQ dimension? By referring to the fundamental concepts used in the quality of products and services, we developed two perspectives (declarative and usage) to study the definitions of DQ dimensions. This conceptual lens helped us to analyse the definitions in a more systematic and focused way to produce a classification with a solid theoretical grounding.

In this study, we developed a two-level classification for DQ dimensions. In the first level we created eight main DQ dimensions (Completeness, Availability and accessibility, Currency, Accuracy, Validity, Reliability and credibility, Consistency and, Usefulness and relevance) and in the second level, we created 33 DQ characteristics such that every DQ dimension contains several DQ characteristics. The eight main DQ dimensions refer to the common vocabulary used in the DQ domain and thus help to organize the 33 DQ characteristics in a systematic manner that supports easy accessibility.

A salient feature in our DQ characteristics is their type (*declarative or usage*). The DQ problems pertaining to *declarative characteristics* can be detected and prevented using a rules-based approach, thus we argue that they are independent of data users. In contrast, the DQ problems pertaining to *usage characteristics* can be detected and prevented only with the human involvement in dealing with data, thus we call them user dependent characteristics. Out of the 33 characteristics, only 10 are declarative whereas 23 are of the usage type. This reveals that a major proportion of the DQ management portfolio should be human-centric and only a lesser proportion can be handled using

automated DQ problem detection and prevention mechanisms. We argue that this insight will be helpful in designing DQ management strategies.

(Batini et al. 2009) revealed that most DQ dimensions defined in the DQ literature lack metrics. This study proposes that every DQ characteristic has two types of associated metrics: verification metrics and validation metrics. The verification metrics measure the current quality of data related to a DQ characteristic, while the validation metric measures the maturity of a DQ solution taken to maintain the DQ characteristic. While the validation metric is new to existing DQ management contexts, we argue that a combination of verification and validation metrics enables DQ to be managed more systematically compared to most of the single metric based DQ management systems. Further, our definitions for verification and validation metrics are based on simple criteria that can be implemented easily in practical situations. Therefore in this thesis, we present a comprehensive classification of DQ characteristics with metrics that can be practically used in managing data quality. Hence, this classification contributes to both theory and practice in DQ management.

10.3.1 Goal oriented DQ requirements modelling

Requirements analysis is a fundamental task in modelling requirements. Owing to the lack of a proper methodology to analyse DQ requirements, we studied the requirements engineering literature with regard to fundamental notions of information systems development. In the recent years, the popularity of goal-oriented requirements engineering approaches has increased quite significantly. The main reason for this is the inadequacy of the traditional systems analysis approaches to deal with more complex organizational environments. DQ dimensional analysis, as mentioned in Chapter 5, revealed that the majority of DQ characteristics are of the usage type, thus the DQ requirements span across various data user needs and are complicated to analyse.

After evaluating prominent goal oriented approaches in the literature, we selected KAOS (Van Lamsweerde 2001) as a suitable one for DQ requirements analysis. With the necessary modifications, KAOS was adapted to the DQ domain. As a result of this study, to the best of our knowledge, KAOS4DQ emerged as the first goal oriented DQ requirements analysis methodology that was proven as capable in analysing DQ requirements in real world organizations.

10.3.2 DQ patterns

The final artefact produced in this thesis is the repository of the DQ patterns to model DQ requirements. Thirty-three DQ patterns were developed by instantiating the meta-meta-model using the classification of refactored DQ dimension. We argue that DQ patterns can be used as constructs

of a meta-model to model real world DQ requirements. DQ patterns are rich in their coverage of domain concepts due to the representational capabilities of their meta-meta-model, and the comprehensive classification of DQ dimensions used in instantiating the meta-meta-model.

Every DQ pattern provides a generic quality requirement for data applicable at a particular data granularity level. The DQ characteristic used in the pattern, on the one hand, represents a DQ user requirement and on the other hand, it represents a probable DQ problem that can occur when the DQ characteristic is not maintained. Thus, the patterns can be used in KAOS4DQ to find the DQ problems as well as defining DQ user requirements.

Further, a pattern provides implementation guidelines for designing a proactive solution to maintain the DQ characteristic of concern and prevent DQ problems. We argue that each DQ characteristic can be maintained either using the rule-based approach (where DQ rules are implemented in the database or application programs) or using the process-based approach (where processes and capabilities are established to improve data usage). Owing to the popular assumption that prevention is better than cure, we argue that DQ problems have to be foreseen and prevented rather than being fixed afterwards.

In light of the management axiom "what gets measured gets managed" (Willcocks and Lester 1996), measurements are an important aspect of management. Therefore, every DQ pattern has measurement criteria attached to it as given by the meta-meta-model. We argue that the DQ patterns support building a comprehensive solution for DQ, starting from problem analysis, solution design and then monitoring and measuring of the data quality. To facilitate the use of the knowledge provided by the patterns, we organised them into a web-portal to provide a publically accessible resource for the viewing and use of the DQ patterns.

We evaluated the sufficiency of the DQ patterns in representing the DQ requirements by performing a descriptive evaluation. The evaluation was performed based on two perspectives: (1) a DQ problem perspective and (2) DQ user requirement perspective since a DQ user requirement and a DQ problem can be seen as the two complementary perspectives of any DQ characteristic. The validation revealed that there is a corresponding pattern to represent each and every DQ requirement and problem considered. In other words, there was no DQ rule or problem that cannot be represented by patterns. Thus we conclude, that the DQ patterns are sufficient in representing the published DQ problems and user requirements.

While the descriptive evaluation proved the representational sufficiency of the patterns, the next concern was whether the pattern based approach is applicable in real world organizations in terms of

analysing and modelling DQ requirements. Rosemann and Vessey (2008) proposed the approach known as applicability checks, where industry experts' judgement is used to evaluate the relevance of research artefacts to practice. Therefore, we posit that performing a requirement analysis and modelling task with a group of industry experts using the DQ patterns and KAOS4DQ methodology will provide necessary evidence about the applicability of the DQ patterns.

We used the DQ patterns successfully and modelled a set of DQ requirements pertaining to a critical dataset for two organization. In this applicability check, we evaluated three perspectives: importance, accessibility, and suitability (Rosemann and Vessey 2008) of the artefact in analysing and modelling DQ requirements. We observed that the DQ patterns were important in analysing the DQ requirements in the organizations effectively and produce the required DQ models. Further, the practitioners considered that it was feasible to use the patterns in DQ requirement analysis and modelling. The validation performed on DQ requirements ensured that the requirement models are accurate and legitimate for these organizations' contexts. Thus, this validation further provides evidence that the patterns can be used in real-world organizational settings.

10.4 Research Limitations

Our work is not without limitations. In each DQ pattern, we have presented some guidelines for the design of its implementation. Even though providing such guidelines was not a part of the scope of our study, we suggest they will be useful in the process of DQ requirements analysis and modelling. These guidelines were developed mainly based on the insight gained from the literature by DQ practitioners (English 2009; Kimball and Caserta 2004; Loshin 2004; Loshin 2011; McGilvray 2008; Redman 2008; Redman 1997). Further we used the published DQ management frameworks by the Canadian Institute for Health Information (CIHI) (Long and Seko 2005), the Health Information and Quality Authority in Ireland (HIQA 2011) and referred to some of the success stories of award-winning organizations for DQ recognised by the International Association for Data and Information Quality (IAIDQ 2015).

We admit that more work is needed in terms of defining and refining these guidelines. Building comprehensive guidelines for thirty-three DQ patterns is an extensive process and the ideal approach is to construct them is through multiple case studies (Yin 2013) of organizations with good DQ management track records. Such an empirically driven set of guidelines would be highly useful for establishing the DQ patterns as a consulting base for DQ management.

In the descriptive evaluation of DQ patterns, we used 213 DQ problems from six different classifications defined in the literature over the past three decades and 197 DQ requirements from

three published rules repositories. Although the six classifications of DQ problems are a sufficient representation, we admit that three rules repositories are not an adequate representation of real world DQ requirements. Most contacted organizations could not share their DQ rules and processes with us due to organizational policies and this has imposed limitations on our study.

In our empirical evaluation, we checked the applicability of the pattern based approach (in KAOSE4DQ). While the study revealed that the DQ patterns are applicable in modelling the DQ requirements and producing accurate DQ requirement models, our study was not sufficient to prove that the pattern based approach is more efficient compared to other traditional DQ analysis methods. We believe that an experiment can be designed where a group of participants analyses and model DQ requirements using the pattern-based approach, while another group performs a control experiment without this approach. Such a study would help to further improve the KAOS4DQ methodology.

10.5 Future work

While the pattern based approach is ready to be applied in practice there are several interesting future studies that can also contribute to further this research.

10.5.1 Relationships between patterns

In the descriptive evaluation in Chapter 7, we presented thirteen observations of relationships between patterns. We further observed the importance of the relationship among patterns in our applicability check. Therefore, it is important to establish all such relationships between the patterns, to support the requirement elicitation process.

In the literature, some researchers have emphasized the relationships between DQ dimensions. Panahy et al. (2013) reveal that the relationship between DQ dimensions has not been studied well in literature. De Amicis et al. (2006) argue that dependencies among the DQ dimensions are essential for improving process quality and promoting effective and comprehensive knowledge discovery. Among limited attempts, De Amicis et al. (2006) propose a data quality dependency model where DQ dimensions can be divided into three major categories: perfect dependency, partial dependency, and independency. Further, trade-off analysis (Gackowski 2005; Madnick et al. 2009; Scannapieco et al. 2005) and logical interdependence analysis (Gackowski 2005) are examples of dependency analysis among different dimensions. Therefore, we posit that a dependency analysis among DQ patterns will help further support the effectiveness of the DQ requirement analysis process.

10.5.2 Tool support for DQ requirements analysis and modelling

Requirements modelling is the starting point of the DQ management process and the requirements gathered at this stage have to be effectively and efficiently managed so that they can be used in designing and implementing solutions in the subsequent management process. In practice, the number of DQ requirements in an organization can be very large and, therefore, it is necessary to have a mechanism to organize this large collection of DQ requirement models. Hence, it is desirable to develop a software tool to support the knowledge management aspect of each DQ requirement modelled in the pattern based approach. This is another focus of our research agenda in future.

In conceptualising a DQ requirement, we used entity relationship grammar to develop the meta-meta-model for a DQ requirement. Rosemann and Green (2002) revealed that when existing modelling grammars are used in the process of meta-modelling, the properties of the resultant meta-models and models will have more or less the same flavour as the modelling grammar. Thus, we intentionally selected ER grammar so that the DQ patterns have the flavour of a logical database to store the DQ requirements. In this case, the organizational DQ requirements can be presented in a database. In modern database systems, the data catalogue is a tool that provides descriptions about each and every data object used in the organization. We argue that, once DQ requirements are inserted into a database they can be viewed through data catalogues by connecting the DQ requirements database with a data catalogue. Developing a tool to model such DQ requirements so that they can be referred to through a data catalogue is a practically valuable extension of our current work.

10.6 Summary

This study contributes to the body of knowledge of DQ management by developing a meta-meta-model for a DQ requirement, a comprehensive classification of DQ dimensions, a repository of DQ patterns and a goal-oriented methodology for DQ requirements analysis. The pattern-based approach is applicable in practice to develop requirement models in organizations.

REFERENCES

Al-Hakim, L. 2007. *Information Quality Management: Theory and Applications*. IGI Global.

Albers, M. J. 1998. "Goal-Driven Task Analysis: Improving Situation Awareness for Complex Problem-Solving," *Proceedings of the 16th annual international conference on Computer documentation*: ACM, pp. 234-242.

Alexander, C. 1979. *The Timeless Way of Building*. New York: Oxford University Press.

Alsaghier, H., Ford, M., Nguyen, A., and Hexel, R. 2011. "Conceptualising Citizen's Trust in E-Government: Application of Q Methodology," *Leading Issues in E-Government* (1), p. 204.

Auerbach, C., and Silverstein, L. B. 2003. *Qualitative Data: An Introduction to Coding and Analysis*. NYU press.

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 and Applications* (1:3), pp. 205-218.

Batini, C., Francalanci, C., Cappiello, C., and Maurino, A. 2009. "Methodologies for Data Quality Assessment and Improvement," *ACM computing surveys* (41:3), pp. 1 - 52.

Batini, C., and Scannapieco, M. 2006. *Data Quality: Concepts, Methodologies and Techniques*. Springer.

Baum, F. 2003. *The New Public Health*. Oxford University Press.

Becker, D., McMullen, W., and Hetherington-Young, K. 2007. "A Flexible and Generic Data Quality Metamodel," *International Conference on Information Quality*.

Booch, G., Jacobson, I., and Rumbaugh, J. 2000. "Omg Unified Modeling Language Specification," *Object Management Group ed: Object Management Group*), p. 1034.

Bourque, P., Dupuis, R., Abran, A., Moore, J. W., and Tripp, L. 1999. "The Guide to the Software Engineering Body of Knowledge," *IEEE software* (16:6), p. 35.

Braun, V., and Clarke, V. 2006. "Using Thematic Analysis in Psychology," *Qualitative research in psychology* (3:2), pp. 77-101.

Bresciani, P., Perini, A., Giorgini, P., Giunchiglia, F., and Mylopoulos, J. 2004. "Tropos: An Agent-Oriented Software Development Methodology," *Autonomous Agents and Multi-Agent Systems* (8:3), pp. 203-236.

Brown, B. B. 1968. "Delphi Process: A Methodology Used for the Elicitation of Opinions of Experts," DTIC Document.

Burton-Jones, A., Wand, Y., and Weber, R. 2009. "Guidelines for Empirical Evaluations of Conceptual Modeling Grammars," *Journal of the Association for Information Systems* (10:6), pp. 495-532.

Byrne, J. K., D. Mccarty, G. Sauter, H. Smith, P Worcester. 2008. "The Information Perspective of Soa Design Part 6:The Value of Applying the Data Quality Analysis Pattern in Soa," IBM corporation.

Carletta, J. 1996. "Assessing Agreement on Classification Tasks: The Kappa Statistic," *Computational linguistics* (22:2), pp. 249-254.

Caudron, J., and Peteghem, D. V. 2015. "Why Data Science Matters and How to Approach It for Your Digital Transformation." Retrieved 30.06.2016, 2016, from

http://www.digitaltransformationbook.com/why-data-science-matters-and-how-to-approach-it-for-your-digital-transformation/

Caro, A., Rodríguez, A., Cappiello, C., and Caballero, I. 2012. "Designing Business Processes Able to Satisfy Data Quality Requirements," in: *17th International Conference on Information Quality (ICIQ)*. Paris France.

Castillo, C., Donato, D., Becchetti, L., Boldi, P., Leonardi, S., Santini, M., and Vigna, S. 2006. "A Reference Collection for Web Spam," *ACM Sigir Forum*: ACM, pp. 11-24.

Chen, M., Mao, S., Zhang, Y., and Leung, V. C. 2014. "Big Data Applications," in *Big Data*. Springer, pp. 59-79.

Chen, P. P. S. 1976. "The Entity-Relationship Model—toward a Unified View of Data," *ACM Transactions on Database Systems (TODS)* (1:1), pp. 9-36.

Cheng, R., Kalashnikov, D. V., and Prabhakar, S. 2003. "Evaluating Probabilistic Queries over Imprecise Data," *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*: ACM, pp. 551-562.

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*: VLDB Endowment, pp. 315-326.

Couture, B. 1986. "Effective Ideation in Written Text: A Functional Approach to Clarity and Exigence," *Faculty Publications--Department of English*), p. 67.

Coyne, Imelda T. 1997. "Sampling in qualitative research. Purposeful and theoretical sampling; merging or clear boundaries?", *Journal of advanced nursing*, 26: 623-30.

Dasu, T., and Loh, J. M. 2012. "Statistical Distortion: Consequences of Data Cleaning," *Proceedings of the VLDB Endowment* (5:11), pp. 1674-1683.

De Amicis, F., Barone, D., and Batini, C. 2006. "An Analytical Framework to Analyze Dependencies among Data Quality Dimensions," *ICIQ*, pp. 369-383.

De Amicis, F., and Batini, C. 2004. "A Methodology for Data Quality Assessment on Financial Data," *Studies in Communication Sciences* (4:2), pp. 115-136.

Dictionaries, O. 2010. "Oxford Dictionaries," in: *Oxford dictionaries Language matters*. Oxford University Press, p. 2012.

Digital Accountability & Transparency Act. 2014. in: *H.R. 2061*, U.S. Congress. Congressional Budget Office.

DOD. 2014. "Us Department of Defence -Business Rules." Retrieved 15/4/2014, 2014, from http://search.defense.gov/search?affiliate=DEFENSE_gov&query=business+rules&x=0&y=0

Drucker, P. F. 1995. *People and Performance: The Best of Peter Drucker on Management*. Routledge.

Elmasri, R. 2008. *Fundamentals of Database Systems*. Pearson Education India.

Endsley, M. R., Bolstad, C. A., Jones, D. G., and Riley, J. M. 2003. "Situation Awareness Oriented Design: From User's Cognitive Requirements to Creating Effective Supporting Technologies," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*: SAGE Publications, pp. 268-272.

English, L. P. 1999. "Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits,").

English, L. P. 2009. *Information Quality Applied: Best Practices for Improving Business Information, Processes and Systems*. Wiley Publishing.

Eppler, M. J. 1999. "Qualitätsstandards—Ein Instrument Zur Sicherung Der Informationsqualität in Multimedia-Produktionen," in *Qualitätssicherung Bei Multimedia-Projekten*. Springer, pp. 129-149.

Eppler, M. J. 2006. *Managing Information Quality: Increasing the Value of Information in Knowledge-Intensive Products and Processes*. Springer.

Eppler, M. J., and Muenzenmayer, P. 2002. "Measuring Information Quality in the Web Context: A Survey of State-of-the-Art Instruments and an Application Methodology," *7th International Conference on Information Quality*: Citeseer, pp. 187-196.

Even, A., and Shankaranarayanan, G. 2005. "Value-Driven Data Quality Assessment," *Tenth International Conference on Information Quality (ICIQ'05)*.

Evoke. 2016. "Evoke Technologies." Retrieved 25/1/2016, 2016, from http://www.evoketechnologies.com/

Favre, J.-M. 2004. "Towards a Basic Theory to Model Model Driven Engineering," *3rd Workshop in Software Model Engineering, WiSME*: Citeseer.

Fereday, J., and Muir-Cochrane, E. 2006. "Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development," *International journal of qualitative methods* (5:1), pp. 80-92.

Fleiss, J. L., and Cohen, J. 1973. "The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability," *Educational and psychological measurement*).

Floridi, L. 2011. *The Philosophy of Information*. Oxford University Press.

Fodor, J. A. 1987. "Psychosemantics:The Problem of Meaning in the Philosophy of Mind,").

Fodor, J. A. 1990. *A Theory of Content and Other Essays*. The MIT press.

Franch, X. 2013. "Software Requirement Patterns," *Proceedings of the 2013 International Conference on Software Engineering*: IEEE Press, pp. 1499-1501.

Friedman, F. 2012. "Magic Quadrant for Data Quality Tools," Gartner Inc.

Gackowski, Z. J. 2005. "Informing Systems in Business Environments: A Purpose-Focused View," *Informing Science: International Journal of an Emerging Transdiscipline* (8), pp. 101-122.

Gamma, E., Vlissides, J., Helm, R., and Johnson, R. 1995. "Design Patterns: Elements of Reusable Object-Oriented Software," *Reading: Addison-Wesley*).

Garvin, D. A. 1987. "Competing on the Eight Dimensions of Quality," *Harvard Business Review*:November-December), pp. 101-109.

Garvin, D. A. 1988. *Managing Quality: The Strategic and Competitive Edge*. Simon and Schuster.

Gatling, C. B., R. Champlin, H. Stefani, G. Weigel. 2007. *Enterprise Information Management with Sap*. Boston: Galileo Press Inc.

Ge, M., and Helfert, M. 2013. "Cost and Value Management for Data Quality," in *Handbook of Data Quality* S. Sadiq (ed.). Springer, pp. 75-92.

Guest, G., MacQueen, K. M., and Namey, E. E. 2011. *Applied Thematic Analysis*. Sage.

Gregor, S., & Hevner, A. R. (2011). Introduction to the special issue on design science. Information Systems and E-Business Management, 9(1), 1-9.

Guizzardi, G. 2007. "On Ontology, Ontologies, Conceptualizations, Modeling Languages, and (Meta) Models," *Frontiers in artificial intelligence and applications* (155), p. 18.Halle, B. V., and Ronald, G. 2001. *Business Rules Applied: Building Better Systems Using the Business Rules Approach*. John Wiley & Sons, Inc.

Halle, B. V., and Ronald, G. 2001. *Business Rules Applied: Building Better Systems Using the Business Rules Approach*. John Wiley & Sons, Inc.

Heinrich, B., Kaiser, M., and Klier, M. 2007. "How to Measure Data Quality? A Metric-Based Approach,").

Hevner, A. R. 2007. "A Three Cycle View of Design Science Research," *Scandinavian journal of information systems* (19:2), p. 4.

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS quarterly* (28:1), pp. 75-105.

HIQA. 2011. "International Review of Data Quality " *Health Information and Quality Authority (HIQA), Ireland. http://www.hiqa.ie/press-release/2011-04-28-international-review-data-quality.*).

Hoffmann, A., Söllner, M., and Hoffmann, H. 2012. "Twenty Software Requirement Patterns to Specify Recommender Systems That Users Will Trust,").

Huang, K. T., Lee, Y. W., and Wang, R. Y. 1998. *Quality Information and Knowledge*. Prentice Hall PTR.

IAIDQ. 2015. "International Association for Information and Data Quality." Retrieved 01/08/2015, 2015, from http://www.iaidq.org/

IEEE. 1990. "Ieee Standard Glossary of Software Engineering Terminology." IEEE, p. 1.

Iivari, J. 2007. "A Paradigmatic Analysis of Information Systems as a Design Science," *Scandinavian journal of information systems* (19:2), p. 5.

Inmon, W. H. 1993. *Building the Data Warehouse*. Wiley, New York.

ISAT. 2004. "Guidelines for the Data Quality Improvement of Localization Data in Public Administration," P. administration (ed.).

ISO. 2000. "9001: 2008 Quality Management Systems. Requirements," *International Organization for Standardization*).

ISO. 2011. "Iso/Ts 8000-1 Data Quality Part 1: Overview." ISO.

ISO. 2012. "Iso 8000-2 Data Quality-Part 2-Vocabulary." ISO.

IT Governance Institute. 2007. "Control Objectives for Information and Related Technologies," IT Governance Institute, USA.

Jackson, M. 2009. "Some Notes on Models and Modelling," *Conceptual Modeling: Foundations and Applications*), pp. 68-81.

Jacobson, I. 1992. "Object Oriented Software Engineering: A Use Case Driven Approach,").

Jayawardene, V., Sadiq, S., and Indulska, M. 2012. "Practical Significance of Key Data Quality Research Areas," *16th Pacific Asia Conference on Information Systems*.

Jayawardene, V., Sadiq, S., and Indulska, M. 2013a. "An Analysis of Data Quality Dimensions."

Jayawardene, V., Sadiq, S., and Indulska, M. 2013b. "The Curse of Dimensionality in Data Quality," *ACIS 2013: 24th Australasian Conference on Information Systems*, pp. 1-11.

Jeusfeld, M., Quix, C., and Jarke, M. 1998. "Design and Analysis of Quality Information for Data Warehouses," *Conceptual Modeling–ER'98*), pp. 349-362.

Johns, D. 2010. "How to Present Data Quality Dimensions for Maximum Impact " Retrieved 14/03/2016, 2016, from http://dataqualitypro.com/data-quality-pro-blog/data-quality-dimensions-techniques

Jones, M. V., Coviello, N., & Tang, Y. K. (2011). International entrepreneurship research (1989–2009): a domain ontology and thematic analysis. Journal of business venturing, 26(6), 632-659.

Juran, J. M. 1962. *Quality Control Handbook*. New York: McGraw-Hill Publishing

Kahn, B. K., and Strong, D. M. 1998. "Product and Service Performance Model for Information Quality: An Update," *IQ*, pp. 102-115.

Khatri, V., and Brown, C. V. 2010. "Designing Data Governance," *Communications of the ACM* (53:1), pp. 148-152.

Kim, W., Choi, B. J., Hong, E. K., Kim, S. K., and Lee, D. 2003. "A Taxonomy of Dirty Data," *Data Mining and Knowledge Discovery* (7:1), pp. 81-99.

Kimball, R., and Caserta, J. 2004. "The Data Warehouse Etl Toolkit: Practical Techniques for Extracting," *Cleaning, Conforming, and Delivering, Digitized Format, originally published*).

Kimball, R., Reeves, L., Ross, M., and Thomthwaite, W. 1998. "The Data Warehouse Lifecycle Toolkit: Tools and Techniques for Designing, Developing, and Deploying Data Warehouses," *Jonh Wiley & Sons, New York*).

Klein, H. K., and Myers, M. D. 1999. "A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems," *MIS quarterly*), pp. 67-93.

Kock, N., Gray, P., Hoving, R., Klein, H., Myers, M. D., and Rockart, J. 2002. "Is Research Relevance Revisited: Subtle Accomplishment, Unfulfilled Promise, or Serial Hypocrisy?," *Communications of the Association for Information Systems* (8:1), p. 23.

Koudas, N., Sarawagi, S., and Srivastava, D. 2006. "Record Linkage: Similarity Measures and Algorithms," *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*: ACM, pp. 802-803.

Kühne, T. 2006. "Matters of (Meta-) Modeling," *Software & Systems Modeling* (5:4), pp. 369-385.

Kung, C., and Soelvberg, A. 1986. "Activity Modeling and Behavior Modeling," *Proc. of the IFIP WG 8.1 working conference on Information systems design methodologies: improving the practice*: North-Holland Publishing Co., pp. 145-171.

Lamsweerde, A. V. 2009. *Requirements Engineering- Fromsystem Goals to Uml Models to Software Specifications*. West Sussex PO19 8SQ England: John Wiley & Sons Ltd.

Landis, J. R., and Koch, G. G. 1977. "The Measurement of Observer Agreement for Categorical Data," *biometrics*), pp. 159-174.

Lapouchnian, A. 2005. "Goal-Oriented Requirements Engineering: An Overview of the Current Research," *University of Toronto*), p. 32.

Lee, Y. W., Pipino, L. L., Funk, J. D., and Wang, R. Y. 2009. *Journey to Data Quality*. The MIT Press.

Lee, Y. W., Strong, D. M., Kahn, B. K., and Wang, R. Y. 2002. "Aimq: A Methodology for Information Quality Assessment," *Information & management* (40:2), pp. 133-146.

Lesca, H., and Lesca, E. 1995. "Gestion De L'information(Qualité De L'information Et Performances De L'entreprise)," *les essentiels de la gestion*).

Li, S., Ragu-Nathan, B., Ragu-Nathan, T., and Rao, S. S. 2006. "The Impact of Supply Chain Management Practices on Competitive Advantage and Organizational Performance," *Omega* (34:2), pp. 107-124.

Liebenau, J., and Backhouse, J. 1990. *Understanding Information: An Introduction*. Palgrave Macmillan.

Lima, L. F. R., Maçada, A. C. G., and Vargas, L. M. 2006. "Research into Information Quality: A Study of the State of the Art in Iq and Its Consolidation," *ICIQ*, pp. 146-158.

Lindsey, E. 2008. *Three-Dimensional Analysis - Data Profiling Techniques*. Data Profiling LLC.

Long, J. A., and Seko, C. E. 2005. "A Cyclic-Hierarchical Method for Database Data-Quality Evaluation and Improvement," *Advances in Management Information Systems-Information Quality (AMIS-IQ) Monograph*).

Loshin, D. 2001. *Enterprise Knowledge Management: The Data Quality Approach*. Morgan Kaufmann Pub.

Loshin, D. 2004. "Enterprise Knowledge Management - the Data Quality Approach." Morgan Kaufmann.

Loshin, D. 2006. "Monitoring Data Quality Performance Using Data Quality Metrics," *Informatica Corporation*).

Loshin, D. 2009. *Master Data Management*. Morgan Kaufmann.

Loshin, D. 2010. *The Practitioner's Guide to Data Quality Improvement*. Morgan Kaufmann.

Loshin, D. 2011. *Practitioner's Guide for Data Quality*. New York: Elsevier Inc.

Lyon, M. 2008. "Assessing Data Quality,Monetary and Financial Statistics," *Bank of England. http://www.bankofengland.co.uk/statistics/Documents/ms/articles/art1mar08.pdf.)*.

Machado, R. J., Ramos, I., and Fernandes, J. M. 2005. "Specification of Requirements Models," in *Engineering and Managing Software Requirements*. Springer, pp. 47-68.

Madnick, S. E., Wang, R. Y., Lee, Y. W., and Zhu, H. 2009. "Overview and Framework for Data and Information Quality Research," *Journal of Data and Information Quality (JDIQ)* (1:1), p. 2.

March, S. T., and Smith, G. F. 1995. "Design and Natural Science Research on Information Technology," *Decision support systems* (15:4), pp. 251-266.

Marshall, Martin N. 1996. 'Sampling for qualitative research', *Family practice*, 13: 522-26.

Maxwell, J. A. 2004. "Causal Explanation, Qualitative Research, and Scientific Inquiry in Education," *Educational researcher* (33:2), pp. 3-11.

McGilvray, D. 2008. *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*. Morgan Kaufmann.

Mettler, T., Rohner, P., and Baacke, L. 2008. "Improving Data Quality of Health Information Systems: A Holistic Design-Oriented Approach," *ECIS*, pp. 1883-1893.

Miles, Matthew B Huberman, A Michael Matthew B Miles, and A Michael Huberman. 1994. *Anexpanded sourcebook qualitative data analysis*.

Moore, G. C., and Benbasat, I. 1991. "Development of an Instrument to Measure the Perceptions of Adopting an Information Technology Innovation," *Information systems research* (2:3), pp. 192-222.

Morgan, D. L. 1997. *The Focus Group Guidebook*. Sage publications.

Morris, C. 1938. "Foundation of the Theory of Signs." London: University of Chicago Press.

Mylopoulos, J. 1992. "Conceptual Modelling and Telos1,").

Nahm, A. Y., Rao, S. S., Solis-Galvan, L. E., and Ragu-Nathan, T. 2002. "The Q-Sort Method: Assessing Reliability and Construct Validity of Questionnaire Items at a Pre-Testing Stage," *Journal of Modern Applied Statistical Methods* (1:1), p. 15.

Nawaz, A. 2012. "A Comparison of Card-Sorting Analysis Methods," *The 10th Asia Pacific Conference on Computer Human Interaction. 2012*, pp. 583-592.

Neely, M. P., and Cook, J. 2008. "A Framework for Classification of the Data and Information Quality Literature and Preliminart Results (1996-2007)," *AMCIS 2008 Proceedings*), p. 131.

Ng, W. S., Watts, P., Lawson, Z., Kemp, A., and Maguire, S. 2012. "Development and Validation of a Standardized Tool for Reporting Retinal Findings in Abusive Head Trauma," *American journal of ophthalmology* (154:2), pp. 333-339. e335.

Noble, J. 1998. "Classifying Relationships between Object-Oriented Design Patterns," *Software Engineering Conference, 1998. Proceedings. 1998 Australian*: IEEE, pp. 98-107.

Nunamaker, J. F., Chen, M., and Purdin, T. D. 1990. "Systems Development in Information Systems Research," *Journal of management information systems* (7:3), pp. 89-106.

O'Brien, J. A. 1990. *Introduction to Information Systems in Business Management*. McGraw-Hill Professional.

Objectiver. 2007. "Kaos Tutorial."

Offermann, P., Levina, O., Schönherr, M., and Bub, U. 2009. "Outline of a Design Science Research Process," *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology*: ACM, p. 7.

OIC. 2012. "Routine Personal Work Information of Public Sector Employees," O.o.t.I. Commissioner (ed.). Office of the Information Commissioner.

Olivé, A. 2007. *Conceptual Modeling of Information Systems*. Springer.

Oliveira, P., Rodrigues, F., and Henriques, P. R. 2005. "A Formal Definition of Data Quality Problems," *IQ*.

OMG. 2004. "Uml 2.0 Infrastructure Specification," *Omg formal document*), pp. 03-09.

Omg, Q. 2008. "Meta Object Facility (Mof) 2.0 Query/View/Transformation Specification," *Final Adopted Specification (November 2005)*).

Otto, B., Wende, K., Schmidt, A., and Osl, P. 2007. "Towards a Framework for Corporate Data Quality Management," *ACIS 2007 Proceedings*), p. 109.

Panahy, P. H. S., Sidi, F., Affendey, L. S., Jabar, M. A., Ibrahim, H., and Mustapha, A. 2013. "A Framework to Construct Data Quality Dimensions Relationships," *Indian Journal of Science and Technology* (6:5), pp. 4422-4431.

Patton, Michael Quinn. 2005. *Qualitative research* (Wiley Online Library).

Purao, S., Storey, V. C., & Han, T. (2003). Improving analysis pattern reuse in conceptual design: Augmenting automated processes with supervised learning. Information Systems Research, 14(3), 269-290.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. "A Design Science Research Methodology for Information Systems Research," *Journal of management information systems* (24:3), pp. 45-77.

Pham Thi, T. T., and Helfert, M. 2007. "Modelling Information Manufacturing Systems," *International Journal of Information Quality* (1:1), pp. 5-21.

Pierce, E., Talburt, J., and Yonke, L. 2013. "The Data Quality Profession," in *Hand Book of Data Quality,* S. Sadiq (ed.). Springer-Verlag Berlin Heidelberg, pp. 397-418.

Pierce, E. M. 2002. "Extending Ip-Maps: Incorporating the Event-Driven Process Chain Methodology," *Proc. 7th International Conference on Information Quality (IQ 2002)*.

Pipino, L. L., Lee, Y. W., and Wang, R. Y. 2002. "Data Quality Assessment," *Communications of the ACM* (45:4), pp. 211-218.

Powell, C. 2003. "The Delphi Technique: Myths and Realities," *Journal of advanced nursing* (41:4), pp. 376-382.

Prasanna, R., Yang, L., and King, M. 2009. "Gdia: A Cognitive Task Analysis Protocol to Capture the Information Requirements of Emergency First Responders," *Proceedings of the 6th International ISCRAM Conference*, pp. 1-10.

Price, R., and Shanks, G. 2004. "A Semiotic Information Quality Framework," *Proceedings of the International Conference on Decision Support Systems DSS04*: Citeseer, pp. 658-672.

Price, R., and Shanks, G. 2005a. "A Semiotic Information Quality Framework: Development and Comparative Analysis," *Journal of Information Technology* (20:2), pp. 88-102.

Price, R. J., and Shanks, G. 2005b. "Empirical Refinement of a Semiotic Information Quality Framework," *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*: IEEE, pp. 216a-216a.

Principles for effective risk data aggregation and risk reporting. 2013. in: *BCBS 239*. Basel Committee on Banking Supervision. Switzerland: Bank for International Settlements.

Randolph, J. 2005. "Free-Marginal Multirater Kappa: An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. Joensuu University Learning and Instruction Symposium 2005," *October 1415th*).

Randolph, J. 2008. "Online Kappa Calculator." Retrieved 10/10/2014, 2914, from http://justus.randolph.name/kappa

Redman, T. 2008. *Data Driven: Profiting from Your Most Important Business Asset*. Boston: Harvard Business Press.

Redman, T. C. 1997. *Data Quality for the Information Age*. Artech House, Inc.

Richards, H., and White, N. 2013. "Ensuring the Quality of Health Information: The Canadian Experience," in *Handbook of Data Quality*. Springer, pp. 321-346.

Richards, L. 2014. *Handling Qualitative Data: A Practical Guide*. Sage.

Riehle, D., and Züllighoven, H. 1996. "Understanding and Using Patterns in Software Development," *TAPOS* (2:1), pp. 3-13.

Robinson, W. N., and Elofson, G. 2004. "Goal Directed Analysis with Use Cases," *Journal of Object Technology* (3:5), pp. 125-142.

Rolland, C., Loucopoulos, P., Grosz, G., and Nurcan, S. 1998. "A Framework for Generic Patterns Dedicated to the Management of Change in the Electricity Supply Industry," *Database and Expert Systems Applications, 1998. Proceedings. Ninth International Workshop on*: IEEE, pp. 907-912.

Rosemann, M., and Green, P. 2002. "Developing a Meta Model for the Bunge–Wand–Weber Ontological Constructs," *Information Systems* (27:2), pp. 75-91.

Rosemann, M., and Vessey, I. 2008. "Toward Improving the Relevance of Information Systems Research to Practice: The Role of Applicability Checks," *MIS Quarterly*), pp. 1-22.

Rosemann, M., and Zur Muehlen, M. 1998. "Evaluation of Workflow Management Systems-a Meta Model Approach," *Australian Journal of Information Systems* (6:1).

Ross, R. 1997. "The Business Rule Book. Classifying, Defining and Modelling Rules. Data Base Research Group," *Inc.-1997*).

Russell, R. S., and Taylor, B. W. 2003. *Operations Management*. Prentice Hall Upper Saddle River, NJ.

Rosenkranz, C., & Holten, R. (2011). The variety engineering method: analyzing and designing information flows in organizations. *Information Systems and E-Business Management, 9*(1), 11-49.

Roser, S., Bauer, B., & Müller, J. P. Model-and Architecture-Driven Development in the Context of Cross-Enterprise Business Process Engineering. In *IEEE SCC, 2006* (pp. 119-126)

Sadiq, S. 2013. *Handbook of Data Quality*. Springer.

Sadiq, S., Indulska, M., and Jayawardene, V. 2011a. "Research and Industry Synergies in Data Quality Management," *16th International Conference on Information Quality (ICIQ2011)*: ICIQ.

Sadiq, S., Yeganeh, N. K., and Indulska, M. 2011b. "20 Years of Data Quality Research: Themes, Trends and Synergies," *The 22nd Australasian Database Conference. Sidney, Australia*.

Sadiq, S., Yeganeh, N. Y., and Indulska, M. 2011c. "An Analysis of Cross-Disciplinary Collaborations in Data Quality Research," *European Conference on Information Systems*, Helsinki Finland.

Scannapieco, M., and Catarci, T. 2002. "Data Quality under a Computer Science Perspective," *Archivi & Computer* (2), pp. 1-15.

Scannapieco, M., Missier, P., and Batini, C. 2005. "Data Quality at a Glance," *Datenbank-Spektrum* (14), pp. 6-14.

Scannapieco, M., Pernici, B., and Pierce, E. 2002. "Ip-Uml: Towards a Methodology for Quality Improvement Based on the Ip-Map Framework," *7th Int'l Conf. on Information Quality (ICIQ-02)*, pp. 8-10.

Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M., and Baldoni, R. 2004. "The Daquincis Architecture: A Platform for Exchanging and Improving Data Quality in Cooperative Information Systems," *Information systems* (29:7), pp. 551-582.

Scheer, A. W. 2009. "Business Process Engineering: Reference Models for Industrial Enterprises,").

Sebastian-Coleman, L. 2012. *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework*. Newnes.

Seidewitz, E. 2003. "What Models Mean," *Software, IEEE* (20:5), pp. 26-32.

Serrano, N., Caballero, I., and García, F. 2009. "Extending Bpmn to Support the Modeling of Data Quality Issues," *Int. Conf. on Inform. Quality (ICIQ)*, pp. 46-60.

Shankaranarayanan, G., and Wang, R. Y. 2007. "Ipmap: Current State and Perspectives," *Proceedings of the 12th International Conference on Information Quality*.

Shankaranarayanan, G., Wang, R. Y., and Ziad, M. 2000. "Ip-Map: Representing the Manufacture of an Information Product," *Proceedings of the 2000 Conference on Information Quality*, pp. 1-16.

Simon, H. A. 1996. *The Sciences of the Artificial*. MIT press.

Smith, B. 2008. "Chapter 11:Ontology," in *The Blackwell Guide to the Philosophy of Computing and Information*, L. Floridi (ed.). Wiley-Blackwell.

Smith, H. 2012. "Using Pre-Built Rule Definitions with Ibm Infosphere Information Analyzer." Retrieved 15/04/2014, 2014, from http://www.ibm.com/developerworks/data/library/techarticle/dm-112isanalyzerrules/#download

Spencer, D. 2009. *Card Sorting: Designing Usable Categories*. Rosenfeld Media.

Stachowiak, H. 1973. "{Allgemeine Modelltheorie},").

Storey, V., and Wang, R. 2001. "Extending the Er Model to Represent Data Quality Requirements." Kluver Academic Publishers.

Strong, D. M., Lee, Y. W., and Wang, R. Y. 1997. "10 Potholes in the Road to Information Quality," *Computer*:8), pp. 38-46.

Stvilia, B., Gasser, L., Twidale, M. B., and Smith, L. C. 2007. "A Framework for Information Quality Assessment," *Journal of the American Society for Information Science and Technology* (58:12), pp. 1720-1733.

Su, Y., and Jin, Z. 2006. "A Methodology for Information Quality Assessment in the Designing and Manufacturing Process of Mechanical Products," *Information Quality Management: Theory and Applications*), pp. 190-220.

Sutcliffe, A. G., Maiden, N. A., Minocha, S., and Manuel, D. 1998. "Supporting Scenario-Based Requirements Engineering," *Software Engineering, IEEE Transactions on* (24:12), pp. 1072-1088.

Talburt, J. R. 2011. *Entity Resolution and Information Quality*. Elsevier.

Tu, S. Y., and Wang, Y. Y. R. 1993. *Modeling Data Quality and Context through Extension of the Er Model*. Total Data Quality Management Research Program, Sloan School of Management, Massachusetts Institute of Technology.

Van Der Aalst, W. M. P., Ter Hofstede, A. H. M., Kiepuszewski, B., and Barros, A. P. 2003. "Workflow Patterns," *Distributed and parallel databases* (14:1), pp. 5-51.

Van Lamsweerde, A. 2001. "Goal-Oriented Requirements Engineering: A Guided Tour," *Requirements Engineering, 2001. Proceedings. Fifth IEEE International Symposium on*: IEEE, pp. 249-262.

Valverde, R., Toleman, M., & Cater-Steel, A. (2011). A method for comparing traditional and component-based models in information systems re-engineering. *Information Systems and E-Business Management, 9*(1), 89-107.

Venkatesh, V., and Davis, F. D. 2000. "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies," *Management science* (46:2), pp. 186-204.

Walls, J. G., Widmeyer, G. R., and El Sawy, O. A. 1992. "Building an Information System Design Theory for Vigilant Eis," *Information systems research* (3:1), pp. 36-59.

Walsham, G. 1993. *Interpreting Information Systems in Organizations*. John Wiley & Sons, Inc.

Wand, Y., Monarchi, D. E., Parsons, J., and Woo, C. C. 1995. "Theoretical Foundations for Conceptual Modelling in Information Systems Development," *Decision Support Systems* (15:4), pp. 285-304.

Wang, R., and V., S. 1998. "Modeling Quality Requirements in Conceptual Database Design," *Conference on Information Quality, Massachusetts Institute of Technology*.

Wang, R. Y. 1998. "A Product Perspective on Total Data Quality Management," *Communications of the ACM* (41:2), pp. 58-65.

Wang, R. Y., Lee, Y. W., Pipino, L. L., and Strong, D. M. 1998. "Manage Your Information as a Product," *Sloan Management Review* (39:4), pp. 95-105.

Wang, R. Y., and Strong, D. M. 1996. "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of management information systems*), pp. 5-33.

Wang, R. Y., Ziad, M., and Lee, Y. W. 2000. *Data Quality*. USA: Kluwer Academic Publishers.

Wende, K. 2007. "A Model for Data Governance–Organising Accountabilities for Data Quality Management," *18th Australasian Conference on Information Systems. The University of Southern Queensland, Toowoomba, Australia*, pp. 417-425.

Willcocks, L., and Lester, S. 1996. "Beyond the It Productivity Paradox," *European Management Journal* (14:3), pp. 279-290.

Winkler, W. E. 2004. "Methods for Evaluating and Creating Data Quality," *Information Systems* (29:7), pp. 531-550.

Yang, L., Prasanna, R., and King, M. 2014. "Gdia: Eliciting Information Requirements in Emergency First Response," *Requirements Engineering*), pp. 1-18.

Yin, R. K. 2013. *Case Study Research: Design and Methods*. Sage publications.

Yonke, C. L., Walenta, C., and Talburt, J. R. 2011. "The Job of the Information/Data Quality Professional," International Association for Information and data Quality (IAIDQ).

Yu, E. 2011. "Modelling Strategic Relationships for Process Reengineering," *Social Modeling for Requirements Engineering* (11), p. 2011.

Yu, E., and Mylopoulos, J. 1998. "Why Goal-Oriented Requirements Engineering," *Proceedings of the 4th International Workshop on Requirements Engineering: Foundations of Software Quality*.

Zhang, R., Jayawardene, V., Indulska, M., Sadiq, S., and Zhou, X. 2014. "A Data Driven Approach for Discovering Data Quality Requirements,").

APPENDIX-A

Protocol for the applicability check

1.0 Selecting participant organizations

Coyne (1997) and Patton (2005) describes that all sampling done in qualitative research are "purposeful sampling". As per Patton (2005), Purposeful sampling is very powerful since it leads to information-rich cases for an in-depth study. Marshall (1996) refers to purposeful sampling as "judgement sample" where he describes as the most common and most intellectual strategy used in academia based on the researchers' practical knowledge of the research area. In this case, the researcher actively selects the most productive sample to answer the research question considering the special expertise of the subjects based on real world evidence.

Based on the above argument we use purposeful samling to select the participant organizations for this study and we use the following criteria to select the sample organizations.

1. Participant organizations should have a dedicated DQ management team
2. Affiliations to professional bodies for DQ
3. Special recognitions/awards achieved for credible DQ initiatives

Based on this criteria we select the best organizations at our proximity to conduct the evaluation.

2.0 Methodology

Once the organizations are selected we perform a DQ requirements elicitation in each organization using the methodology (KAOS4DQ). The study is conducted in three phases where the first two phases are focussed on eliciting DQ requirements using DQ patterns and KAOS4DQ methodology. Hence it is quite identical to the two phases in KAOS4DQ which enable us to elicit DQ requirements systematically.

Phase-1: Top-down analysis (Analysis of DQ context).

Phase-2: Bottom-up analysis (Analysis of DQ problems).

Phase-3: Validation of findings

Phase three is focussed on the validation of the DQ requirements identified as a result of the first two phases by referring to the DQ practitioners' viewpoint, thus confirming the accuracy of the findings by eliminating any subjectivity caused by researcher misinterpretations etc.

2.1 Phase-1: Top-down analysis (Analysis of DQ context)

In this phase, the prime focus is to examine the nature of DQ problems experienced by the organization in concern. As per Rosemann and Vessey (2008), it is important to design the applicability check interview catering to the organizational context. Hence, in turn, this survey fulfilled a vital requirement in applicability checks by identifying the nature DQ problems which will be used an input to conduct the applicability check interview in the next phase.

Participants: In order to collect this high-level information, the head of the data quality team is selected as a suitable participant considering his\her broad view about the entire data quality function in the organization.

The scope of the survey: In order to limit the scope of the study to a manageable corpus, in answering the survey, the participant is asked to focus on a single critical data set (product data, customer data etc.) in the organization. Hence we can restrict our investigation to that particular data set in this study.

Data collection: The information is collected through an online survey as suggested by the methodology KAOS4DQ using the survey tool designed for that (Appendix-B). The results of the survey reveal the most applicable DQ characteristics for the organization that helped us to plan the next phase.

Data analysis: The survey response is analysed as per the data analysis criteria given in Appendix-B and most relevant top 10 DQ patterns are selected.

The outcome of this phase: A list of data quality patterns applicable for a particular data set in the organization.

2.2 Phase-2: Bottom-up analysis (Analysis of DQ problems)

In this phase, the objective is to elicit the real DQ problems in the organization. We select semi-structured group interviews, as the data collection methodology in this session considering its support towards inter-participant interaction (Rosemann and Vessey 2008).

2.2.1 Participants for phase-2

The participants for Phase 2 are the head of the DQ team and some members of the DQ team with an in-depth understanding about DQ problems in the organizational dataset in concern. It should be noted that in KAOS4DQ, all data stakeholders are used to elicit DQ requirements in a data set. But here we

use DQ professionals since this is an applicability check and in-depth professional knowledge could be used in evaluating our artefact (Rosemann and Vessey 2008).

These members are identified with the support of the DQ manager based on the following selection criteria.

- DQ professionals having more than 3 years of experience in DQ management / active participation in at least 3 data quality projects. The number of participants varied depending on the size of the DQ team in an organization and it was expected to have at least two participations from each organization. (Group interviews need at least two participants)

Individual invitations are sent to participants and their participation is considered voluntary. E-mail consent is obtained from each participant prior to the study and formal consent form is signed at the interview.

In validating MIS research artefacts in industry environments, Rosemann and Vessey (2008) pointed out that, it is important to communicate the research artefact to the practitioners, well before the group interview. Then they will be well equipped with the required knowledge to participate in the experiment/interview/focus group. Hence few days prior to the interview we send them the URL[8] for our online patterns repository where the participants could gain an insight into DQ patterns.

2.2.2 Data collection for phase-2

The list of DQ characteristics identified in phase-1 is considered as input to the discussion and we elicit information about the DQ problems pertaining to the most important ten DQ characteristics (but not limited to ten) identified in the relevant data set in phase-1.

The setup for the semi-structured group interview is as follows.

1. Participants are seated at a round table set up.
2. Introduction to DQ patterns referring to eight DQ dimensions and respective characteristics to educate the participants about the artefact. This is a brief refreshing presentation since we have already provided this material to the participants few days prior to the interview.
3. Participants are asked to introduce their roles in DQ team and their interaction with the data set identified in phase 1.

[8] http://dke.uqcloud.net/DataQualityPatterns/

4. Elicitation of DQ problems: Top ten DQ characteristics displayed in the presentation are taken one by one and the participants are asked to talk about the DQ problems related to each characteristic.
5. One interviewer acted as a moderator to drive the discussion towards eliciting required information as per the data collection sheet provided in Appendix C.
6. Two other interviewers supported to maintain the floor of the discussion. They use the relationships between the DQ characteristics (from Table 7.32 chapter-7) and introduce new DQ characteristics to the discussion forum where necessary.
7. Following context regarding each DQ problem is discussed and elicit.
    - DQ Problem.
    - Data associated with the DQ problem (Semantics of data, data providers)
    - How the DQ problem was detected.
    - The root cause of the DQ problem.
    - What is the main organizational activity affected by the DQ problem and its consequences and the responsible parties?
    - How the problem can be solved (or was resolved in the past) or any solutions proposed to resolve the problem so far.
8. The interview is audio recorded in order to prepare transcripts to be used for data analysis.

It should be noted that in this phase, we conduct the group interview in the same fashion as describe in KAOS4DQ, the only exception is that we do not assign the responsibilities of DQ goals and objects since it was not an important task in evaluating the artefact. (But the validation of requirements is performed separately in phase 3)

After the interview, the audio recording is converted to a transcript of text and all collected data is consolidated for analysis purpose.

The outcome of the interview process: Interview transcripts and manually collected data about DQ problems and the context explained in 7 above.

2.2.3 Data analysis of phase-2

In this step, the data collected in phase-2 are analysed systematically to define DQ goals, data objects, and DQ operational requirements. This is performed after the interview. We analyse the data using DQ patterns as follows.

I. **DQ problem context: The facts relating to the DQ problem in concern.**

Data quality problem: The data quality problem and the consequences explained to the participants using their own words and organizational terminology.

Applicable DQ pattern: An interpretation of the data quality problem from the viewpoint of data quality patterns. (Each data quality problem is a violation of the data quality characteristic of which the pattern is based on). In case the given data quality problem by the participants was complex where there were multiple underlying characteristics associated with the problem, we decomposed such problems into atomic problems based on underlying DQ characteristics.

Data quality goal: DQ characteristic in violation in each DQ problem was used to define a DQ goal.

Data object: Data associated with the problem recognised.

II. **Data quality solution context:** The facts relating to the solution of each DQ problem

Implementation form: Based on the implementation guidelines in an applicable DQ pattern a high-level solution is developed to address the DQ problem.

Operational DQ requirement: Contextual evidence provided by the interview data about the actual solution taken to resolve a problem in terms of rules or processes.

Duel coder approach: This analysis is performed by two different researchers individually, and afterwards the individual results were discussed jointly in order to reach consensus about the analysis.

The outcome of the analysis: DQ operational requirements for the data set in concern.

2.3 Phase-3: Validation of findings

As a result of phase-2, we are able to elicit the DQ requirements of the organization pertaining to a particular data set. Therefore we demonstrate how DQ patterns can be used to elicit DQ requirements using KAOS4DQ methodology. In this phase, the DQ operational requirements found are validated from the DQ practitioners. The prime focus of this validation is to find out whether we have correctly understood the DQ problems and produced the DQ operational requirements. (it should be noted that validation of requirements is an essential aspect in KAOS4DQ). Further, we want to collect feedback

from the practitioners about other similar DQ scenarios where the respective DQ patterns can be applicable. Data collection in this phase is done using an online survey.

DQ manager is selected as the participant of this phase, considering his broad insight into the DQ domain as well as his active participation in phase-1 and phase-2. In answering the survey, he is asked to consult his team members who participated in phase-2 where necessary.

In the survey, the DQ manager is provided with a template which contained the following information with regards to the data analysis. He/she had to agree or disagree (with reasons) with the following facts with regards to the formalization of each DQ requirement analysed using the DQ pattern based approach.

1. Data quality problem
2. Data quality characteristic in violation
3. Data quality goal
4. Data object and granularity
5. Characteristic type
6. Implementation form / DQ requirements
7. Probable verification metrics
8. Probable validation metrics

Further, the participant was asked about his opinion on similar data quality problems that could have a similar formalization of DQ requirements using the pattern-based approach.

APPENDIX-B

Survey tool to analyze DQ context

1.0 Phase-1: Top down analysis (Analysis of DQ context)

The following questions present 8 dimensions of data quality (DQ), including different data quality problems related to each dimension. Please indicate if any of the DQ problems faced by your organization are related to these. If so, indicate the problem frequency and severity of the impact.

The scale definition for problem frequency is as follows.

- 0% that is data quality problem not experienced
- <10% of data does not meet quality aspect in concern,
- 10-50% of data does not meet the quality aspect in concern
- 50-80% of data does not meet the quality aspect in concern
- >80 of data does not meet the quality aspect in concern

	Problem Frequency					**Severity of Impact**		
Q1. Dimension: Data completeness	**Not Experienced**	**<10%**	**10-50%**	**50-80%**	**>80%**	**Low**	**Medium**	**High**
Completeness of mandatory attributes: The attributes which are necessary for a complete representation of a real world entity contain null values	☐	☐	☐	☐	☐	☐	☐	☐
Completeness of optional attributes: Nonmandatory attributes contain invalid null values.(Valid null value: NULL is the right value - ZIP code of an address when the country in concern does not use ZIP codes . Invalid null value: Value is actually missing - ZIP code is missing for an instance where the country uses ZIP codes)	☐	☐	☐	☐	☐	☐	☐	☐
Completeness of records: Not every real world entity instance that is relevant for the organization is found in the data.	☐	☐	☐	☐	☐	☐	☐	☐

Data volume: The volume of data is deficient or overwhelming to perform an intended task.	☐	☐	☐	☐	☐	☐	☐	☐

Q2. Dimension: Data Availability and Accessibility	**Problem Frequency**					**Severity of Impact**		
	Not Experienced	**<10%**	**10-50%**	**50-80%**	**>80%**	**Low**	**Medium**	**High**
Continuity of data access: The technology infrastructure prohibits the speed and continuity of access to the data for the users.	☐	☐	☐	☐	☐	☐	☐	☐
Data maintainability: Data is not accessible to perform necessary updates and maintenance operations during its entire life-cycle.	☐	☐	☐	☐	☐	☐	☐	☐
Data awareness: The data users are not aware of all available data and its location	☐	☐	☐	☐	☐	☐	☐	☐
Ease of data access: The data is not easily accessible in a form that is suitable for its intended use. (users have to further process data to make it usable for the task at hand)	☐	☐	☐	☐	☐	☐	☐	☐
Data punctuality: Data is not available at the time of its intended use.	☐	☐	☐	☐	☐	☐	☐	☐
Data access control: The access to the data is not controlled to prevent damage or unauthorised access to data.	☐	☐	☐	☐	☐	☐	☐	☐

Q3. Dimension: Data Currency	Problem Frequency					Severity of Impact		
	Not Experienced	<10%	10-50%	50-80%	>80%	Low	Medium	High
Data timeliness: Data which refers to time (eg: currency exchange rate) is not available for use within an acceptable time relative to its time of creation.	☐	☐	☐	☐	☐	☐	☐	☐
Data freshness: Data which is subjected to changes over time (eg: address) is not fresh and up-to-date with respect to its intended use.	☐	☐	☐	☐	☐	☐	☐	☐

Q4. Dimension: Data Accuracy	Problem Frequency					Severity of Impact		
	Not Experienced	<10%	10-50%	50-80%	>80%	Low	Medium	High
Accuracy to reference source: Data does not agree with an identified source.	☐	☐	☐	☐	☐	☐	☐	☐
Accuracy to reality: Data does not truly reflect the real world.	☐	☐	☐	☐	☐	☐	☐	☐
Precision: Attribute values are not accurate as per linguistics or granularity (i.e. from language point of view or numerical point of view)	☐	☐	☐	☐	☐	☐	☐	☐

Q5. Dimension: Data Validity	Problem Frequency					Severity of Impact		
	Not Experienced	<10%	10-50%	50-80%	>80%	Low	Medium	High
Business rules compliance: Calculations on data do not comply with business rules.	☐	☐	☐	☐	☐	☐	☐	☐

Meta-data compliance: Data do not comply with its meta-data.	☐	☐	☐	☐	☐	☐	☐	☐
Standards and Regulatory compliance: Not all data processing activities comply with the policies, procedures, standards, industry benchmark practices and regulatory requirements that the organization is bound by.	☐	☐	☐	☐	☐	☐	☐	☐
Statistical validity: Computed data are not statistically valid.	☐	☐	☐	☐	☐	☐	☐	☐

	Problem Frequency					**Severity of Impact**		
Q6. Dimension: Data Reliability	**Not Experienced**	**<10%**	**10-50%**	**50-80%**	**>80%**	**Low**	**Medium**	**High**
Source Quality: Data used is not from trusted and credible sources.	☐	☐	☐	☐	☐	☐	☐	☐
Objectivity: Data is biased and partial.	☐	☐	☐	☐	☐	☐	☐	☐
Traceability: The lineage of the data is not verifiable.	☐	☐	☐	☐	☐	☐	☐	☐

	Problem Frequency					**Severity of Impact**		
Q7. Dimension: Data consistency	**Not Experienced**	**<10%**	**10-50%**	**50-80%**	**>80%**	**Low**	**Medium**	**High**
Uniqueness: The data is not uniquely identifiable. (Duplicates exist)	☐	☐	☐	☐	☐	☐	☐	☐
Non-redundancy: The data is recorded in exactly one place.	☐	☐	☐	☐	☐	☐	☐	☐
Semantic consistency: Data is not semantically consistent (i.e. the meaning of data is not consistent)	☐	☐	☐	☐	☐	☐	☐	☐
Value consistency: Data values are not consistent and hence provide conflicting or heterogeneous instances.	☐	☐	☐	☐	☐	☐	☐	☐

Format consistency: Data formats are not consistently used.	☐	☐	☐	☐	☐	☐	☐	☐
Referential integrity: Data relationships are not represented through referential integrity rules.	☐	☐	☐	☐	☐	☐	☐	☐

	Problem Frequency					**Severity of Impact**		
Q8. Dimension: Data Usability & Interpretability	**Not Experienced**	**<10%**	**10-50%**	**50-80%**	**>80%**	**Low**	**Medium**	**High**
Usefulness and relevance: The data is not useful and relevant for the task at hand.	☐	☐	☐	☐	☐	☐	☐	☐
Understandability: The data is not understandable.	☐	☐	☐	☐	☐	☐	☐	☐
Appropriate Presentation: The data presentation is not aligned with its use.	☐	☐	☐	☐	☐	☐	☐	☐
Interpretability: Data cannot be interpreted.	☐	☐	☐	☐	☐	☐	☐	☐
Information value: Data provides no business value to the organization.	☐	☐	☐	☐	☐	☐	☐	☐

2.0 Phase 1: Data Analysis

	Problem Frequency					**Severity of Impact**			**Frequency * Severity**	**Rank**
	Not Experienced	**<10%**	**10-50%**	**50-80%**	**>80%**	**Low**	**Medium**	**High**		
Weighting factors for responses	0	1	2	3	4	1	2	3		

The five levels of problem frequency is assigned with the weights 0 to 4 respectively while the levels of severity of impact is assigned with weights 1 to 3 respectively.

The product of problem frequency and severity of impact is calculated. The ranks are assigned based on the values for frequency * severity where the lowest rank (1) is given to the highest value and so on.

APPENDIX-C

Data collection format for Group Interview

Characteristic Name	DQ Problem:
Data associated with the problem	
Semantics of data	Granularity level
Characteristic type	
How the problem was originally detected?	
What other information / material is used to detect the problem?	
What are the metrics implemented to measure the problems?	

Root cause of the problem	
Business impact of the problem	
Affected business activity:	Consequences
Solutions taken to resolve the problem	
Solution	How to evaluate the progress?

APPENDIX-D

DQ characteristics and their candidate themes

Characteristic	Completeness of mandatory attributes	
Consolidated Definition	The attributes which are mandatory for a complete representation of a real world entity must contain values and cannot be null.	
Candidate Themes	”A given data element has a full value stored for all records that should have a value”	(English 2009)
	“Data element is always required to be populated…….”	(Byrne 2008)
	“Completeness refers to the expectation that certain attributes should have assigned values in a data set”	(Loshin 2001)
	” Determine the extent to which data is not missing. For example, an order is not complete without a price and quantity”	(Gatling 2007)

Table 1: Completeness of mandatory attributes

Characteristic	Completeness of optional attributes .	
Consolidated Definition	Optional attributes should not contain invalid null values	
Candidate Themes	“A null value might actually represent an unavailable value which can be either, an attribute that is not applicable for this entity, or there is no value in the attribute’s domain that correctly	(Loshin 2001)

	classifies this entity, or the value may actually be missing".	
	"Ability to distinguish…. null and default values from applicable values of the domain".	(Redman 1997)
	"In the case of null, a special element of an attribute's domain can be assigned as the attribute's value".	(Redman 1997)
	"…there should be a recognizable form for presenting that null value that does not conflict with any valid value".	(Loshin 2001)

Table 2: Completeness of optional attributes

Characteristic	Completeness of records	
Consolidated Definition	Every real world entity instance that is relevant for the organization can be found in the data	
Candidate Themes	"Every real world phenomenon is represented".	(Price and Shanks 2005)
	"Data is complete if no piece of information is missing, Anti-example: The Beatles were John Lennon, George Harrison and Ringo Starr".	(Kimball and Caserta 2004)
	"A record exists for every real world object or event, the enterprise needs to know about".	(English 2009)
	" Monitoring for incomplete lists of eligible records or missing data items will identify data quality problems".	(HIQA 2011)

	"…having all data that existed in the possession..."	(ISO 2012)

Table 3: Completeness of records

Characteristic	Data volume	
Consolidated Definition	The volume of data is neither deficient nor overwhelming to perform an intended task	
Candidate Themes	"Is the scope of information adequate? Not too much nor too little…"	(Eppler 2006)
	"A measure of the availability and comprehensiveness of data compared to the total data universe or population of interest"	(McGilvray 2008)
	" The quantity or volume of available data is appropriate"	(Wang and Strong 1996)
	" Degree of presence of data in a given collection"	(Scannapieco and Catarci 2002)

Table 4: Data volume

Characteristic	Continuity of data access	
Consolidated Definition	The technology infrastructure should not prohibit the speed and continuity of access to the data for the users.	
Candidate Themes	"Is there a continuous and unobstructed way to get to the information?"	(Eppler 2006)
	"Can the infrastructure match the user's working pace?"	(Eppler 2006)

	"Data is ……quick to retrieve"	(Price and Shanks 2005)
	"The frequency of failures of a system, its fault tolerance"	(Scannapieco and Catarci 2002)
	"Time interval between the submission of a query and the answer"	(Scannapieco and Catarci 2002)

Table 5: Continuity of data access

Characteristic	Data maintainability	
Consolidated Definition	Data should be accessible to perform necessary updates and maintenance operations in its entire lifecycle.	
Candidate Themes	"…. the degree to which data can be accessed, updated, maintained and managed".	(McGilvray 2008)
	"Can all of the information be organized and updated on an on-going basis?"	(Eppler 2006)

Table 6: Data maintainability

Characteristic	Data awareness
Consolidated Definition	Data users should be aware of all available data and its location.

Candidate Themes	"The awareness of data users of what data is being collected and knowing where it is located"	(HIQA 2011)

Table 7: Data awareness

Characteristic	Ease of data access	
Consolidated Definition	Data should be easily accessible in a form that is suitable for its intended use.	
Candidate Themes	"….ease of obtaining information objects relative to a particular activity."	(Stvilia et al. 2007)
	"Data are…. easily retrieved and used"	(Wang and Strong 1996)
	"Accessibility refers to the physical conditions in which users can obtain data easily……."	(Lyon 2008)

Table 8: Ease of data access

Characteristic	Data Punctuality	
Consolidated Definition	Data should be available at the time of its intended use.	
Candidate Themes	"Punctuality refers to the time lag between the release date of data and the target date when it should have been delivered"	(Lyon 2008)
	"The Characteristic of the Information being accessible when it is needed"	(English 2009)

	"Is the information processed and delivered without delays?"	(Eppler 2006)
	"Timeliness refers to the time expectation for accessibility and availability of information"	(Loshin 2006)

Table 9: Data Punctuality

Characteristic	Data access control	
Consolidated Definition	The access to data should be controlled to ensure it is secure against damage or unauthorised access.	
Candidate Themes	"Is the information protected against loss or unauthorized access?"	(Eppler 2006)
	"The extent to which information is protected from harm in the context of a particular activity".	(Stvilia et al. 2007)
	"Access to data can be restricted and hence kept secure"	(Wang and Strong 1996)
	"Data is appropriately protected from damage or abuse"	(Price and Shanks 2005)

Table 30: Data access control

Characteristic	Data timeliness	
Consolidated Definition	Data which refers to time, should be available for use within an acceptable time relative to its time of creation	
Candidate Themes	"The entity represents the most current information resulting from the output of a business event."	(Byrne 2008)

	"Timeliness of data refers to the extent to which data is collected within a reasonable time period from the activity or event and is available within a reasonable timeframe to be used for whatever purpose it is intended"	(HIQA 2011)
	"The age of the data is correct for the Knowledge Worker's purposes. Purposes such as inventory control for Just-in-Time Inventory require the most current data"	(English 2009)
	"The amount of time the information remains valid in the context of a particular activity that generates data"	(Stvilia et al. 2007)
	"….data are current and available for use as specified and in the time frame in which they are expected."	(McGilvray 2008)
	"The age of the data is appropriate for the task at hand"	(Wang and Strong 1996)
	"How long data remains valid"	(Scannapieco and Catarci 2002)

Table 41: Data timeliness

Characteristic	Data freshness
Consolidated Definition	Data which is subjected to changes over the time, should be fresh and up-to-date with respect to its intended use.

Candidate Themes	"A datum value is up-to-date if it is correct in spite of a possible discrepancy caused by time related change to the correct value."	(Redman 1997)
	"Currency can measure how "up-to-date" information is, and whether it is correct despite possible time-related changes."	(Loshin 2006)
	"A measure of the rate of negative change to the data."	(McGilvray 2008)
	"Is the information up to-date and not obsolete?"	(Eppler 2006)
	"Degree to which information is current with the world that it models"	(Loshin 2001)
	" ….data is accurate but not up to date: President of the USA is Bill Clinton"	(Kimball and Caserta 2004)
	"…data is sufficiently up-to-date for the task at hand."	(Gatling 2007)

Table 52: Data freshness

Characteristic	Accuracy to reference source	
Consolidated Definition	Data should agree with an identified source.	
Candidate Themes	"Accuracy of datum <e, a, v> refers the nearness of the value v to some value v' in the attribute domain, which is considered as the correct one for the entity e and the attribute a….. If the datum's value v coincides value v', the datum is said to be correct"	(Redman 1997)

	"Data accuracy refers to the degree with which data values agree with an identified source of correct information."	(Loshin 2001)
	"A measure of the correctness of the content of the data which requires an authoritative source of reference to be identified and accessible."	(McGilvray 2008)
	"The extent to which data are correct reliable and certified free of error."	(Wang and Strong 1996)
	"In many cases, accuracy is measured by how the values agree with an identified source of correct information."	(Loshin 2006)
	"The data agrees with an original, corroborative source record of data…."	(English 2009)
	"Accuracy of data refers to how closely the data correctly captures what it was designed to capture"	(HIQA 2011)
	"Degree of correctness of a value when comparing with a reference one"	(Scannapieco and Catarci 2002)
	"…..extent to which data is collected consistently over time….."	(HIQA 2011)

Table 63: Accuracy to reference source

Characteristic	Accuracy to reality
Consolidated Definition	Data should truly reflect the real world.

Candidate Themes	"Determines the extent to which data objects correctly represent the real-world values for which they were designed"	(Gatling 2007)
	"The data value correctly reflects the real-world condition"	(Byrne 2008)
	"Is the information…..? close enough to reality"	(Eppler 2006)
	"The degree to which an information object correctly represents another information object, process, or phenomenon in the context of a particular activity or culture"	(Stvilia et al. 2007)
	"Each identifiable data unit maps to the correct real-world phenomenon"	(Price and Shanks 2005)
	"Each identifiable data unit represents at least one specific real-world phenomenon"	(Price and Shanks 2005)
	"The extent to which the correctness of information is verifiable or provable in the context of a particular activity"	(Stvilia et al. 2007)
	("The data correctly reflects the characteristics of a Real-World object or event being described"	(English 2009)

Table 74: Accuracy to reality

Characteristic	Precision
Consolidated Definition	Attribute values should be accurate as per linguistics and granularity.

Candidate Themes	"Data is correct if it conveys a lexically, syntactically and semantically correct statement"	(Kimball and Caserta 2004)
	"Data values are correct to the right level of detail or granularity."	(English 2009)
	"Is the information to the point, void of unnecessary elements"	(Eppler 2006)
	"The granularity or precision of the model or content values of an information object according to some general-purpose IS-A ontology such as WordNet"	(Stvilia et al. 2007)
	"…information is legitimate… according to some stable reference source such as a dictionary or set of domain norms"	(Stvilia et al. 2007)
	"…If the domain is infinite (the rational numbers, for example), then no string format of finite length can represent all possible values and hence will not provide the precision to meet user needs"	(Redman 1997)
	"The degree of precision of an attribute's value should reasonably match the degree of precision of the value being displayed"	(Loshin 2001)
	"The extent to which an information object matches the precision."	(Stvilia et al. 2007)

Table 85: Precision

Characteristic	Business rules compliance	
Consolidated Definition	Data should comply with business rules.	
Candidate Themes	"Data values conform to the specified business rules"	(English 2009)
	"A derived or calculated data value is produced correctly according to a specified calculation formula or a set of derivation rules."	(English 2009)
	"Determines the extent to which data is not missing important relationship linkages. For example, the launch date for a new product must be valid and must be the first week of any quarter, since all new products are launched in the first week of each quarter."	(Gatling 2007)

Table 96: Business rules compliance

Characteristic	Metadata compliance	
Consolidated Definition	Data should comply with its metadata.	
Candidate Themes	"Data values are consistent with the attribute definition"	(English 2009)
	"The metadata of the data element clearly states or defines the purpose of the data element, or the values used in the data element can be understood by metadata or data inspection"	(Byrne 2008)
	A data value is a valid value or within a specified range of valid values for this data element	(English 2009)

	"Instances of data are either store, exchanged, or presented in a format that is consistent with the domain of values, as well as consistent with other similar attribute values"	(Loshin 2006)
	"Instances of data are represented in a format that is consistent with the domain of values."	(English 2009)
	"Determine the extent to which data confirms to a specified format."	(Gatling 2007)
	"…whether physical instances of data are in record with their formats."	(Redman 1997)
	"Appropriate metadata is available to define, constrain, and document data"	(Price and Shanks 2005)

Table 107: Metadata compliance

Characteristic	Standards and Regulatory compliance	
Consolidated Definition	All data processing activities should comply with the policies, procedures, standards, industry benchmark practices and all regulatory requirements that the organization is bound by.	
Candidate Themes	"Data is handled in accordance with any definition or standard which enable benchmarking over time"	(HIQA 2011)
	"A measure of the existence of…. data standards and models…"	(McGilvray 2008)
	"The data element has a standardized enterprise business definition…."	(Byrne 2008)

	Signs and other Information-bearing mechanisms like traffic lights should be universally used across the broadest audience possible	(English 2009)

Table 18: Standards and Regulatory compliance

Characteristic	Statistical validity	
Consolidated Definition	Computed data should be statistically valid.	
Candidate Themes	"Coherence of statistics is their adequacy to be reliably combined in different ways and for various uses"	(Lyon 2008)
	"Coherence of data refers to the internal consistency of the data…. Coherence is promoted through the use of standard concepts, classifications and target populations.."	(HIQA 2011)
	"Accuracy in the general statistical sense denotes the closeness of computations or estimates to the exact or true values"	(Lyon 2008)

Table 19: Statistical validity

Characteristic	Source Quality	
Consolidated Definition	Data used is from trusted and credible sources.	
Candidate Themes	"The source of information guarantees the quality of information it provides with remedies for non-compliance……."	(English 2009)

	"….the degree to which different departments conform to the usage of the enterprise data instead of relying on their own data sources"	(Loshin 2001)
	"The degree of reputation of an information object in a given community or culture."	(Stvilia et al. 2007)
	"Data are trusted or highly regarded in terms of their source and content"	(Wang and Strong 1996)

Table 110: Source Quality

Characteristic	Objectivity	
Consolidated Definition	Data is unbiased and impartial.	
Candidate Themes	" The degree to which Information is presented without bias…"	(English 2009)
	"Data are unbiased and impartial"	(Wang and Strong 1996)
	" "Is the information free of distortion, bias or error"	(Eppler 2006)
	"Extent to which data are unbiased (unprejudiced) and impartial"	(Scannapieco and Catarci 2002)

Table 121: Objectivity

Characteristic	Traceability
Consolidated Definition	The lineage of the data is verifiable.

Candidate Themes	”Is the background of the information visible and traceable?”	(Eppler 2006)
	“The extent to which the correctness of information is verifiable or provable by tracing back in the context of a particular activity.”	(Stvilia et al. 2007)
	“A data provenance record can include information about creation, update, transcription, abstraction, validation and transforming ownership of data”	(ISO 2012)

Table 132: Traceability

Characteristic	Uniqueness	
Consolidated Definition	The data is uniquely identifiable.	
Candidate Themes	“The entity is unique and there are no duplicate values.”	(Byrne 2008)
	“There is a key that can be used to uniquely access each entity.”	(Loshin 2006)
	“Each real-world phenomenon is either represented by at most one identifiable data unit or by multiple but consistent identifiable units or by multiple identifiable units whose inconsistencies are resolved within an acceptable time frame”	(Price and Shanks 2005)

Table 143: Uniqueness

Characteristic	Non-redundancy	
Consolidated Definition	The data is recorded in exactly one place	
Candidate Themes	"Unwanted duplication existing within or across systems for a particular field, record or data set"	(McGilvray 2008)
	"Only one record exists in a given data store that represents a single real-world object or event"	(English 2009)
	" Table columns(fields) are not repeated"	(Gatling 2007)

Table 154: Redundancy

Characteristic	Semantic consistency	
Consolidated Definition	Data is semantically consistent	
Candidate Themes	"….clearly states or defines the purpose of the data element."	(Byrne 2008)
	"Data about an object or event in one data store is semantically equivalent to the Meta-data of the same object in another data store."	(English 2009)
	"….using vocabulary control for elements to convey the same concepts and meanings in an information object."	(Stvilia et al. 2007)

Table 165: Redundancy

Characteristic	Value consistency	
Consolidated Definition	Data values are consistent and do not provide conflicting or heterogeneous instances	
Candidate Themes	"Consistency means that two or more things do not conflict with one another. With regards to data values, it means that a bit of added discipline is desired"	(Redman 1997)
	"….consistency refers to data values in one data set being consistent with values in another data set…"	(Loshin 2001)
	"Data is consistent if it doesn't convey heterogeneity in… contents…"	(Kimball and Caserta 2004)
	"Extent to which distinct data instances provide non-conflicting information about the same underlying data object."	(Gatling 2007)
	"Consistency can also reflect the regular use of standardized values….."	(Byrne 2008)
	"….consistency specifies that two data values drawn from separate data sets must not conflict with each other…"	(Loshin 2006)
	".. Consistency among different data values (e.g. Sex and Name).."	(Scannapieco and Catarci 2002)

Table 176: Value consistency

Characteristic	Format consistency	
Consolidated Definition	Data formats are consistently used.	
Candidate Themes	"A measure of the equivalence of data standards and formats used in various data stores, applications, and systems…."	(McGilvray 2008)
	"…..elements of an information object are consistently represented using the same structure and format."	(Stvilia et al. 2007)
	"…structured Attributes like dates, time, telephone number, tax ID number, product code, and currency amounts should be presented in a consistent, standard way	(English 2009)

Table 187: Format consistency

Characteristic	Referential integrity	
Consolidated Definition	Data relationships are represented through referential integrity rules	
Candidate Themes	"…an object identifier is used as foreign keys within a data set to refer to the core representation…"	
Candidate Themes	An integrity exists between entities in different tables	(Loshin 2006)
	"Data follows specified database integrity rules."	(English 2009)

Table 28: Referential integrity

Characteristic	Usefulness and relevance	
Consolidated Definition	Data is useful and relevant for the task at hand	
Candidate Themes	"The extent to which the information is informative in the context of an activity…"	(Stvilia et al. 2007)
	"Data are applicable and useful for the task at hand"	(Wang and Strong 1996)
	Relevance is the degree to which statistics meet current and potential users' needs	(Lyon 2008)
	"….extent to which the data meets the needs of users…."	(HIQA 2011)
	"The extent to which information is applicable in a given activity."	(Stvilia et al. 2007)
	"Can the information be applied? Is it useful?"	(Eppler 2006)
	"Does the information provision correspond to the user's needs and habits?"	(Eppler 2006)
	"The amount of information contained in an information object"	(Stvilia et al. 2007)
	"Data are of sufficient depth, breath and scope for the task at hand"	(Wang and Strong 1996)

	"Knowledge Workers have all the facts they need to perform their processes and improve the processes."	(English 2009)
	"The data includes all of the types of information important for its use"	(Price and Shanks 2005)
	"Data are beneficial and provide advantages for their users."	(Wang and Strong 1996)
	"Can the information be adapted by the information consumer?"	(Eppler 2006)
	"Data are relevant if they satisfy user-specified criteria"	(Scannapieco and Catarci 2002)

Table 29: Usefulness and relevance

Characteristic	Understandability	
Consolidated Definition	Data is understandable.	
Candidate Themes	"Data are in appropriate language and unit and data definitions are clear."	(Wang and Strong 1996)
	"Is there any ambiguity in understanding the data and is there information available to help the user understand the terminology"	(HIQA 2011)
	"When there is any possibility of ambiguity, a key or legend should be included."	(Loshin 2001)

	"Data are compactly represented without being overwhelmed"	(Wang and Strong 1996)
	"Data are clear without ambiguity and easily comprehended"	(Wang and Strong 1996)
	"Information is easily read and understood"	(English 2009)
	"Information is presented with clear labels, footnotes, and/or other explanatory notes, with references or links to definitions or documentation the clearly communicates the meaning"	(English 2009)
	"….extent to which data can be understood…"	(HIQA 2011)
	"Is the information understandable or comprehensible to the target group?"	(Eppler 2006)
	"The content of an object is focused on one topic facilitating comprehension."	(Stvilia et al. 2007)
	"The extent of cognitive complexity of an information object…."	(Stvilia et al. 2007)
	"….data characteristics are well understood….."	(Byrne 2008)
	"….content of an information object is expressed by conventional, typified terms and forms according to some general-purpose reference source……."	(Stvilia et al. 2007)
	"…format in which data are specified, including language spoken, units, etc. and to the clarity (non-ambiguity) of data definitions"	(Scannapieco and Catarci 2002)

Table 190: Understandability

Characteristic	Appropriate Presentation	
Consolidated Definition	Presentation of data is aligned with its use.	
Candidate Themes	"….Data is presented consistently in a standardized or consistent format across different media…"	(English 2009)
	" Good format, like good views, are flexible so that changes in user need and recording medium can be accommodated "	(Redman 1997)
	" How well the format and presentation of the data match the user needs"	(Loshin 2001)
	"….presentation is customized as needed, with respect to aggregating data and changing the data format, precision, or units."	(Price and Shanks 2005)
	"Flexibility in presentation describes the ability of the system to adapt to changes in both the represented information and in user requirements for presentation of information."	(Loshin 2001)
	"….a portable interface is important so that as applications are migrated from one platform to another, still the presentation of data is familiar to the users"	(Loshin 2001)
	"Good presentation formats are portable or universal. This means that they can be applied to as wide a range of situations as possible."	(Redman 1997)
	"Data is presented in an intelligible manner "	(Price and Shanks 2005)

	"...format and appearance of presentation support appropriate use of information "	(McGilvray 2008)
	"Data is presented in a manner appropriate for its use with respect to format……."	(Price and Shanks 2005)
	"The degree to which Information is presented in an Intuitive and relevant way."	(English 2009)
	"Data are always presented in the same format and are compatible with the previous data."	(Wang and Strong 1996)
	"…..information being presented in the right technology Media…."	(English 2009)
	"Data are always presented in the same format"	(Scannapieco and Catarci 2002)

Table 201: Appropriate presentation

Characteristic	Interpretability	
Consolidated Definition	Data should be interpretable.	
Candidate Themes	"Comparability aims at measuring and interpreting the impact of differences in applied statistical concepts……"	(Lyon 2008)
	"Extent to which data is consistent between organisations and over time allowing comparisons to be made"	(HIQA 2011)

	"…helps the user to interpret values correctly…"	(Redman 1997)
	Interpretation should be supported through the usage of appropriate technology media and tools. (rephrased definition)	(Redman 1997)
	"Data is not ambiguous if it allows only one interpretation…."	(Kimball and Caserta 2004)

Table 212: Interpretability

Characteristic	Information value	
Consolidated Definition	Quality information should provide a business value to the organization	
Candidate Themes	"…a measure the importance, value, and relevance of the data to the business"	(McGilvray 2008)
	Organization use shared data to create value as the organization matures (rephrased definition)	(Loshin 2001)
	"Value-added is related to how much data provide benefits for the users…."	(Scannapieco and Catarci 2002)
	"…..degree to which data will produce the desired business transaction or outcome"	(Stvilia et al. 2007)

Table 223: Information value

APPENDIX-E

Validation of DQ characteristics

1.0 Card sorting summary for goal 1

Dimension	Total Item placement Ratio	Kappa	Characteristic	Individual item placement ratio
Completeness	84.00%	0.63	Completeness of mandatory attributes	80.00%
			Completeness of optional attributes	90.00%
			Completeness of records	90.00%
			Data volume	80.00%
Availability & Accessibility	82.22%	0.64	Continuity of data access	86.67%
			Data maintainability	80.00%
			Data awareness	80.00%
			Ease of data access	73.33%
			Data punctuality	76.00%
			Data access control	100.00%
Currency	76.00%	0.3	Data age	74.29
			Data freshness	77.5%
Accuracy	86.40%	0.61	Accuracy to reference source	88.57%
			Accuracy to reality	91.11%

			Precision	80.00%
Validity	85.26%	0.69	Business rules compliance	80.00%
			Mata-data compliance	82.50%
			Standards and regulatory compliance	89.47%
			Statistical validity	100.00%
Reliability and credibility	91.11%	0.76	Source quality	80.00%
			Objectivity	100.00%
			Traceability	100.00%
Consistency	85.00%	0.65	Uniqueness	80.00%
			Non-redundancy	80.00%
			Semantic consistency	93.33%
			Value consistency	86.67%
			Format consistency	90.00%
			Referential integrity	80.00%
Usability & Interpretability	83.4%	0.62	Usefulness and relevance	85.71%
			Understandability	87.50%
			Appropriate presentation	84.29%
			Interpretability	86.96%
			Information value	86.67%

Table 1: Summary of the first round of card sorting for the goal 1

Dimension	Total Item placement Ratio	Kappa	Characteristic	Individual item placement ratio
Currency	90.00%	0.64	Data age	85.71%
			Data freshness	93.75%

Table 2: Summary of the second round of card sorting for the goal 1

2.0 Card sorting summary for goal 2

Dimension	Total Item placement Ratio	Kappa	Characteristic	Individual item placement ratio
Completeness	95.00%	0.86	Completeness of mandatory attributes	100.00%
			Completeness of optional attributes	100.00%
			Completeness of records	80.00%
			Data volume	100.00%
Availability & Accessibility	96.67%	0.91	Continuity of data access	100.00%
			Data maintainability	80.00%
			Data awareness	100.00%
			Ease of data access	100.00%

			Data punctuality	100.00%
			Data access control	100.00%
Currency	0%	1	Data age	0%
			Data freshness	0%
Accuracy	100%	1	Accuracy to reference source	100.00%
			Accuracy to reality	100.00%
			Precision	100.00%
Validity	95.00%	0.8	Business rules compliance	100.00%
			Mata-data compliance	100.00%
			Standards and regulatory compliance	80.00%
			Statistical validity	100.00%
Reliability and credibility	100.00%	1	Source quality	100.00%
			Objectivity	100.00%
			Traceability	100.00%
Consistency	100.00%	1	Uniqueness	100.00%
			Non-redundancy	100.00%
			Semantic consistency	100.00%
			Value consistency	100.00%
			Format consistency	100.00%
			Referential integrity	100.00%

Usability & Interpretability	100.00%	1	Usefulness and relevance	100.00%
			Understandability	100.00%
			Appropriate presentation	100.00%
			Interpretability	100.00%
			Information value	100.00%

Table 23: Summary of the first card sorting round for goal 2.

Dimension	Total Item placement Ratio	Kappa	Characteristic	Individual item placement ratio
Currency	100%	1	Data timeliness	100%
			Data freshness	100%

Table 4: Summary of the second round of card sorting for goal 2