An Analysis of Data Quality Dimensions

Vimukthi Jayawardene
School of Information Technology and Electrical Engineering
The University of Queensland
w.jayawardene@uq.edu.au

Shazia Sadiq
School of Information Technology and Electrical Engineering
The University of Queensland
shazia@itee.uq.edu.au

Marta Indulska
Business School
The University of Queensland
m.indulska@business.uq.edu.au

Abstract. Data quality (DQ) has been studied in significant depth over the last two decades and has received attention from both the academic and the practitioner community. Over that period of time a large number of data quality dimensions have been identified in due course of research and practice. While it is important to embrace the diversity of views of data quality, it is equally important for the data quality research and practitioner community to be united in the consistent interpretation of this foundational concept. In this paper, we provide a step towards this consistent interpretation. Through a systematic review of research and practitioner literature, we identify previously published data quality dimensions and embark on the analysis and consolidation of the overlapping and inconsistent definitions. We stipulate that the shared understanding facilitated by this consolidation is a necessary prelude to generic and declarative forms of requirements modeling for data quality.

1 Introduction

Data quality (DQ) has been widely researched over the past several decades (Sadiq et al., 2011) and by now has developed into a professional discipline (Yonke et al., 2011), with a prominent focus within organizational strategy. Advancements in data quality management have resulted in contributions from researchers as well as practitioners. A wealth of knowledge exists in the realm of the practitioner community (eg:- (Redman, 1997), (Loshin, 2001), (English, 2009), (McGilvray, 2008)), including initiatives such as the International Association of Information and Data Quality and its Information Quality Certification Program (www.iaidq.org). Although the diversity of contributions is valuable, some fundamental aspects of data quality management, in particular those relating to data quality dimensions, and consequently measures and metrics, have regressed into a level of disparity that does not support a shared understanding of the core knowledge of the discipline. In this paper, we address this area of concern and present the results of an analysis and consolidation of the main contributions of data quality dimensions stemming from research, vendor and practitioner communities.

In light of the management axiom "what gets measured gets managed" (Willcocks and Lester, 1996), dimensions of data quality signify a crucial management element in the domain of data quality. On these grounds, over the last two decades researchers and practitioners have suggested several classifications of data quality dimensions many of which have overlapping, and sometimes conflicting interpretations (eg. (Wang and Strong, 1996), (Redman, 1997), (English, 2009), (Loshin, 2001)). Despite the numerous classifications, few studies to date have embarked on an effort to consolidate these view-points. For example, Eppler (Eppler, 2006) provides a useful analysis of several of the existing classifications of data quality dimensions and recognizes sixteen mutually exclusive dimensions. This analysis is very useful, however the selection of classifications is incomplete and the coverage of the study does not span academic and practitioner contributions. Further, the basis for selection (or exclusion) of the classifications and their constituent dimensions has not been established. Yet, a comprehensive classification of the data quality dimensions is instrumental in the pursuit of developing a streamlined and unified set of dimensions that can assist in a shared understanding within the broader community and provide a basis for modeling of data quality requirements.

To bridge this gap in the body of knowledge, in this paper we undertake a study of existing body of knowledge on data quality dimensions. Our study spans both academic and industry contributions and incorporates both the semiotic and the product perspective on data quality. We believe that such an analysis is essential to create a shared understanding of the multiple and often conflicting interpretations of data quality dimensions as currently found in the broader research and practice body of knowledge. Broad convergence on the understanding and interpretations of a foundational concept such as data quality dimensions is a necessary prelude to the development of generic data quality requirements modeling and enforcement frameworks, particularly as the scale, availability and usage of data increases exponentially.

2 Background

2.1 Data & Data Quality

Before moving to the notion of data quality dimensions, let us revisit the first order questions arising from the background of this domain. What is data and what is data quality? In (Liebenau and Backhouse, 1990) Liebenau and Backhouse used modern semiotic theory principles developed by Morris (Morris, 1938) to explain data as "…language, mathematical or other symbolic surrogates which are generally agreed upon to represent people, objects, events and concepts". In its simplest form, data is a representation of objects or phenomena in the real world. Thus, when it comes to the discussion of quality of data, we can say that poor quality data is a result of poor representation of the real world. In the context of information systems, this representation of the real world is moderated by the needs of the system users, and 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. The semiotic perspective of data has been adopted by DQ researchers as well, for example, Price and Shanks (Price and Shanks, 2004) defined 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 the study of data and its quality. To date, however, the semiotic perspective has not become popular among researchers and practitioners. When it comes to supporting processes for managing DQ, a prominent approach, proposed by Wang (Wang, 1998), uses a product perspective of data as the underlying approach. By considering that 'information is processed data', Wang argues that information is analogous to products and data is analogous to raw materials in a typical product manufacturing process. Based on this argument, Wang considers information as a product of an information system and recognizes an information manufacturing process analogous to a product manufacturing process (Wang, 1998).

Since traditional product quality is a well explored concept, researchers have attempted to use product quality management models claiming 'fitness for use' as the principle for distinguishing good quality data and poor quality data. The 'fitness for use' approach is based on the general definition for quality introduced by Juran (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. In line with this perspective, Wang and Strong (Wang and Strong, 1996) have defined dimensions for data in a way that can represent customer expectations and can be used in creating a data specification.

2.1 Quality Dimensions

The term dimension is defined as "a measurable extent of a particular kind, such as length, breadth, depth, or height". Dimensions deal with measurements or, in other words, are quantifications of characteristics of an object or phenomenon. The essence of this definition is apparent in many classifications of dimensions in various quality domains. For example, Garvin (Garvin, 1987) defines eight dimensions of product quality, *viz.* performance, features, reliability, conformance, durability, serviceability, aesthetics, and perceived quality.

Table 1. Product quality dimensions (Garvin, 1987)

Dimension	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 image, advertising, or brand names.

From this classification it is evident that the dimensions lead to a measurable perspective of the product itself. The underlying idea is that once the specification for the product is created using these dimensions, product quality can be measured by evaluating the extent to which the prescribed values for the dimensions 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 perceptional measures (perceived quality, serviceability, aesthetics) facilitating a judgment of the *usage* of the product that depends on its users.

Similarly Russell and Taylor (Russell and Taylor, 2003) define the dimensions of service quality as time and timeliness, completeness, courtesy, consistency, accessibility and convenience, accuracy, and responsiveness.

Table 2. Service quality dimensions by Russell and Taylor (Russell and Taylor, 2003).

Dimension	Definition
Time & Timeliness	Customer wait 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

In this classification the dimensions have been defined using the *declarative* perspective to explain the service (completeness, accuracy, time and timeliness) as well as the perceptional perspective, facilitating the perceptional judgment of the *usage* the service (courtesy, consistency, accessibility and timeliness, responsiveness).

Thus, we observe that studies on product and service quality consider both the declarative and usage perspectives. These declarative and usage perspectives similarly play a fundamental role in identifying and defining DQ dimensions. Hence in this paper we use the following two criteria to identify and analyze DQ dimensions, and exclude published definitions that do not fall into the two categories of dimensions:

Declarative Perspective (D): Focuses on user independent characteristics of data which explains data itself, or in other words data definitions by meta-data, schema standards and business rules imposed based on the operational aspects of organizations.

Usage Perspective (U): Focuses on user dependent characteristics of data related to effective and efficient data creation and usability that contribute to users' judgment about the data's fitness for use.

2.2 Granularity of data

When reasoning about the characteristics of data quality dimension, it is also important to consider at which data granularity level they are applicable. Otherwise practically it is hard to use the dimensions in managing data quality. In literature where data quality dimensions are used in assessing data quality (Batini et al., 2009, Pipino et al., 2002, Lee et al., 2002, Eppler and Muenzenmayer, 2002), the authors do not explicitly mention the granularity of data and consider it to be a context dependent fact due to the broad nature of the definition of dimensions (eg. Completeness: The extent to which data is not missing and is of sufficient breadth and depth for the task at hand). But we believe when it comes to the characteristics of data quality dimensions, a clear granularity level can be defined for each characteristic so that it will be practically useful.

Even and Shankaranarayanan (2005) provide a valuable insight into granularity levels of data quality dimensions by considering a hierarchy of data as, data items (elements), data records, datasets, databases and organizational database collections when describing data quality dimensions and metrics.

In studying data quality dimensions we observe that some characteristics (e.g.: completeness of records, understandability) are applicable at higher granularity levels, such as a record or a collection of records. Further we observe that granularity may depend on the type of the characteristic (D/U). Declarative characteristics are primarily defined on data *elements* and *records* while the usage characteristics may be defined on any arbitrary abstraction of data elements and records retrieved from the same relation or from different relations are considered, that is an information *object*. Thus 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 records used to accomplish a task.

We will return to this concept and associate the granularity levels with each defined characteristic of the consolidated data quality dimensions in section 4.

3 Approach

In our review of the classifications, we observe that most approaches appear to be influenced by the classification of Wang and Strong (Wang and Strong, 1996), while also incorporating individual experience. Due to the contextual nature of many studies, these classifications are quite diverse. This diversity, while important, makes it difficult to build a unified and shared understanding of the DQ domain from a dimension and consequently measurement perspective. Accordingly, a synthesis of the various definitions is required to cater for the multiplicity of DQ dimensions. For this analysis we identified four relevant sources of data quality dimension classifications, ensuring coverage of the academic, practitioner, vendor and business communities, and developed a four-step methodology as described below.

First we reviewed existing literature and identified prominent DQ dimension classifications that fit the following perspectives:

a) Perspectives from industry practitioners involved in consulting on large data quality projects and contributing to DQ body of knowledge by publishing books and an apparent prominence in industry. Relevant sources within the practitioner perspective were identified by examination of citations in public forums and professional training programs by professional bodies such as DAMA ((DAMA)) and IAIDQ (Watson-Manheim et al., 2002). Within these sources we identified several prominent contributions (Redman, 1997), (English, 2009), (McGilvray, 2008), (Loshin, 2001), (Kimball and Caserta, 2004).

b) Perspectives from market leaders of DQ management tools, as identified by Gartner's Magic Quadrant (Friedman, 2012). These market leaders include: SAP (G. Gatling, 2007), IBM (B. Byrne, 2008), and Informatica (Loshin, 2006).

c) Data Quality standards, as identified by ISO 8000 - a standard for data quality (ISO, 2012).

d) Perspectives from organizations that have recognized the importance of DQ and developed their own DQ frameworks to manage DQ. Although many organizations conduct DQ projects, only few have made available their DQ dimensions publicly with sufficient level of information suitable for an analysis. In our search we found Bank of England (Lyon, 2008) and Health Information and Quality Authority (HIQA) (HIQA, 2011), the latter representing an international study on DQ practices of healthcare organizations in England, Wales, Canada and New Zealand.

e) Perspectives from academia with rigorous research based findings and a high level of citations: In our earlier work [31] we analysed DQ research contributions over the last 2 decades and created a bibliographic database[1] of over one thousand publications. We used this resource to identify research articles that focus on data quality criteria or dimensions. Consequently, we identified 36 publications focussing on DQ dimensions in sufficient depth and breath. Based on citation analysis, the most prominent DQ dimensions classification was developed by Wang & Strong (Wang and Strong, 1996), with the majority of other

[1] This database can be accessed through http://dqm.cloud.itee.uq.edu.au/

classifications being derivatives of this original work. On this basis we selected the original work by Wang and strong (Wang and Strong, 1996) and three additional classifications that have significant and contrasting differences (Price and Shanks, 2005), (Eppler, 2006), (Stvilia et al., 2007) Scannapieco and Catarci 2002.

Altogether we selected fourteen publications that fairly represent the above four perspectives, and thus provide a broad scope for the analysis.

In the second stage of the analysis, the 16 papers (or parts thereof, in case of books) were loaded into NVIVO[2] – a qualitative data analysis tool. We employed a multi-coder approach to facilitate a rigorous identification of the dimensions within the text of the 16 documents. The text was reviewed and individually coded by two researchers to ensure all dimensions were identified. Each coder independently coded the relevant text in NVIVO[2], creating a node for each dimension and its definition. The coding structures were then consolidated between the two researchers to arrive at a final coding (after resolving coding disagreements through discussion) that identified 129 terms as dimensions with 189 definitions. It was noticed that these terms and definitions have many overlaps and contrasts. Different authors have used the same term (as a dimension) to refer to contrasting aspects of data quality where as some authors have used different terms to refer to the same aspect of data quality. Hence it was apparent that among the 189 definitions there are many common themes and a necessity exists to consolidate these definitions towards reaching a consensus in this domain. From this coding process we were able to identify the contextual meaning of the dimensions, based on which we could elicit the underlying theme behind each dimension.

In the third step, we analyzed the definitions of each dimension with respect to their reflection of a declarative or a Usage characteristic as per the theoratical lens explained in section 2.1 above. In particular, for each definition, two researchers individually coded the definitions as being usage (U), declarative (D), a mixture of both (D/U) or neither (X). The aim of this task was to refine the list of dimensions by eliminating those that do not represent characteristics of data or users' view of data. The independent ratings were evaluated using Cohen's Kappa , with a result of 0.81, indicating high confidence about raters' agreement (Carletta, 1996). Coding disagreements were then discussed between the three researchers until a consensus was reached. In this analysis, out of 189 definitions, 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 by Redman (1997) and Loshin (2001) respectively, which are focused on the utilization of disk space and memory space of computers, and 'Stewardship' (Loshin, 2001) which is focused on assigning the people responsibility for data, and

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

represents a management function rather than a declarative or usage perspective of data quality. As a result of this step we identified 186 definitions which confirm to the theoretical lens as data quality characteristics. In the final step the researchers analyzed the themes in every definitions and created a classification of dominant themes of data quality characteristics. Braun and Clarke (2006) has explained the importance of creating thematic maps in consolidating themes in qualitative data. Hence one researcher clustered the definitions based on evident themes and created a thematic map. In this effort every definition was analyzed for the theme behind it and similar themes were clustered together and created consolidated themes and termed them as data quality characteristics. Then each characteristic was given a definition considering the original definitions by authors and a representative term considering the original dimension names given by the authors. Then these consolidate themes were further clustered into broader clusters and the clusters were termed as data quality dimensions and a representative term was given for each cluster. Following this step, two researchers individually reviewed the thematic map. The three researchers then met to finalize the clustering, definitions and representative terms, leading to an agreement of eight main clusters (dimensions) and thirty three data quality characteristics spread across the eight clusters.

4 Analysis & Results

The sixteen sources of dimensions selected for this study revealed 127 dimensions. These dimensions are expressed using one or more representative terms, together with the authors' own definitions. It should be noted that some dimensions were referred to by the same term in different classifications; in the lists presented below such terms are presented together.

Following the 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 these clusters are presented in detail with the individual terms and the definitions given by various authors. Further each individual definition is classified into declarative perspective (D) or usage perspective (U) based on the contextual meaning of the author's definition.

Completeness:

Table 3: Definitions relating to completeness.

Ability to represent null values	Ability to distinguish neatly (without ambiguities) null and default values from applicable values of the domain. (Redman, 1997)	D
Null values	A null value is a missing value. However, a value that is missing may provide more information than one might think because there may be different reason that it is missing. A null value might	D

	actually represent an unavailable value, an attribute that is not applicable for this entity, or no value in the attribute’s domain that correctly classifies this entity. Of course, the value may actually be missing (Loshin, 2001)	
Representation of null values	When the null value (or absence of a value) is required for an attribute, there should be a recognizable form for presenting that null value that does not conflict with any valid values. (Loshin, 2001)	D
Value existence	A given data element (fact) has a full value stored for all records that should have a value (English, 2009)	D
Completeness	Completeness refers to the degree to which values are present in a data collection, as for as an individual datum is concerned, only two situations are possible: Either a value is assigned to the attribute in question or not. In the latter case, null, a special element of an attribute’s domain can be assigned as the attribute’s value. Depending on whether the attribute is mandatory, optional, or inapplicable, null can mean different things. (Redman, 1997)	D
	Completeness refers to the expectation that certain attributes are expected to have assigned values in a data set. Completeness rules can be assigned to a data set in three levels of constraints: 1. Mandatory attributes that require a value 3. Inapplicable attributes (such as maiden name for a single male), which may not have a value.2. Optional attributes, which may have a value (Loshin, 2001)	D
	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)	D
	Determined the extent to which data is not missing. For example, an order is not complete without a price and quantity (G. Gatling, 2007)	D
	An expectation of completeness indicates that certain attributes should be assigned values in a data set. Completeness rules can be assigned to a data set in three levels of constraints:1. Mandatory attributes that require a value, 2. Optional attributes, which may have a value based on some set of conditions, and 3. Inapplicable attributes,	D

	(such as maiden name for a single male), which may not have a value.(Loshin, 2006)	
	Completeness of data refers to the extent to which the data collected matches the data set that was developed to describe a specific entity. Monitoring for incomplete lists of eligible records or missing data items will identify data quality problems.(HIQA, 2011)	U
	Degree of presence of data in a given collection (Scannapieco and Catarci, 2002)	U
Mapped completely	Every real-world phenomenon is represented (Price and Shanks, 2005)	D
Appropriate amount of data	The quantity or volume of available data is appropriate (Wang and Strong, 1996)	U
Comprehensiveness	Is the scope of information adequate? (not too much nor too little) (Eppler, 2006)	D
Data Coverage	A measure of the availability and comprehensiveness of data compared to the total data universe or population of interest (McGilvray, 2008)	
Value completeness	A given data element (fact) has a full value stored for all records that should have a value (English, 2009) (not included in survey)	D
Record existence	A record exists for every Real-World Object or Even the Enterprise needs to know about (English, 2009)	U
Complete	Domain Level: Data element is 1. Always required be populating and not defaulting; or 2. Required based on the condition of another data element. Entity Level: The required domains that comprise an entity exist and are not defaulted in aggregate.(B. Byrne, 2008)	D
Data completeness	Quality of having all data that existed in the possession of the sender at time the data message was created (ISO, 2012)	U

Completeness is considered in a broad sense and contains several themes. Namely, it focuses on handling of null values, representing real world objects without omission and maintaining right volume of data for intended usage can be considered as dominating themes.

Several authors have pointed out that null values should be given special consideration in managing data quality. For example, *"ability to distinguish neatly (without ambiguities) null and default values from applicable values of the domain"* (Redman, 1997) Null values have multiple implications such as unknown, missing or not applicable values, thus causing ambiguity in their interpretation.

Different granularity levels (field, record, and table) may define completeness in different ways. For example, "*data are of sufficient depth, breath and scope for the task at hand"* (Wang and Strong, 1996), and *"knowledge workers have all the facts they need to perform their processes or make their decisions"* (English, 2009). Thus, a snapshot view of the database may not indicate if the data is complete or not. Completeness cannot be judged merely by looking at the existing records of a database - there can be missing data objects altogether. This problem relates back to the fundamental notion of closed world vs. open world assumptions for digital information systems (Batini and Scannapieco, 2006). For example, *"a record exists for every Real-World Object or the Event the Enterprise needs to know about"* (English, 2009) and *"every real-world phenomenon is represented"* (Price and Shanks, 2005).

In light of the above themes we identified and consolidated the main themes in the above definitions and thereby defined the following data quality characteristics of completeness.

Table 4: Characteristics of completeness.

Characteristic	Description	Granularity	Type
Completeness of mandatory values	The attributes which are necessary for a complete representation of a real world entity must contain values and cannot be null	E	D
Completeness of optional values	Non-mandatory 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:

Table 5: Dimensions relating to Availability & Accessibility.

Accessibility	Data are available or easily or quickly retrieved (Wang and Strong, 1996)	U
	Is there a continuous and unobstructed way to get to the information? (Eppler, 2006)	U
	Accessibility of data refers to how easily it can be accessed; the awareness of data users of what data is being collected and knowing where it is located. (HIQA, 2011)	U

	Speed and ease of locating and obtaining an information object relative to a particular activity (Stvilia et al., 2007)	U
	Accessibility expresses how much data are available or quickly retrievable. (Scannapieco and Catarci, 2002)	
Accessibility and clarity	Accessibility refers to the physical conditions in which users can obtain data Clarity refers to the data’s information environment including appropriate metadata (Lyon, 2008)	U
Accessibility timeliness	The characteristic of getting or having the Information when needed by a process or Knowledge Worker (English, 2009)	U
Availability	The Characteristic of the Information being accessible when it is needed (English, 2009)	U
	availability of a data source or a system (Scannapieco and Catarci, 2002)	
Ease of Use and maintainability	A measure of the degree to which data can be accessed and used and the degree to which data can be updated, maintained, and managed (McGilvray, 2008)	U
Security	Is the information protected against loss or unauthorized access? (Eppler, 2006)	U
	The extent to which information is protected from harm in the context of a particular activity (Stvilia et al., 2007)	U
Timeliness and punctuality	Timeliness reflects the length of time between availability and the event or phenomenon described. Punctuality refers to the time lag between the release date of data and the target date when it should have been delivered (Lyon, 2008)	D
Maintainability	Can all of the information be organized and updated on an on-going basis? (Eppler, 2006)	U
Speed	Can the infrastructure match the user’s working pace? (Eppler, 2006)	U
Timeliness	Is the information processed and delivered rapidly without delays? (Eppler, 2006)	U
	Timeliness refers to the time expectation for accessibility and availability of information. Timeliness can be measured as the time between when information is expected and when it is readily available for use. For example, in the financial industry, investment product pricing data is often provided by third-	U

	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 can be defined and compliance with those timeliness constraints can be measured (Loshin, 2006)	
Accessible	Data is easy and quick to retrieve (Price and Shanks, 2005)	U
Access Security	Access to data can be restricted and hence kept secure (Wang and Strong, 1996)	D
Secure	Data is appropriately protected from damage or abuse (including unauthorized access, use, or distribution) (Price and Shanks, 2005)	U
Reliability	The frequency of failures of a system, its fault tolerance (Scannapieco and Catarci, 2002)	U

In this cluster, a broad range of definitions combining timeliness, availability and accessibility of data can be observed. Availability of data when needed and the security perspective of data are the dominating aspects of this cluster.

In existing classifications timeliness and currency are two terms that have a significant interplay and overlap. However, we observe some fundamental differences in their interpretation (timely availability of data vs. correct aging of data\freshness of data) when analysing the various definitions and hence currency, together with other related dimensions, is a cluster in and of itself.

On-time availability of data is a major consideration of this cluster, as evidenced by several closely related definitions. For example, (Loshin, 2006) consider that timeliness "*refers to the time expectation for accessibility and availability of information*". Similarly, (English, 2009) discuss *"the characteristic of getting or having the Information when needed by a process or Knowledge Worker"*. In both of these definitions the focus is on the efficient retrieval of data when needed, whereas (McGilvray, 2008) broadens the focus towards efficient database management: "*a measure of the degree to which data can be accessed and used and the degree to which data can be updated, maintained, and managed"*.

On the other hand, several authors have aligned accessibility of data with security giving more prominence to the security perspective of data – e.g. *"access to data can be restricted and hence kept secure"* (Wang and Strong, 1996) and *"is the information protected against loss or unauthorized access?"* (Eppler, 2006).

Based on the above definitions we identified and consolidated the main themes in the above definitions and thereby defined the following data quality characteristics of completeness.

Table 6: Characteristics of availability and accessibility.

Characteristic	Description	Granularity	Type
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 unauthorised access.	IO	U

Currency:

Table 7: Dimensions relating to Currency

Currency	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 values; a datum is outdate at time t if it is incorrect at t but was correct at some time preceding t. currency refers to a degree to which a datum in question is up-to-date. (Redman, 1997)	D
	The "age" of the data is correct for the Knowledge Worker's purpose . Purposes such as inventory control for Just-in-Time Inventory require the most current data. Comparing sales trends for last period to period one-year ago requires sales data from respective periods.(English, 2009)	U
	Is the information upto-date and not obsolete? (Eppler, 2006)	U

	Currency refers to the degree to which information is current with the world that it models. Currency can measure how “up-to-date” information is, and whether it is correct despite possible time-related changes. Data currency may be measured as a function of the expected frequency rate at which different data elements are expected to be refreshed, as well as verifying that the data is up to date. For example, one might assert that the contact information for each customer must be current, indicating a requirement to maintain the most recent values associated with the individual’s contact data (Loshin, 2006)	U
	The age of an information object (Stvilia et al., 2007)	U
Currency/Timeliness	Currency refers to the degree to which information is current with the world that it models. Currency can measure how up to date information is and whether is it correct despite possible time-related changes. Timeliness refers to the time (Loshin, 2001)	D
Data Decay	A measure of the rate of negative change to the data (McGilvray, 2008)	D
Timely	Domain Level: The data element represents the most current information resulting from the output of a business event. Entity Level: The entity represents the most current information resulting from the output of a business event. (B. Byrne, 2008)	U
	The currency (age) of the data is appropriate to its use. (Price and Shanks, 2005)	U
Volatility	The amount of time the information remains valid in the context of a particular activity (Stvilia et al., 2007)	U
	How long data remains valid (Scannapieco and Catarci, 2002)	U
Timeliness and availability	A measure of the degree to which data are current and available for use as specified and in the time frame in which they are expected (McGilvray, 2008)	U
Timeliness	Data is accurate if it is up to date – antiexample: “Current president of the USA: Bill Clinton”. (Kimball and Caserta, 2004)	U

	The age of the data is appropriate for the task at hand (Wang and Strong, 1996)	U
	Determines the extent to which data is sufficiently up-to-date for the task at hand. For example, hats, mittens, and scarves are in stock by November (G. Gatling, 2007)	U
	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. Data should be made available at whatever frequency and within whatever timeframe is needed to support decision making. (HIQA, 2011)	U
	Timeliness can be defined in terms of currency (how recent data are)(Scannapieco and Catarci, 2002)	U

With change being a constant phenomenon in the real world, it is not surprising that most interpretations of data currency are based on the most up-to-date reality. Hence in this cluster the main consideration is managing the right age of data for the intended purposes. For example, (English, 2009) discuss age of data with respect to a user’s need: “*the age of the data is correct for the Knowledge Worker’ purpose”. Similarly, (B. Byrne, 2008, Price and Shanks, 2005) consider the importance of currency:* “*the data element represents the most current information resulting from the output of a business event”.* Numerous other authors also share this vision, with (Loshin, 2001) considering that “*currency refers to the degree to which information is current with the world that it models”*, and (Redman, 1997) agreeing that *“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 values”.* Hence the focus of these definitions is on the prevention of the negative consequences of outdated data being used for the task at hand.

Some changes to data are outside the control of the system (e.g. market statistics) where as some data gets obsolete due to lack of proper system updates. Hence both these cases need to be taken care of with right policies and procedures to refresh the data at suitable times. Several authors have defined timeliness (HIQA, 2011, G. Gatling, 2007, Wang and Strong, 1996, Kimball and Caserta, 2004) with an emphasis on aging of data with reference to users’ perception towards catering to the task at hand while others have emphasized on policies and procedures to maintain the right aging of data for the task.

In our analysis the following two characteristics were identified within this cluster.

Table 8: Characteristics of data currency.

Characteristic	Description	Granularity	Type
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:

Table 9: Dimensions relating to accuracy

Accuracy	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 (or maybe only a) correct one for the entity e and the attribute a. In some cases, v' is referred to as the standard. If the datum's value v coincides value v', the datum is said to be correct. (Redman, 1997)	U
	Data accuracy refers to the degree with which data values agree with an identified source of correct information. There are different sources of correct information: database of record, a similar, corroborative set of data values from another table, dynamically computed values, the result of a manual workflow, or irate customers. (Loshin, 2001)	U
	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)	U
	The extent to which data are correct reliable and certified free of error (Wang and Strong, 1996)	U
	Is the information precise enough and close enough to reality? (Eppler, 2006)	U
	Determines the extent to which data objects correctly represent the real-world values for which they were designed. For example, the sales orders for the Northeast region must be assigned a Northeast sales representative	U

	(G. Gatling, 2007)	
	The data value correctly reflects the real-world condition. (B. Byrne, 2008)	U
	Data accuracy refers to the degree with which data correctly represents the “real-life” objects they are intended to model. In many cases, accuracy is measured by how the values agree with an identified source of correct information (such as reference data). There are different sources of correct information: a database of record, a similar corroborative set of data values from another table, dynamically computed values, or perhaps the result of a manual process (Loshin, 2006)	U U
	Accuracy of data refers to how closely the data correctly captures what it was designed to capture. Verification of accuracy involves comparing the collected data to an external reference source that is known to be valid. Capturing data as close as possible to the point of activity contributes to accuracy. The need for accuracy must be balanced with the importance of the decisions that will be made based on the data and the cost and effort associated with data collection. If data accuracy is compromised in any way then this information should be made known to the data users. (HIQA, 2011)	U
	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)	U
	Degree of correctness of a value when comparing with a reference one. (Scannapieco and Catarci, 2002)	
	Closeness of agreement between a property value and the true value (value that characterizes a characteristic perfectly defined in the conditions that exists when the characteristic is considered. (ISO, 2012)	U
Accuracy to reality	The data correctly reflects the Characteristics of a Real-World Object or Event being described. Accuracy and Precision represent the highest degree of inherent Information Quality possible (English, 2009)	U
Accuracy to surrogate source	The data agrees with an original, corroborative source record of data, such as a notarized birth certificate, document, or unaltered electronic data received from a party outside the control of the organization that is demonstrated to be a reliable source. (English, 2009)	U

Correctness	Data is correct if it conveys a lexically, syntactically and semantically correct statement – e.g.,the following pieces of information are not correct:“Germany is an African country” (semantically wrong);Book.title: ‘De la Mancha Don Quixote’ (syntactically wrong); UK’s Prime Minister: ‘Toni Blair’ (lexicallywrong). (Kimball and Caserta, 2004)	D
Precision	Data values are correct to the right level of detail or granularity, such as price to the penny or weight to the nearest tenth of a gram (English, 2009)	U
Phenomena mapped correctly	Each identifiable data unit maps to the correct real-world phenomenon. (Price and Shanks, 2005)	U
Conciseness	Is the information to the point, void of unnecessary elements? (Eppler, 2006)	D
Properties mapped correctly	Non-identifying (i.e. non-key) attribute values in an identifiable data unit match the property values for the represented real-world phenomenon (Price and Shanks, 2005)	U
Precision/completeness	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)	D
	The extent to which an information object matches the precision and completeness needed in the context of a given activity (Stvilia et al., 2007)	D
Mapped meaningfully	Each identifiable data unit represents at least one specific real-world phenomenon (Price and Shanks, 2005)	U
Mapped unambiguously	Each identifiable data unit represents at most one specific real-world phenomenon (Price and Shanks, 2005)	U
Verifiability	The extent to which the correctness of information is verifiable or provable in the context of a particular activity (Stvilia et al., 2007)	U
Accuracy/Validity	The extent to which information is legitimate or valid according to some stable reference source such as a dictionary or set of domain constraints and norms (soundness) (Stvilia et al., 2007)	U
Reliability	Reliability of data refers to the extent to which data is collected consistently over time and by different organisations either manually or electronically. (HIQA, 2011)	U
Format precision	The set S should be sufficiently precise to distinguish among elements in the domain that must be	U

	distinguished by users. This dimension makes clear why icons and colors are of limited use when domains are large. But problems can and do arise for the other formats as well, because many formats are not one-to-one functions. For example, if the domain is infinite (the rational numbers, for example), then no string format of finite length can represent all possible values. The trick is to provide the precision to meet user needs.(Redman, 1997)	
	The degree of precision of the presentation of an attribute’s value should reasonably match the degree of precision of the value being displayed. The user should be able to see any value the attributer may take and also be able to distinguish different values. (Loshin, 2001)	U

Accuracy is the first and foremost requirement that many users expect from data. Hence it is not surprising that many authors have a common understanding of accuracy. Accuracy is evaluated by comparing data with their original sources in reality. For example, “*data accuracy refers to the degree with which data values agree with an identified source of correct information”* (Redman, 1997, Loshin, 2001). Hence in fact accuracy in this sense is related to the process of data creation.

The level of accuracy is another aspect which is driven by the consumer need, for example, “*data values are correct to the right level of detail or granularity, such as price to the penny or weight to the nearest tenth of a gram”* (English, 2009). Conciseness (Eppler, 2006), on the other hand, which has a component relating to user opinion (*“... is the information to the point, void of unnecessary elements....”)* is a perceptual measure.

The following data quality characteristics were identified in this cluster.

Table 10: Characteristics of accuracy

Characteristic	Description	Granularity	Type
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:

Table 11: Dimensions relating to Validity

Business rule validity	Data values conform to the Specified Business Rules (English, 2009)	D
Derivation validity	A derived or calculated data value is Produced Correctly according to a specified Calculation Formula or set of Derivation Rules (English, 2009)	D
Validity	Validity of data refers to data that has been collected in accordance with any rules or definitions that are applicable for that data. This will enable benchmarking between organisations and over time.(HIQA, 2011)	D
Integrity	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.(G. Gatling, 2007)	D
Value validity	A data value is a Valid Value or within a specified range of valid values for this data element (English, 2009)	D
Conformance	This dimension refers to whether 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. Each column has numerous metadata attributes associated with it: its data type, precision, format patterns, use of a predefined enumeration of values, domain ranges, underlying storage formats, etc. (Loshin, 2006)	D
Valid	Data element passes all edits for acceptability and is free from variation and contradiction based on the condition of another data element (a valid value combination). (B. Byrne, 2008)	D
Data Specifications	A measure of the existence, completeness, quality, and documentation of data standards, data models, business rules, metadata, and reference data (McGilvray, 2008)	D
Representation consistency	Representation consistency refers to whether physical instances of data are in record with their formats. For example, an EMPLOYEE's salary cannot be represented "$AXT," as there is (or should be) no such element in S. One would often like to know whether a physical instance is the proper representation for the intended (correct) value. But in practice this is rarely	D

	possible, as the intended value is conceptual and not known. So one is left with the issue of whether the representation conflicts with S. (Redman, 1997)	
	This dimension refers to whether instances of data are represented in a format that is consistent with the domain of values and with other similar attribute values. For example, the display of time in a non-military (12-hour) format may be confusing if all other instances of times in the system are displayed in the 24-hour military format (Loshin, 2001)	D
Signage Accuracy and Clarity	Signs and other Information-Bearing Mechanisms like Traffic Signals should be standardized and universally used across the broadest audience possible.(English, 2009)	D
Allowing access to relevant metadata	Appropriate metadata is available to define, constrain, and document data (Price and Shanks, 2005)	D
Coherence	Coherence of data refers to the internal consistency of the data. Coherence can be evaluated by determining if there is coherence between different data items for the same point in time, coherence between the same data items for different points in time or coherence between organisations or internationally. Coherence is promoted through the use of standard data concepts, classifications and target populations. (HIQA, 2011)	U
	Coherence of statistics is their adequacy to be reliably combined in different ways and for various uses. (Lyon, 2008)	U
Conformity	Determines the extent to which data conforms to a specified format. For example, the order date must be in the format YYYY/MM/DD. (G. Gatling, 2007)	D
Definition Conformance	Data values are consistent with the Attribute (Fact) definition (English, 2009)	D
Semantic definition	The data element has a commonly agreed upon enterprise business definition and calculations (B. Byrne, 2008)	D
Accuracy	Accuracy in the general statistical sense denotes the closeness of computations or estimates to the exact or true values. (Lyon, 2008)	U
Understood	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.	D

	The metadata of the entity clearly states or defines the purpose of the entity and its required attributes/domains (B. Byrne, 2008)	

The main consideration in this cluster is the conformance of data to business rules, For example, *“validity of data refers to data that has been collected in accordance with any rules or definitions that are applicable for that data”* (HIQA, 2011). It also refers to conformance to metadata: “*Data values are consistent with the Attribute (Fact) definition”* (English, 2009). According to (McGilvray, 2008), validity *“A measure of the existence, completeness, quality, and documentation of data standards, ”*, emphasizes that adherence to data standards is another aspect of validity. In this cluster the following themes were identified as quality characteristics.

Table 12: Characteristics of validity

Characteristic	Description	Granularity	Type
Business rules compliance	Calculations on data must comply with business rules	E	D
Meta-data compliance	Data should comply with its metadata	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 and Credibility:

Table 13: Dimensions relating to Reliability and Credibility.

Believability	Data are accepted or regarded as true real and credible (Wang and Strong, 1996)	U
	Believability is the extent to which data are accepted or regarded as true, real and credible (Scannapieco and Catarci, 2002)	U
Source	The source of information (1) guarantees the quality	U

Quality and Security Warranties or Certifications	of information it provides with remedies for non-compliance; (2) documents its certification in its Information Quality Management capabilities to capture, maintain, and deliver Quality Information; (3) provides objective and verifiable measures of the Quality of Information it provides in agreed-upon Quality Characteristics; and (4) guarantees that the Information has been protected from unauthorized access or modification (English, 2009)	
Reputation	Data are trusted or highly regarded in terms of their source and content (Wang and Strong, 1996)	U
Objectivity	Data are unbiased and impartial (Wang and Strong, 1996)	U
	Objectivity is the extent to which data are unbiased (unprejudiced) and impartial.(Scannapieco and Catarci, 2002)	U
Presentation Objectivity	The degree to which Information is presented without bias, enabling the Knowledge Worker to understand the meaning and significance without misinterpretation. (English, 2009)	U
Perceptions	Perceptions of the syntactic and semantic criteria defined earlier (Price and Shanks, 2005)	U
Traceability	Is the background of the information visible (author, date etc.)? (Eppler, 2006)	U
Verifiability	The extent to which the correctness of information is verifiable or provable in the context of a particular activity (Stvilia et al., 2007)	U
Authority	The degree of reputation of an information object in a given community or culture (Stvilia et al., 2007)	U
Enterprise Agreement of Usage	The notion of abstracting information into a data domain implies that there are enough users of the same set of data that it makes sense to manage their own versions. The dimension of enterprise agreement of usage measures the degree to which different organizations conform to the usage of the enterprise data domain of record instead of relying on their own data set. (Loshin, 2001)	U
Data Provanance	A data provenance record can include information about creation, update, transcription, abstraction, validation and transforming ownership of data (ISO, 2012)	U
Credibility	How much information is accurate, complete, consistent and non-fictiousness (Scannapieco and Catarci, 2002)	U

Reputation	Reputation is the extent to which data are trusted or highly regarded in terms of their source or content (Scannapieco and Catarci, 2002)	U

The main focus of the definitions in cluster is assurance of the trustworthiness of data. Aspects relating to confidence of data are emphasized in (McGilvray, 2008) under the dimension of Perception Relevance and Trust: "*a measure of the perception of and confidence in the quality of the data; the importance, value, and relevance of the data to business need".* Similarly in (Wang and Strong, 1996), under objectivity, authors relate to the credibility of data: *"data are unbiased and impartial".* However, under believability (Wang and Strong, 1996) emphasizes the credibility and truthfulness of data by referring to the original data sources through lineage and provenance.

English (English, 2009) presents the credibility and trustworthiness of data by referring to some broader aspects: "*The source of information (1) guarantees the quality of information it provides with remedies for non-compliance; (2) documents its certification in its Information Quality Management capabilities to capture, maintain, and deliver Quality Information; (3) provides objective and verifiable measures of the Quality of Information it provides in agreed-upon Quality Characteristics; and (4) guarantees that the Information has been protected from unauthorized access or modification".*

In this cluster majority of the dimensions have been defined based on user judgement regarding the trustworthiness of data and hence belong to the perceptional perspective. The dimensions verifiability and traceability however has a declarative component in its definition, as it refers to a mechanism in facilitating the correctness of data thereby improving the credibility, that is *".... the extent to which the correctness of information is verifiable or provable in the context of a particular activity"* (Stvilia et al., 2007), "*Is the background of the information visible.*" (Eppler, 2006). The Following characteristics were identified in this cluster.

Table 14: Characteristics of reliability and credibility

Characteristic	Description	Granularity	Type
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:

Table 15: Dimensions relating to Consistency.

Duplication /Non-duplication	A measure of unwanted duplication existing within or across systems for a particular field, record, or data set (McGilvray, 2008)	D
	There is only one record in a given data store that represents a Single Real-World Object or Event (English, 2009)	D
Uniqueness/Unique	Determines the extent to which the columns are not repeated. (G. Gatling, 2007)	D
	The entity is unique — there are no duplicate values (B. Byrne, 2008).	D
	Asserting uniqueness of the entities within a data set implies that no entity exists more than once within the data set and that there is a key that can be used to uniquely access each entity. For example, in a master product table, each product must appear once and be assigned a unique identifier that represents that product across the client applications (Loshin, 2006)	D
Equivalence of redundant or distributed data	Data about an object or event in one data store is semantically Equivalent to data about the same object or event in another data store (English, 2009)	D
Consistency/Consistent	Consistency, in popular usage, means that two or more things do not conflict with one another. This usage extends reasonably well to data values, although a bit of added discipline is desired. (Redman, 1997)	D
	Consistency can be curiously simple or dangerously complex. In its most basic form, consistency refers to data values in one data set being consistent with values in another data set. Two data values drawn from separate data sets may be consistent with each other, yet both can be incorrect (Loshin, 2001)	D
	Is the information free of contradictions or convention breaks? (Eppler, 2006)	D
	Data is consistent if it doesn’t convey heterogeneity, neither in contents nor in form –	D

	antiexamples: Order.Payment. Type = ‘Check’; Order. Payment. CreditCard_Nr = 4252… (inconsistency in contents); Order.requested_by: ‘European Central Bank’;Order.delivered_to: ‘ECB’ (inconsistency in form,because in the first case the customer is identified by the full name, while in the second case the customer’s acronym is used). (Kimball and Caserta, 2004)	
	Determines the extent to which distinct data instances provide nonconflicting information about the same underlying data object. For example, the salary range for level 4 employees must be between $40,000 and $65,000 (G. Gatling, 2007)	D
	Domain Level: The data values persist from a particular data element of the data source to another data element in a second data source. Consistency can also reflect the regular use of standardized values, articularly in descriptive elements. Entity Level: The entity’s domains and domain values either persist intact or can be logically linked from one data source to another data source. Consistency can also reflect the regular use of standardized values particularly in descriptive domains (B. Byrne, 2008)	D
	In its most basic form, consistency refers to data values in one data set being consistent with values in another data set. A strict definition of consistency specifies that two data values drawn from separate data sets must not conflict with each other, although consistency does not necessarily imply correctness (Loshin, 2006)	D
	Consistency among different data values (e.g. Sex and Name).(Scannapieco and Catarci, 2002)	D
Referential integrity	Assigning unique identifiers to objects (customers, products, etc.) within your environment simplifies the management of your data, but introduces new expectations that any time an object identifier is used as foreign keys within a data set to refer to the core representation, that core representation actually exists. (Loshin, 2006)	D

Consistency and Synchronization	A measure of the equivalence of information stored or used in various data stores, applications, and systems, and the processes for making data equivalent (McGilvray, 2008)	D
Structured Valued Standardization	Structured Attributes like dates, time, telephone number, tax ID number, product code, and currency amounts should be presented in a consistent, standard way in any presentation. When number and identifiers are separated into natural groups, such as standard U.S. phone number formats [+1(555)999-1234], they are easier to remember and use (English, 2009)	D
Data Integrity fundamentals	A measure of the existence, validity, structure, content, and other basic characteristics of the data (McGilvray, 2008)	D
Semantic Consistency	The extent of consistency in using the same values (vocabulary control) and elements to convey the same concepts and meanings in an information object. This also includes the extent of semantic consistency among the same or different components of the object (Stvilia et al., 2007)	D
Structural Consistency	The extent to which similar attributes or elements of an information object are consistently represented using the same structure, format, and precision (Stvilia et al., 2007)	D
Mapped consistently	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)	D
Concurrency of redundant or distributed data	The Information Float or Lag Time is acceptable between (a) when data is knowable (create or changed) in one data store to (b) when it is also knowable in a redundant or distributed data store, and concurrent queries to each data store produce the same result. (English, 2009)	D

In (McGilvray, 2008) and (English, 2009) the dimension of *Duplication/Non-Duplication* emphasizes maintaining non-redundant data sets within the organizational landscape including all multiple sources of data available. The same point of view is also presented by IBM and Informatica in (B. Byrne, 2008) and (Loshin, 2006) respectively under the dimension *Uniqueness/Unique*.

In (Loshin, 2006), the term consistency as a dimension is defined referring to multiple data sources as, " *….in its most basic form, consistency refers to data values in one data set being consistent with values in another data set. A strict definition of consistency specifies that two data values drawn from separate data sets must not conflict with each other, although consistency does not necessarily imply correctness".*

The definitions given for the term consistency by SAP (G. Gatling, 2007) and IBM (B. Byrne, 2008), also follow a similar approach to that of the above definitions. In (HIQA, 2011), the dimension coherence is defined as "*Comparability of data refers to the extent to which data is consistent between organisations and over time allowing comparisons to be made".* This definition emphasizes that data should be consistent between the organizations to make comparisons. All dimensions in this cluster are based on declarative perspective referring to the consistent representation of real world objects and database integrity fundamentals.

Table 16: Characteristics of consistency

Characteristic	Description	Granularity	Type
Uniqueness	The data is uniquely identifiable.	R	D
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 & Interpretability:

Table 17: Dimensions relating to Usability and Interpretability.

Comparability	Comparability aims at measuring the impact of differences in applied statistical concepts and measurement tools/procedures when statistics are compared between geographical	U

	areas, non-geographical domains, or over time. (Lyon, 2008)	
	Comparability of data refers to the extent to which data is consistent between organisations and over time allowing comparisons to be made. This includes using equivalent reporting periods. (Lyon, 2008, HIQA, 2011)	U
Interpretability	A good format is one that helps the user interpret values correctly. Consider a domain consisting of three values and two candidate representations: (1, 2, 3) and (poor, good, excellent). Obviously the second format is superior because it is less likely to be misinterpreted. This point is one where the connection of data quality to the user is most clear. Data are being presented to users so they may be used properly. Formats that hinder correct interpretation may increase rework and lower downstream, drastically lowering the utility of data given by such a format.(Redman, 1997)	U
	Data are in appropriate language and unit and data definitions are clear (Wang and Strong, 1996)	U
	Interpretability of data refers to the ease at which the user can understand the data. Is there any ambiguity in understanding the data and is there information available to help the user understand the terminology? (HIQA, 2011)	U
Correct Interpretation	A good presentation provides the user with everything required for the correct interpretation of information. When there is any possibility of ambiguity, a key or legend should be included. (Loshin, 2001)	U
Unambiguity	Data is not ambiguous if it allows only one interpretation – anti-example: Song.composer = ‘Johann Strauss’ (father or son?). (Kimball and Caserta, 2004)	U
Concise representation	Data are compactly represented without being overwhelmed (Wang and Strong, 1996)	U
Ease of understanding	Data are clear without ambiguity and easily comprehended (Wang and Strong, 1996)	U
Format precision	The set S should be sufficiently precise to	U

	distinguish among elements in the domain that must be distinguished by users. This dimension makes clear why icons and colors are of limited use when domains are large. But problems can and do arise for the other formats as well, because many formats are not one-to-one functions. For example, if the domain is infinite (the rational numbers, for example), then no string format of finite length can represent all possible values. The trick is to provide the precision to meet user needs. (Redman, 1997, Loshin, 2001)	
	The degree of precision of the presentation of an attribute's value should reasonably match the degree of precision of the value being displayed. The user should be able to see any value the attributer may take and also be able to distinguish different values. (Redman, 1997, Loshin, 2001)	U
Understandable	Data is presented in an intelligible manner (Price and Shanks, 2005)	U
Presentation Standardization	The Characteristic in which formatted data is presented consistently in a standardized or consistent way across different media, such as in computer screens, reports, or manually prepared reports (English, 2009)	U
Format flexibility	Good format, like good views, are flexible so that changes in user need and recording medium can be accommodated. (Redman, 1997)	U
Appropriateness	The most important quality characteristic of a format is its appropriateness. One format is more appropriate than another if it is better suited to users' needs. The appropriateness of the format depends upon two factors: user and medium used. Both are of crucial importance. The abilities of human users and computers to understand data in different formats are vastly different. For example, the human eye is not very good at interpreting some positional formats, such as bar codes, although optical scanning devices are. On the other hand, humans can assimilate much data from a graph, a format that is relatively hard for a computer to interpret. Appropriateness is	U

	related to the second quality dimension, interpretability. (Redman, 1997, Loshin, 2001)	
	Appropriateness is the dimension we use to categorize how well the format and presentation of the data match the user needs. In our example, there is a difference between a high-level monthly sales report that is supplied to senior management and the daily product manifests that are handed to the shipping department for product packaging. (Redman, 1997, Loshin, 2001)	U
Structured Valued Standardization	Structured Attributes like dates, time, telephone number, tax ID number, product code, and currency amounts should be presented in a consistent, standard way in any presentation. When number and identifiers are separated into natural groups, such as standard U.S. phone number formats [+1(555)999-1234], they are easier to remember and use (English, 2009)	U
Document Standardization	Periodic Reports, such as Financial Statements, Annual Reports, and Policy and Procedure Manuals should have a standard format with a style sheet that presents the information in a consistent and easily read and understood format. (English, 2009)	U
Suitably presented	Data is presented in a manner appropriate for its use, with respect to format, precision, and units. (Price and Shanks, 2005)	U
Flexibly presented	Data can be easily manipulated and the presentation customized as needed, with respect to aggregating data and changing the data format, precision, or units (Price and Shanks, 2005)	U
Presentation Quality	A measure of how information is presented to and collected from those who utilize it. Format and appearance support appropriate use of information (McGilvray, 2008)	U
Representational consistency	Data are always presented in the same format and are compatible with the previous data (Wang and Strong, 1996)	U
Informativeness /Redundancy	Intrinsic: The extent to which the information is new or informative in the context of a particular activity or community (Stvilia et al.,	U

	2007)	
	Relational Contextual:The amount of information contained in an information object. At the content level, it is measured as a ratio of the size of the informative content (measured in word terms that are stemmed and stopped) to the overall size of an information object. At the schema number of elements in the objectlevel it is measured as a ratio of the number of unique elements over the total (Stvilia et al., 2007)	U
Interactivity	Can the information process be adapted by the information consumer? (Eppler, 2006)	U
Presentation media appropriateness	The Characteristic of Information being presented in the right technology Media, such as online, hardcopy report, audio, or video. (English, 2009)	U
Presentation Utility	The degree to which Information is presented in a way Intuitive and appropriate for the task at hand. The Presentation Quality of Information will vary by the individual purposes for which it is required. Some users require concise presentation, whereas others require a complete, detailed presentation, and yet others require graphic, color, or other highlighting techniques (English, 2009)	U
Presentation Clarity	The Characteristic in which Information is presented in a way that clearly communicates the truth of the data. 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 and any anomalies in the Information (English, 2009)	U
Relevance/ Relevancy	Data are applicable and useful for the task at hand (Wang and Strong, 1996)	U
	Relevance is the degree to which statistics meet current and potential users' needs. It refers to whether all statistics that are needed are produced and the extent to which concepts used (definitions, classifications etc.) reflect user needs (Lyon, 2008)	U
	Relevance of data refers to the extent to which	U

	the data meets the needs of users. Information needs may change and is important that reviews take place to ensure data collected is still relevant for decision makers. (HIQA, 2011)	
	The extent to which information is applicable in a given activity (Stvilia et al., 2007)	U
	The Characteristic in which the Information is the right kind of Information that adds value to the task at hand, such as to perform a process or make a decision. (English, 2009)	U
Transactability	A measure of the degree to which data will produce the desired business transaction or outcome (McGilvray, 2008)	U
Usability	Usability of data refers to the extent to which data can be accessed and understood. (HIQA, 2011)	U
Value added	Data are beneficial and provide advantages for their use (Wang and Strong, 1996)	U
Appropriate amount of data	The quantity or volume of available data is appropriate (Wang and Strong, 1996)	U
Clarity	Is the information understandable or comprehensible to the target group? (Eppler, 2006)	U
Applicability	Can the information be directly applied? Is it useful? (Eppler, 2006)	U
Convenience	Does the information provision correspond to the user's needs and habits? (Eppler, 2006)	U
Cohesiveness	The extent to which the content of an object is focused on one topic (Stvilia et al., 2007)	U
Complexity	The extent of cognitive complexity of an information object measured by some index or indices (Stvilia et al., 2007)	U
Informativeness/Redundancy	The amount of information contained in an information object. At the content level, it is measured as a ratio of the size of the informative content (measured in word terms that are stemmed and stopped) to the overall size of an information object. At the schema number of elements in the objectlevel it is measured as a ratio of the number of unique elements over the total (Stvilia et al., 2007)	U
Naturalness	The extent to which the model or schema and content of an information object are expressed by conventional, typified terms and forms according to some general-purpose reference	U

	source (Stvilia et al., 2007)	
Flexibility	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. For example, a system that display different counties; currencies may need to have the screen presentation change to allow for more significant digits for prices to be displayed when there is a steep devaluation in one county's currency (Loshin, 2001)	U
Ubiquity	As a data quality-oriented organization matures, the agreement of usage will move from a small set of "early adopters" to gradually encompass more and more of the enterprise, Ubiquity measures the degree to which different departments in an organization use shared reference data. (Loshin, 2001)	U
Precise	The data element is used only for its intended purpose, that is, the degree to which the data characteristics are well understood and correctly utilized. (B. Byrne, 2008)	U
Portability	In an environment that makes use of different kinds of systems and applications, a portable interface is important so that as applications are migrated from one platform to another, the presentation of data is familiar to the users. Also, when dealing with a system designed for international use, the user of international standards as well as universally recognized icons is a sign of system designed with presentation portability in mind. (Loshin, 2001)	U
	Good formats are portable or universal. This means that they can be applied to as wide a range of situations as possible. The male and female icons mentioned earlier are excellent for this reason. Portability is especially important in situations similar to those employing these icons-a variety of users that portability levels of skill in understanding the format. It can be expected that portability will be of increased importance as worldwide telecommunications continue to	U

	improve.(Redman, 1997)	

The dimensions grouped into this cluster are a combination of the characteristics which help the utilization of data for its intended purposes. Some definitions emphasize factors to improve interpretability of data such as good formats and documents to present data for interpretation purposes. For example, "*good format, like good views, are flexible so that changes in user need and recording medium can be accommodated"* (Redman, 1997). Further, (English, 2009) and (Loshin, 2001) emphasize the same aspect. Some definitions focus on unambiguity, conciseness and clarity related aspects, and others contribute towards richness of interpretation. As per (Kimball and Caserta, 2004), *"data is not ambiguous if it allows only one interpretation"*. In (HIQA, 2011) the authors defines interpretability as: " *...the ease at which the user can understand the data"*. Similarly the same point is expressed in (Wang and Strong, 1996). Usefulness of data is emphasized by some authors some authors (McGilvray, 2008) who define the term Transactability as "*a measure of the degree to which data will produce the desired business transaction or outcome"* (McGilvray, 2008) . Whereas in (HIQA, 2011) and (Wang and Strong, 1996) define the terms Usability and value added with a similar focus on usefulness of data. English (English, 2009) has also emphasized the usability and interpretability aspects through definitions for Presentation Utility, Presentation Clarity and Presentation media appropriateness.

The characteristics identified in this cluster are as follows.

Table 18: Characteristics of usability and interpretability

Characteristic	Description	Granularity	Type
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

5 Summary

In our analysis we applied a rigorous multi-coder approach to categorize 127 data quality dimensions from 16 sources using thematic analysis, providing a consolidated view of the related DQ dimensions. The classification resulted in eight main clusters and a set of dominant quality characteristics within each cluster. Altogether thirty such quality characters were identified within the eight main clusters and we provided a definition and representative term for each characteristic. For each main cluster, we selected an umbrella term that best represents the cluster. Further in this analysis, we have classified each individual definition using the two perspectives (declarative and usage) to provide further characterization for each definition, as well as identify definitions that do not exhibit either of the two perspectives. In our analysis we found three such definitions that could not be convincingly explained from either perspective, nor fit into any of the above clusters based on their underlying motivations and definitions. These are ‘Efficient use of memory’ and ‘Use of storage’ defined in (Redman, 1997) and (Loshin, 2001) respectively, which focus on the utilization of disk space and memory space of computers while referring to logical and physical data modelling aspects to take proactive measures at the very early stages of IS analysis and design. In addition, ‘Stewardship’ (Loshin, 2001) is focused on assigning the responsibility for data, and represents more of a management function rather than a declarative or usage perspective of data quality.

This consolidated view and analysis of DQ dimensions aims to resolve the increasing proliferation of a plethora of DQ dimensions that share the same title with a differing focus, or, vice versa, that are reborn by authors as new DQ dimensions when, in fact, they have the same focus as that put forth by prior DQ researchers. Indeed, an agreement on the core dimensions of DQ is central to effective communication about DQ expectations in organisations, as well as being central to any efforts that focus on formal data quality requirements modelling.

6 Conclusion and Future Work

DQ dimensions are a foundational concept in the study of data quality and data quality management. Though data quality is a widely researched topic, in more recent years significant contribution to this body of knowledge has stemmed from practitioners. The practitioner viewpoints are a substantial value-add, evident from the large customer bases they support. However, 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 in the body of knowledge.

In this paper we have analysed data quality dimensions defined in sixteen credible sources into eight common clusters and thirty three characteristics of data quality providing new definitions. This classification provided a basis on which a shared understanding of DQ dimensions can be achieved, by removing overlaps, redundancies, and conflicts, while embracing the diversity and importance of contextual interpretations. The shared understanding developed is an essential prelude for DQ requirements modelling.

Currently, we are extending the explanations of the dimensions defined within the clusters using practical examples with the help of data professionals and managers who deal with data quality issues on a daily basis. This extended work will identify which definitions are more prominent in practice, and which are rarely used, and provide meaningful use cases for each definition. The extended work is expected to generate patterns of usage for a wide variety of data quality dimensions and will provide much needed baseline knowledge for data quality requirements modelling, and consequently, data quality assessment and enforcement frameworks.

References

Oxford Dictionaries [Online]. Oxford University Press. Available: http://oxforddictionaries.com/definition/english/dimension 2013.

(DAMA), D. M. A. *Data Management Association (DAMA)* [Online]. Data Management Association (DAMA): Data Management Association (DAMA). Available: http://www.dama.org.au/ [Accessed 20/10/2012 2012].

B. 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.

BATINI, C., FRANCALANCI, C., CAPPIELLO, C. & MAURINO, A. 2009. Methodologies for data quality assessment and improvement. *ACM computing surveys*, 41**,** 1 - 52.

BATINI, C. & SCANNAPIECO, M. 2006. *Data quality: concepts, methodologies and techniques*, Springer.

BRAUN, V. & CLARKE, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology,* 3**,** 77-101.

CARLETTA, J. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics,* 22**,** 249-254.

ENGLISH, L. P. 2009. *Information quality applied: Best practices for improving business information, processes and systems*, Wiley Publishing.

EPPLER, M. J. 2006. *Managing information quality: increasing the value of information in knowledge-intensive products and processes*, Springer.

EPPLER, M. J. & MUENZENMAYER, P. 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, 2002. Citeseer, 187-196.

EVEN, A. & SHANKARANARAYANAN, G. Value-Driven Data Quality Assessment. Tenth International Conference on Information Quality (ICIQ'05), 2005.

FRIEDMAN, F. Magic Quadrant for Data Quality Tools. 2012. Gartner Inc.

G. GATLING, C. B., R. CHAMPLIN, H. STEFANI, G. WEIGEL 2007. *Enterprise Information Management with SAP*, Boston, Galileo Press Inc.

GARVIN, D. A. 1987. Competing on the Eight Dimensions of Quality. *Harvard Business Review***,** 101-109.

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.

ISO 2012. ISO 8000-2 Data Quality-Part 2-Vocabulary. ISO.

JURAN, J. M. 1962. *Quality control handbook,* New York, McGraw-Hill Publishing

KIMBALL, R. & CASERTA, J. 2004. The data warehouse ETL toolkit: practical techniques for extracting. *Cleaning, Conforming, and Delivering, Digitized Format, originally published*.

LEE, Y. W., STRONG, D. M., KAHN, B. K. & WANG, R. Y. 2002. AIMQ: a methodology for information quality assessment. *Information & management,* 40**,** 133-146.

LIEBENAU, J. & BACKHOUSE, J. 1990. *Understanding information: an introduction*, Palgrave Macmillan.

LOSHIN, D. 2001. *Enterprise knowledge management: The data quality approach*, Morgan Kaufmann Pub.

LOSHIN, D. 2006. Monitoring Data quality Performance using Data Quality Metrics. *Informatica Corporation*.

LYON, M. 2008. Assessing Data Quality,Monetary and Financial Statistics. *Bank of England.* http://www.bankofengland.co.uk/statistics/Documents/ms/articles/art1mar08.pdf.

MCGILVRAY, D. 2008. *Executing data quality projects: Ten steps to quality data and trusted information*, Morgan Kaufmann.

MORRIS, C. 1938. Foundation of the theory of signs. London: University of Chicago Press.

PIPINO, L. L., LEE, Y. W. & WANG, R. Y. 2002. Data quality assessment. *Communications of the ACM,* 45**,** 211-218.

PRICE, R. & SHANKS, G. A semiotic information quality framework. Proceedings of the International Conference on Decision Support Systems DSS04, 2004. Citeseer, 658-672.

PRICE, R. J. & SHANKS, G. Empirical refinement of a semiotic information quality framework. System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on, 2005. IEEE, 216a-216a.

REDMAN, T. C. 1997. *Data quality for the information age*, Artech House, Inc.

RUSSELL, R. S. & TAYLOR, B. W. 2003. *Operations management*, Prentice Hall Upper Saddle River, NJ.

SADIQ, S., YEGANEH, N. Y. & INDULSKA, M. An Analysis of Cross-Disciplinary Collaborations in Data Quality Research. European Conference on Information Systems, 2011 Helsinki Finland.

SCANNAPIECO, M. & CATARCI, T. 2002. Data quality under a computer science perspective. *Archivi & Computer,* 2**,** 1-15.

STVILIA, B., GASSER, L., TWIDALE, M. B. & SMITH, L. C. 2007. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology,* 58**,** 1720-1733.

WANG, R. Y. 1998. A product perspective on total data quality management. *Communications of the ACM,* 41**,** 58-65.

WANG, R. Y. & STRONG, D. M. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems***,** 5-33.

WATSON-MANHEIM, M. B., CHUDOBA, K. M. & CROWSTON, K. 2002. Discontinuities and continuities: A new way to understand virtual work. *Inform. Technol. People*.

WILLCOCKS, L. & LESTER, S. 1996. Beyond the IT productivity paradox. *European Management Journal*, 14**,** 279-290.

YONKE, C. L., WALENTA, C. & TALBURT, J. R. 2011. The job of the Information/Data Quality Professional. International Association for Information and data Quality (IAIDQ).