An Evaluation of the Conformed Dimensions of Data Quality in Application to an Existing Information Quality-Privacy-Trust Research Framework

(Research-in-Progress)

Dan Myers, MBA, IQCP
DQ Matters
dan@dqmatters.com

Brian P. Blake
Univ. of Arkansas at Little Rock
bpblake@ualr.edu

Abstract: This research compares a proposed standard, cross-industry set of dimensions of data quality, the Conformed Dimensions of Data Quality, to the data quality dimensions previously espoused by Pipino, Lee, and Wang. Further, these Conformed Dimensions of Data Quality are evaluated in application to a proposed Information Quality-Privacy-Trust research framework that currently utilizes the related Lee, Pipino, and Wang data quality dimensions as a foundation. This evaluation serves as a validation of the benefits of the existing conformed dimensions and highlights possible extensions of the conformed dimensions needed for application in contexts that require the inclusion of subjective dimensions of data quality. The need for future research to address information system level data quality dimensions is also identified. Finally, continued evaluation of the CDDQ against addition frameworks and applications is recommended to increase the overall robustness of the proposed framework.

Keywords: Conformed Dimensions, Data Quality, Information Quality Research Frameworks

INTRODUCTION

Information quality (also known as data quality) is a multidisciplinary field with research spanning a wide range of topics, but existing researchers are primarily operating in the disciplines of Management Information Systems and Computer Science [11]. Within quality literature, the concept of "fitness for use" has been widely adopted as a definition for data quality [8][11]-[14]. But to be applicable, this definition of fitness for use must be contextualized [8]. In this regard, previous writings and research have presented data quality as a multi-dimensional concept [11]-[15]. But this has resulted in confusion regarding which set of dimensions an organization should use to measure quality.

The Conformed Dimensions of Data Quality (CDDQ) was established based on a desire to standardize the language used by many authors and organizations regarding dimensions that describe data quality. Communication is inefficient and frustrating if not based on a standard, so the CDDQ is focused on finding agreement and simplicity of terminology within existing research and increasing future agreement in application through the buy-in and use of the proposed conformed dimensions. The CDDQ list a set of high level dimensions and their underlying concepts which provide a granular breakdown of ideas within each dimension. This conformed dimension framework also provides includes example metrics for each of the underlying concepts, but this paper will not directly address that level of detail.

In this research, the Conformed Dimensions of Data Quality are evaluated in application to a proposed Information Quality-Privacy-Trust research framework that currently utilizes the Lee, Pipino, Funk and Wang data quality dimensions. This framework matrix is part of ongoing dissertation research toward a Computer and Information Sciences Ph.D. with a focus on Information Quality and has been a component of several previously published papers [1][2].

BACKGROUND

Conformed Dimensions of Data Quality

In 2013, comparative research was completed for six information quality author/organization's lists of the dimensions of data quality and methodology to align the definitions was proposed [4]. In 2016, a standard set of dimensions, called the Conformed Dimensions of Data Quality [3], was published online[1] based on the 2013 research. For the last three years, a survey about the use of the dimensions of data quality in general, and interest in a standard, has been conducted. Each year, companion whitepapers have been published on the CDDQ website charting interest in such a standard and observing industry trends [6].

As noted previously, the Conformed Dimensions of Data Quality list a set of high level dimensions and their more detailed underlying concepts. The CDDQ was established based on a desire to standardize the language used by many authors and organizations regarding dimensions that describe data quality. By identifying underlying concepts, or ideas that formed the basis of a dimension, it was found that different authors' contributions could be disassembled and compared at the concept level and reassembled in the most commonly understood way- as a conformed set of dimensions. This can be illustrated as follows.

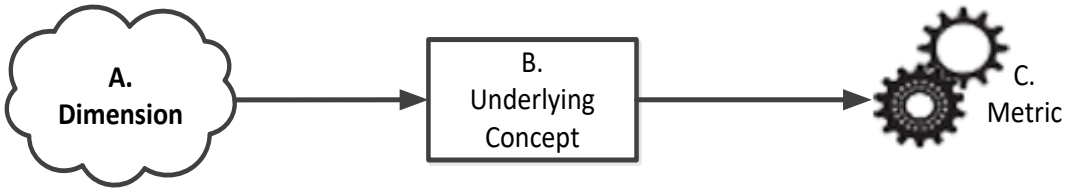

Figure 1 – CDDQ Dimension-Concept-Metric Conceptualization

A given dimension, such as Completeness, can be broken down into a finite number of concepts such as Attribute Population, Record Population, etc. These concepts then form the basis for definition of example metrics. This is highlighted in Table 1.

	Description
Dimension: Completeness	Completeness measures the degree of population of data values in a data set.
Underlying Concept: Attribute Population	This measures whether a value is present (not null) for an attribute (column).
Metric: Attribute Population	For a given column, the count of not null rows divided by the total number of rows in the set.
Metric Formula:	Column Population = for a given column, the number of row values not null / number of rows in set
Example Scenario for Metric:	A retail company sells T-shirts, Diapers, and Pants, but only directly ships Pants to end-customers via its website and T-shirts and Diapers are distributed through grocery store chains. The transactional table that lists sales of items has three rows in it (albeit small) with one T-shirt sale and one Diaper package sale and one Pants sale. The DIRECT_SHIP column of this table stores a "Y" when the item is shipped directly to the end-customer. In this scenario, we see that the DIRECT_SHIP column is 1/3 (33.3%) Not Null or has an Attribute Population of 33.3%. See Appendix F.

Table 1 – CDDQ Dimension to Metric Example (r3.3)

[1] Online at: http://dimensionsofdataquality.com

Additional Background on Prior CDDQ Development

In preparation for the 2013 publication [4], clarification of previous published material and professional input regarding the dimensions of data quality were solicited from Danette McGilvray, Thomas Redman, and others in the information quality community, as well as practitioners in the field. Case studies about implementing the Conformed Dimensions are under development, and the authors encourage more organizations to leverage the proposed standard. To help practitioners and researchers identify the appropriate application of the Conformed Dimensions, a set of principles has been developed and a glossary of terms used in the standard. Table 2 lists these principles.

#	Principle	Explanation	Discussion
1	Quantifiable Objective Focus	The Conformed Dimensions are focused on providing standard language for objective and quantifiable measures of data quality.	The Conformed Dimensions all have explicit definitions, and at least one underlying concept that further characterizes the aspect of quality. The goal is to ensure scientifically measurable criteria that enable repeatability through standardization.
2	Independent of System	The Conformed Dimensions are independent of storage or system specific constraints.	The ISO/IEC 25012:2008 Dimensions of Data Quality include dimensions like "Portability" or "Recoverability" that focus on system specific constraints. The Conformed Dimensions, in contrast, are independent of Information Systems platform and physical data storage.

Table 2: CDDQ Principles

Because, in practice, each organization tends to define subjective dimensions in a different way, the CDDQ does not include definitions for subjective dimensions such as Believability or Trustworthiness. Instead of directly including a highly subjective dimension, it is recommended that, if required, organizations define an enterprise standard definition of the desired subjective dimension based on underlying objective dimensions as proposed later in this paper, or a composition of objective and measurable subjective input. For example, "Believability: exists for an attribute which is sourced from the governed system of record, and has Validity- Domain of Predefined Values at 99.9% for the last two months" where system of record is documented, Validity is objectively defined in the Conformed Dimensions.

Stewardship of the Conformed Dimensions

The Conformed Dimensions of Data Quality are maintained at www.dimensionsofdataquality.com and updated on a periodic basis to remedy issues identified by users such as awkward language and, although infrequent, addition of underlying concepts. Prior versions are maintained in PDF form on the website. Table 3 presents the current version, as of paper submission, of the Conformed Dimensions of Data Quality (r3.3) and their underlying concepts.

Conformed Dimension	Conformed Dimension Definition	Underlying Concepts	Non-Standard Terminology for Dimension
Completeness	Completeness measures the degree of population of data values in a data set.	Record Population, Attribute Population, Truncation, Existence	Fill Rate, Coverage, Usability, Scope
Accuracy	Accuracy measures the degree to which data factually represents its associated real-world object, event, concept or alternatively matches the agreed upon source(s).	Agree with Real-world, Match to Agreed Source	Consistency
Consistency	Consistency measures whether or not data is equivalent across systems or location of storage.	Equivalence of Redundant or Distributed Data, Format Consistency	Integrity, Concurrence, Coherence

Validity	Validity measures whether a value conforms to a preset standard.	Values in Specified Range, Values Conform to Business Rule, Domain of Predefined Values, Values Conform to Data Type, Values Conform to Format	Accuracy, Integrity, Reasonableness, Compliance
Timeliness	Timeliness is a measure of time between when data is expected versus made available.	Time Expectation for Availability, Manual Float	Currency, Lag Time, Latency, Information Float
Currency	Currency measures how quickly data reflects the real-world concept that it represents.	Current with World it Models	Timeliness
Integrity	Integrity measures the structural or relational quality of data sets.	Referential Integrity, Uniqueness, Cardinality	Validity, Duplication
Accessibility	Accessibility measures how easy it is to acquire data when needed, how long it is retained, and how access is controlled.	Ease of Obtaining Data, Access Control, Retention	Availability
Precision	Precision measures the number of decimal places and rounding of a data value or level of aggregation.	Precision of Data Value, Granularity	Coverage, Detail
Lineage	Lineage measures whether factual documentation exists about where data came from, how it was transformed, where it went and end-to-end graphical illustration.	Source Documentation, Segment Documentation, Target Documentation, End-to-End Graphical Documentation	
Representation	Representation measures ease of understanding data, consistency of presentation, appropriate media choice, and availability of documentation (metadata).	Easy to Read & Interpret, Presentation Language, Media Appropriate, Metadata Availability	Presentation

Table 3: List of Conformed Dimensions of Data Quality

Each conformed dimension has one or more underlying concepts. The definitions for all underlying concepts of the current version of the Conformed Dimensions is in Appendix C. An example of that is provided here in Table 4 to see how granular the CDDQ are documented.

Conformed Dimension	Underlying Concepts	Definition of Underlying Concept
Completeness	Record Population	This measures whether a row is present in a data set (table).
	Attribute Population	This measures whether a value is present (not null) for an attribute (column).
	Truncation	This measures whether the value contains all characters of the correct value.
	Existence	Existence identifies whether a real-life fact has been captured as data.

Table 4: List of Underlying Concepts for Completeness in CDDQ

Prospective Applications of CDDQ:

We've included a list of prospective applications of the CDDQ in Table 5. These include use within professional associations, use by software venders, use for simplification in education, use by governmental or regulatory bodies, and use by data providers.

Application	Discussion
Use within professional associations (e.g. DAMA, IQ International, EDM Council) to reduce confusion and complexity in the IQ industry.	Clearly acceptance by each of these organizations would require review, validation, and even likely enhancements to the CDDQ, the net result would be a stronger and more thoroughly understood standard.
Use by software vendors to standardize names of measures provided in their software.	Based on informal discussions with a few vendors, if a single standard becomes available they would likely enhance their product by leveraging IQ industry accepted names and formulas.
Use for Simplification in education	Simplification of education through use of a single body of explanation, rather than having to learn various versions of the dimensions of data quality espoused by authors over the last 25+ years.
Use by Governmental or Regulatory Bodies	As future regulatory standards are developed, regulators and standards bodies will naturally look for vendor neutral/free methods of measuring data quality. The CDDQ (as a non-proprietary tool and vendor agnostic framework) is a good option.
Use by Data Providers	Given that data providers have service level agreements regarding the frequency (accessibility), currency, formats, and other measures of quality the use of the CDDQ provides a method to standardize data products across industries, clients, and geographies which should reduce cost, improve communication and reusability of data preparation software code.

Table 5 – Prospective Applications of CDDQ

Acknowledgement of Challenges:

Although all the prospective applications above are positive in nature, we should call attention to items not solved by the development of a set of Conformed Dimensions. Some of these include:

- **Broader Adoption:** Creation of a standard does not ensure acceptance and broad adoption. If the result of normalization of the dimensions of data quality into a Conformed Set of Dimensions is not ultimately used across organizations and companies, then it's value will be limited.

- **Re-Education & Change Management:** As with any change, change management principles will need to be followed to facilitate re-education of professionals in the industry. This will be challenging for some who do not see personal value in using the standard or simply don't want to spend time learning a different language. For that reason, it may be more easily accepted by new industry entrants and students.

Wang & Strong / Pipino, Lee, & Wang Data Quality Dimensions

In 1996, Wang and Strong published an empirical framework to capture the multi-dimensional aspects of information quality that are most important to data consumers [13]. This research was presented in application by Strong, Lee and Wang in "Data Quality in Context" the following year [14]. Since that time, their framework has been widely cited in information quality literature by hundreds of authors. The Wang and Strong Quality Framework [13] contains four categories of data quality: Intrinsic DQ, Contextual DQ, Representational DQ, and Accessibility DQ. These four categories contain fifteen data quality dimensions in Table 5.

DQ Category	DQ Dimensions
Intrinsic DQ	Accuracy, Objectivity, Believability, Reputation
Accessibility DQ	Accessibility, Access Security
Contextual DQ	Relevancy, Value-Added, Timeliness, Completeness, Amount of Data
Representational DQ	Interpretability, Ease of Understanding, Concise Representation, Consistent Representation

Table 5: Wang Strong Quality Framework [13]

Intrinsic data quality dimensions "have quality in their own right" [13]. Fisher, Lauria, Chengalur-Smith, and Wang [12] describe these as non-contextual self-contained quality aspects. Accessibility data quality includes the dimensions of Access and Security [13] and deals with the availability and protection of data [12]. Contextual dimensions "must be considered within the context of the task at hand" [13] and are "specifically tied to the particular use or user in order to determine quality" [12]. Representational data quality relates to the format and meaning of the data [13] and focus on the importance of the presentation and usability of data [12]. In addition, definitions of these data quality dimensions as presented by Pipino, Lee, and Wang [15] can be found in Table 6.

Dimensions	Definitions
Accessibility	The extent to which data are available, or easily and quickly retrievable
Appropriate Amount of Data	The extent to which the quantity and volume of available data is appropriate
Believability	The extent to which data are accepted or regarded as true, real, and credible
Completeness	The extent to which data are of sufficient depth, breadth, and scope for the task at hand
Concise Representation	The extent to which data are compactly represented
Consistent Representation	The extent to which data is presented in the same format
Ease of Manipulation	The extent to which data is easy to manipulate and apply to different tasks
Free-of-Error	The extent to which data is correct and reliable
Interpretability	The extent to which data is in appropriate languages, symbols, and units, and the definitions are clear
Objectivity	The extent to which data is unbiased, unprejudiced, and impartial

Relevancy	The extent to which data is applicable and helpful for the task at hand
Reputation	The extent to which data is highly regarded in terms of its sources or content
Security	The extent to which access to data is restricted appropriately to maintain its security
Timeliness	The extent to which the data is sufficiently up-to-date for the task at band
Understandability	The extent to which data is easily comprehended
Value-Added	The extent to which data is beneficial and provides advantages from its use

Table 6: Pipino, Lee, and Wang Data Quality Dimensions [15]

Information Quality-Privacy-Trust Research Framework Matrix

Prior research [1][2] focuses on the general overlap of the multi-faceted dimensions, aspects, and properties of trust, privacy, information quality, and online social networks. It seeks to identify where these areas overlap regarding both online social networks and each other. This research hypothesizes that:

H1: The multi-faceted dimensions, aspects, and properties of trust, privacy, and information quality can be effectively overlaid within a series of related matrices.

H2: An understanding of intersections of these sub-aspects lends itself to a broader understanding of the relationship of these concepts.

H3: An understanding of intersections of these sub-aspects lends itself to specific target areas for future research.

As a starting point for this research, a framework matrix has been developed to map the points of intersection between four aspects of prior research. These include: 1) Solove's [9] taxonomy of privacy, 2) Schneier's [10] divisions of social network data, 3) Wang and Strong's [13] multiple dimensions of information quality, and 4) the trustworthiness characteristics of Ability, Benevolence, and Integrity as presented by Mayer, Davis, and Schoorman [19] and Gefen [17]. This Information Quality-Privacy-Trust research framework matrix, as presented in prior research [1][2], is highlighted in Table 7.

The development and validation of select relationship matrices for data privacy, online social networks, information quality, and trust as a research framework is the first deliverable from this research. This is accomplished in part through a validation in current literature. Hogben [20], for example, highlighted specific online social network privacy threats that include digital dossier aggregation, secondary data collection, recognition and identification, data permanence, infiltration of networks, profile squatting and ID theft related reputation slander, and cyberstalking/cyberbullying. These can be shown to align neatly with the privacy components within the framework matrix. In addition, a corresponding validation survey has been created and is being implemented for select professionals and topic experts. Their opinions in relation to the framework matrices will be gathered and reconciled. The framework matrix will be further validated through structured equation modeling as the trade-offs between certain framework relationships are measured. Further, this current comparative research to the CDDQ framework will both confirm viability of the utilized quality dimensions and well as will help identify underlying concepts for measurement in the development of this planned structured equation model.

	Types of Social Networking Data					
	Service Data	**Disclosed Data**	**Entrusted Data**	**Incidental Data**	**Behavioral Data**	**Derived Data**
	Data you give the social network site in order to use it	What you post on your own pages	What you post on other people's pages	What other people post about you	Data the site collection about your habits by recording what you do and who you do it with	Data about you that is derived from all other data
Data Privacy Issues	Insecurity Secondary use Breach of Confidentiality	Increased Accessibility Insecurity Appropriation Secondary Use	Increased Accessibility Secondary use Identification Exclusion Breach of Confidentiality Disclosure Exposure Distortion Intrusion (onto their pages)	Identification Exclusion Breach of Confidentiality Disclosure Exposure Distortion Intrusion (onto your pages) Increased Accessibility Secondary use	Aggregation Insecurity Secondary Use Breach of Confidentiality Identification Exclusion	Aggregation Insecurity Secondary Use Breach of Confidentiality Identification Exclusion
Information Quality Dimensions	Accuracy Appropriate Amount Relevancy Security Accessibility Concise Representation Consistent Representation	Accuracy Appropriate Amount Relevancy Security Believability Reputation Understandability Accessibility Objectivity Ease of Operation	Accuracy Appropriate Amount Relevancy Security Believability Reputation Understandability Accessibility Objectivity Ease of Operation	Accuracy Appropriate Amount Relevancy Security Believability Reputation Understandability Accessibility Objectivity Ease of Operation	Accuracy Appropriate Amount Relevancy Security Timeliness Concise Representation Completeness Consistent Representation Accessibility Understandability Interpretability	Accuracy Appropriate Amount Relevancy Security Accessibility Understandability Interpretability Consistent Representation Concise Representation
Trust	Ability Benevolence Integrity	Benevolence Integrity	Benevolence Integrity	Benevolence Integrity	Ability Benevolence Integrity	Ability Benevolence Integrity

Table 7: Framework Matrix: Information Quality, Data Privacy, and Trust in Social Media Networks [1]

A further application of this framework matrix is in understanding which aspects of information quality, privacy, and trust relate to each other in regard to the concept of Information Quality Modification found within the prior related research [1][2]. This is based on the hypothesis that:

H4: Behavioral intent to share information is not a simple binary response. Instead it is a degree based response that uses information quality modification to mitigate privacy and trust concerns between the thresholds of open disclosure and full non-disclosure (see Fig. 2).

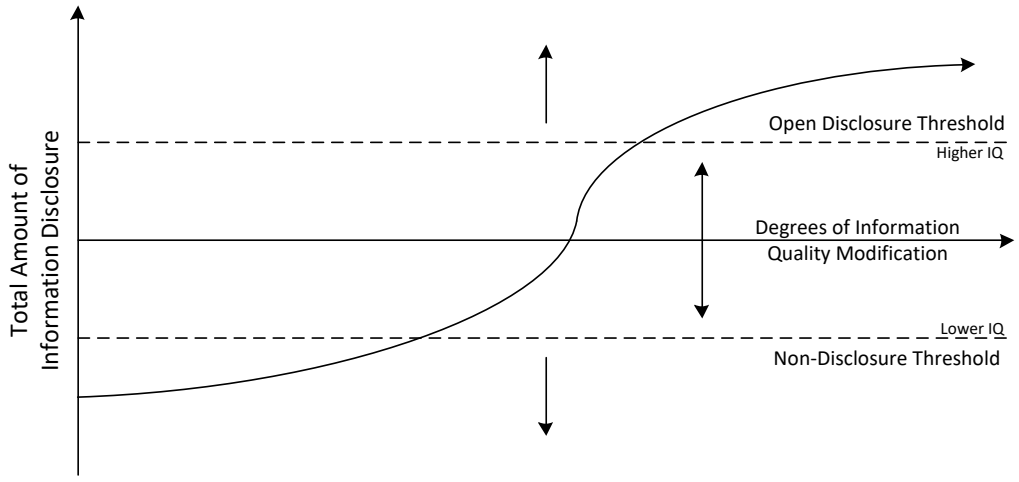

Figure 2: Initial Information Quality Modification Concept

Hypothesis 4 is an extension of Marsh's Positive and Negative Thresholds for Trust [16] and Kosa's Proposed Thresholds for Privacy [18] as applied to Information Quality. It is expected that trade-offs are present between specific trust characteristics, information quality dimensions, and data privacy aspects

found within the proposed framework matrix. It should also be noted that any modification of Accessibility IQ dimensions mitigates privacy and trust concerns by changing the visibility of a given piece of information rather than changing the shared information itself. For example, certain demographics (e.g. birthdate) or events may be willingly disclosed if limited to close friends, but if a user knows that these same demographic/events will be made public they are more likely to modify or withhold the information.

RATIONALE & PURPOSE OF THE CONFORMED DIMENSIONS

In general, we believe the following rationale should encourage the adoption of a conformed cross-industry standard set of DQ dimensions:

- If the dimensions are created with the purpose of 'communicating' the characteristics of data then why would we want there to be dimensions with conflicting definitions, or overlapping terminology. The answer is that having a single generally agreed-upon standard is preferred.

- Arguing about what should be in a set of enterprise, or even department level, DQ dimensions wastes time and confuses people who are beginning to learn about DQ. With a standard set of dimensions, organizations can skip over the first wave of arguments and can begin using the terminology and concepts to measure data quality from day one.

- As new concepts are defined and used there are reasons to protect one's intellectual property, which in this case of the dimensions of data quality ended 20 or more years ago. There is little value in using one author's version over another, and on the contrary, use of the most easily understood and conceptually robust is desirable.

- One of the primary reasons that methodologies such as Six Sigma and Lean are so valuable is that they seek to maximize repeatability and standardize (e.g. Lean/Kaizen 4th S, Seiketsu), inputs, outputs, and processes. Any organization that is test-and-learn focused is also focused on the scientific method of controlling for change in an environment in order to measure change and thereby improvement. This is nearly impossible without standardization, such as provided by the Conformed Dimensions of Data Quality.

- If organizations use this set of Conformed Dimensions of Data Quality over an extended period, then comparison between departments, companies and even perhaps industries may be feasible.

This paper's comparison of the CDDQ and Lee, Pipino, Funk, and Wang [7] serves to both validate the benefits of the existing conformed dimensions and to highlight possible extensions of the conformed dimensions needed for application in contexts that require the inclusion of subjective dimensions of data quality.

METHODS

This research compares the Conformed Dimensions of Data Quality, a proposed standard cross-industry set of dimensions of data quality, to the data quality dimensions previously espoused by Wang & Strong [13] and Lee, Pipino, Funk & Wang [15]. Further, these Conformed Dimensions of Data Quality are evaluated in application to a proposed Information Quality-Privacy-Trust research framework that currently utilizes the Lee, Pipino, Funk and Wang data quality dimensions.

COMPARATIVE FINDINGS

The first task in this research was to compare the proposed CDDQ dimensions to the list of data quality dimensions espoused by Lee, Pipino, Funk & Wang [7][15]. As noted previously, the CDDQ were first conceived of in 2013 [4] and published online in 2016. In this previous research, data quality dimensions from Lee, Pipino, Funk & Wang as presented in Journey to Data Quality were considered. Because of this, many definitions presented in our comparison are very similar.

Concepts covered by Pipino, Lee, and Wang, but not found in CDDQ

1. **Schema Completeness:** Pipino, Lee, and Wang include a metric called Schema Completeness, the Conformed Dimensions do not include this as an underlying concept for a few reasons:

    - Before relational databases popularized the concept of a schema, simple 2-D tabular datasets (aka VSAM or text files) were the construct of choice and, short of a folder structure or naming convention, didn't include another layer of hierarchy. So, concepts needed to only handle rows and columns.
    - Ironically, now in a modern era of non-relational solutions such as, NoSQL, Key-Value, and Documents repositories, schemas aren't of relevance. So, like in the early days, only rows and columns are required.
    - Lastly, but even of more distinction, because entities (tables) are composed of attributes (columns) and records (rows), and both of those are addressed with Attribute Population and Record Population (addressed in the Conformed Dimensions) there seems to be no need for Schema Completeness. Said another way, the CDDQ treats a missing table the same as an empty table (from a "data" completeness perspective). In this context, Completeness is a measure of data not of schema (metadata). If we want to measure metadata, then the CDDQ proposes use of the Representation dimension and underlying concept of Metadata Availability.

2. **Accuracy:** Although the Conformed Dimensions don't include an underlying concept explicitly stating 'Free of Error' this is equivalent with the CDDQ underlying concept of "Agree with the Real-World.

Concepts covered in CDDQ, but not found in Pipino, Lee, and Wang:

1. **Completeness:** Although Pipino, Lee, and Wang [15] call out a separate dimension for Appropriate Amount of Data, it isn't clear whether that means count of attributes needed, count of rows needed or something else. The Conformed Dimensions covers the two former meanings and includes another underlying concept called Existence, which is used to measure whether "real-life facts have been captured as data".

2. **Accuracy - Match to Agreed Source:** From the beginning Wang and Strong [13] espoused a dimension called Accuracy and based on prior research [4] most authors identify Accuracy as containing two primary concepts: 'Agreement with the Real-world' and 'Matching to Agreed Source' (see Table 8). For that reason, the Conformed Dimensions include both as underlying concepts, whereas Pipino, Lee, and Wang [15] only include the former equivalent (Free-of-Error).

3. **Consistency:** Although Pipino, Lee, and Wang [15] identified the Consistent Representation dimension, the metrics didn't include the most commonly understood concept, Referential Consistency, until later publication by Lee, Pipino, Funk, and Wang [7]. The Conformed Dimensions includes this as 'Equivalence of Redundant or Distributed Data,' which simplifies the

language into a form that lay persons are more likely to understand. Note that Referential Integrity still remains under the integrity dimension.

	Agree with Real-World	Match to Agreed Source	Precision of Data Value	Values in Specified Range of Valid Values
Tom Redman [27]	"Agree with the real world"	"Source agreed to be correct"		
Larry English [28]	"Characteristic of real world"	"Accuracy to surrogate source"	"Accuracy and precision represent the highest degree of inherent IQ possible"	
TDWI [29]	"Matches reality"			
DMBOK 1 [30]	"Represents real-life entities"	"Values agree with an identified reference source"		
David Loshin [31]		"Definition of system of record"	"Precision of data value"	"Domain Definition"
Lee, Pipino, Funk, Wang [15]	"Free of error"			

Table 8: Concepts within the Accuracy Dimension

4. **Integrity:** Similarly, Lee, Pipino, Funk, and Wang [7] call out Codd's Integrity constraints in their 2006 work, although not exactly as a dimension. The Conformed Dimensions has added these as underlying concepts with the Integrity dimension adding detailed descriptions for each.

5. **Validity:** Lee, Pipino, Funk, and Wang [7] call out a concept under Integrity relating to business rules, but the Conformed Dimensions goes further to name a separate dimension for Validity and places this and other underlying concepts (see below) based on ideas found in industry [4]. Most authors that discuss the dimensions of data quality include a dimension for validity [32][30]. Redman originally placed the concept inside of consistency, but later agreed that it fits into its own dimension [4]. Loshin [4] placed them within Accuracy and Batini and Scannapieco [26] include validity within the Accuracy Cluster.

   The Conformed Dimensions specifically define the Validity dimension as a "measure of whether a value conforms to a present standard" and includes the following underlying concepts:

   - **Values in Specified Range**- Values must be between some lower number and some higher number.
   - **Values Conform to Business Rule**- Validity measures whether values adhere to some declarative formula.
   - **Domain of Predefined Values**- This is a set of permitted values.
   - **Values Conform to Data Type**- Validity measures whether values have a specific characteristic (e.g. Integer, Character, Boolean). Data types restrict what values can exist, the operations that can be use on it, and the way that the data is stored.
   - **Values Conform to Format**- Validity measures whether the data are arranged or composed in a predefined way.

	Values in Specified Range of Valid Values	Values Conform to Business Rule	Conform to Other Attribute Types
Tom Redman [27]		In Consistency as "Degree to which a set of data satisfies business rules"	
Larry English [28]	"Values in Specified Range of Valid Values"	"Values Conform to Business Rule" and "Derivation Correct"	
TDWI [29]			
DMBOK 1 [30]	"Consistent with domain of Values"		"Values conform to numerous attributes associated: data type, precision, format, etc."
David Loshin [31]	In Accuracy as "Value Acceptance"		
Lee, Pipino, Funk, Wang [15]		In Integrity as "Codd introduced a fifth, all-purpose category that he labeled business rules."	

Table 9: Concepts within the Validity Dimension

6. **Currency and Timeliness:** The Conformed Dimensions call out a separate dimension called Currency for what Pipino, Lee, and Wang call Timeliness. After review of other author's underlying concepts, it made more sense for the Conformed Dimensions to align these as Currency.

	Current with World it Models	Concurrence of Distributed Data
Tom Redman [27]	"Current if up-to-date"	
Larry English [28]	"Age is correct for purpose"	"Lag time between when data in system(a) is queried in system(b)"
TDWI [29]	"Lag between business event and data record"	
DMBOK 1 [30]	"Info current with world it models"	
David Loshin [31]	"Age/Freshness"	"Synchronization/replication"
Lee, Pipino, Funk, Wang [15]	Included in Timeliness as "How up-to-date the data is"	

Table 10: Concepts within the Currency and Timeliness Dimensions

The Conformed Dimensions then define Timeliness as, "a measure of time between when data is expected versus made available," with the following three underlying concepts.

- **Time Expectation for Availability-** The measure of time between when data is expected versus made available.
- **Manual Float-** Manual float is a measure of the time from when an observation is made to the point it is recorded in electronic format.

- **Electronic Float-** Electronic float is a measure of the time from when data is captured in an electronic format until it is accessed by a person.[2]

Finally, Appendix A presents the full matrix of the dimensions espoused by Lee, Pipino, Funk, and Wang as mapped to the most appropriate Conformed Dimension and associated Underlying Concept.

APPLICATION FINDINGS

The second task in this research is to evaluate the Conformed Dimensions of Data Quality in application to a proposed Information Quality-Privacy-Trust research framework that currently utilizes the Lee, Pipino, Funk, and Wang data quality dimensions.

Information Quality-Privacy-Trust Research Framework Matrix

For this aspect of our research, the findings from our comparative analysis were applied to this existing research framework. A summarized version of this is presented in Table 11 In addition, a full research framework with mapping of conformed dimensions can be found in Appendix B.

	Existing Dimensions	Conformed Dimensions	Conformed Dimensions (Underlying Concepts)
Information Quality Dimensions	Accuracy	Accuracy	Accuracy
	Appropriate Amount	Completeness	Completeness (Existence / Appropriate Amount)
	Completeness	Accessibility	Completeness (Attribute/Record)
	Accessibility	Representation	Accessibility (Access Control / Security)
	Security	Consistency	Accessibility (Ease of Obtaining)
	Concise Representation	Timeliness	Representation (Understandability)
	Consistent Representation	Currency	Representation (Interpretability)
	Timeliness		Representation (Concise Representation)
	Understandability		Consistency (Consistent Representation)
	Interpretability		Timeliness
	Believability		Currency
	Objectivity	Validity	Validity
	Relevancy	Integrity	Integrity
	Reputation	Lineage	Lineage
	Ease of Manipulation	Believability	Believability
		Objectivity	Objectivity
		Relevancy	Relevancy
		Reputation	Reputation
		Ease of Manipulation	Ease of Manipulation

Comparison Legend	
Direct CDDQ to Pipino, Lee, & Wang Mapping	In CDDQ, Not in Pipino, Lee, & Wang
Subjective Attribute (Excluded from CDDQ)	System Attribute (Excluded from CDDQ)

Table 11: Non-Conformed and Conformed IQ Dimensions within the Research Framework

Most IQ dimensions, specifically the more objectively measurable dimensions, map directly between Pipino, Lee, and Wang and the Conformed Dimensions of Data Quality. In some cases, multiple dimensions from the existing framework are combined into a single Conformed Dimension. These mappings can be displayed in either summarized form or expanded form through reference to the relevant underlying concepts. Further, the CDDQ makes three additional dimensions available (Validity, Integrity,

[2] The Electronic Float underlying concept was added to the CDDQ in release 3.4.

and Lineage). It should be mentioned that Lee, Pipino, Funk, and Wang [7] discuss Integrity in relation to Codd's Integrity constraints, but they do not include it as a dimension in their listing of data quality dimensions.

There are two areas where CDDQ does not map well to the needs of the existing Information Quality-Privacy-Trust research framework matrix. These include the more subjective dimensions of Believability, Objectivity, Relevancy, and Reputation that are excluded by CDDQ Principle #1 and the system related dimensions of Ease of Manipulation, and possibly Security, that are excluded by CDDQ Principle #2. It may be possible to treat Ease of Manipulation as an aggregate measure of available dimensions such as Accessibility, Understandability, Interpretably, Completeness, and Consistent Representation that would correspond to aspects that represent the manipulability of data. In turn, Security may be mapped to the underlying concept of Access Control with Accessibility. This could resolve the issues with system related dimensions, but how to approach the subjective dimensions needed for this research framework is still in question.

DISCUSSION

As part of the development of the Conformed Dimensions of Data Quality, six authors' definitions of the dimensions of data quality were compared [4]. One of those six works, used as a foundation, included the dimensions proposed by Pipino, Lee, and Wang [15]. Because of this, we would expect the Conformed Dimensions to be relatively complete in our comparison. As expected, they were found to align well. The major exception to this are the subjective dimensions of Believability, Objectivity, Relevancy, and Reputation that were intentionally excluded based on the expressed CDDQ principles.

Regarding our Application Findings, the detailed definitions and underlying concepts found within the CDDQ are of benefit. Inclusion of the relevant underlying concepts that go with each dimension within the Information Quality-Privacy-Trust research framework allow for relationship details to be highlighted within the matrix that were not directly shown by listing the names of the dimensions alone. This may offer a reason to use the CDDQ instead of the Pipino, Lee, and Wang [15] set of dimensions. The inclusion of related underlying concepts and general metrics in the CDDQ framework will be of benefit in future aspects of the planned Information Quality-Privacy-Trust research.

We believe the Application Findings show that, for certain situations, it may be beneficial to include the subjective dimensions of data quality. For the Conformed Dimensions of Data Quality to be more complete, these subjective dimensions, where possible, should also be addressed in some way. This may be accomplished by directly adding subjective dimensions or through the identification of underlying objective dimensions that may be combined to define a subjective measure. Additional research and validation is still required, but in our initial ideas regarding subjective dimensions in presented in the following section

There is also a question as to the need for a Conformed Dimension related to system issues. Our Application Findings noted this as possible issue when determining on how to define and map Ease of Manipulation to the existing CDDQ framework. This question also arises in research [21] comparing CDDQ to ISO dimensions of data quality such as Portability, Recoverability, and Efficiency.

Subjective versus Objectives Dimensions

Table 12 includes definitions for Subjective and Objective, both generically using an English dictionary and as used in this discussion about the dimensions of data quality. By definition, objective dimensions can only be based on objective measures (e.g. count of null values). A subjective dimension can be a human defined measure based on experience and opinion or it may also have its inputs based on completely objective measures. For example, we can define the Believability dimension as a composite

measure of recorded levels of Consistency greater than 98% for greater than 1 month, and Validity at 99.9% for 2 months, where Consistency and Validity are objective dimensions.

	Subjective	Objective
Oxford English Dictionary	Based on or influenced by personal feelings, tastes, or opinions.	Not influenced by personal feelings or opinions in considering and representing facts.
IQ Domain Specific	"Subjective data quality assessments reflect the needs and experiences of stakeholders: the collectors, custodians, and consumers of data products" [15] citing [Ballou et al, 1998 and Wand and Wang, 1996]	"Objective measurements based on the data set in question" [15]
Source of Measures	Human verbal (usually written) input	Defined by humans, but of logical construction that can be repeated, programmed, and executed without human input

Table 12: Subjective verses Objective Definitions

Because, in practice, each organization tends to define subjective dimensions in a different way, the CDDQ does not include definitions for subjective dimensions such as Believability, Relevance, Reputation, etc. Instead of directly including a subjective dimension, it is recommended that organizations define an enterprise standard definition of the desired subjective dimension based on underlying objective dimensions found within the CDDQ.

For example, Lee, Pipino, Funk, and Wang identify two underlying concepts within the Believability dimension:

1. Measures collected via survey of data consumer's opinion about the quality of the data [15]
2. Measures defined as a function of other measures [7, p.57]

In both scenarios listed above, it is preferable to maintain scientific rigor by basing subjective measures, both those taken through surveys and through composite functions, upon defined objective measures. In accrual accounting, "the matching principle states that expenses should be recorded during the period in which they are incurred, regardless of when the transfer of cash occurs." [33] In the same manner, information quality related survey questions (however subjective due to human opinion) should to the extent possible be based upon objective measures. Collection of data consumers' subjective opinions of quality should be based upon thoroughly documented objective measures (preferably the CDDQ) and compared with objective measures collected computationally (aka with a data profiler). This ensures a one-to-one comparison.

Defining Subjective Dimensions of Data Quality

Pipino, Lee, and Wang [15] cite five subjective dimensions of data quality- Believability, Objectivity, Reputation, Relevance, and Value-added. Later work by Lee, Pipino, Funk, and Wang reduced that to only one, Believability [7]. Of these five dimensions, Believability is by far the most well researched [15, 7, 25, 31, 34]. So, if research that requires the inclusion subjective dimensions, such as the Information Quality-Privacy-Trust matrix, is to rely upon the CDDQ as the framework of dimensions of data quality, Believability is an important place to start.

A comparison of underlying concepts within Believability and Reasonability is presented in Table 13. In this we see that the most cited underlying concept within Believability is whether the data is from an authoritative source. Of the eleven, or more, authors that discuss the dimensions of data quality researched for this paper, four included concepts relating to the source. Second only to that number, was

the suggestion that Believability (and other subjective dimensions) are a function of multiple variables. These sources either directly stated this [7][34] or listed multiple metrics used to evaluate the Believability dimension which implicitly supports the fact that they believe this dimension to be a composite, requiring a number of facets. The third concept, temporal consistency and temporal validity were cited almost as many times, but with a wide variation of time-related aspects (see Appendix D which lists all the authors' definitions of each dimension, and associated metrics). The last two concepts identified didn't have as much consensus, only two authors, but make a lot of sense.

Author/ Source	Dimension Named	Concept 1	Concept 2	Concept 3	Concept 4	Concept 5	Concept 6
		From Authoritative Source	**Function of Multiple Variables**	**Temporal Consistency**	**Temporal Validity**	**Consistency between sources**	**Likely/ Possible (Subjective)**
Lee et al [7]	Believability		Function of multiple variables				
DMBOK 2 [34]	Reasonability		[Composite of subjective]	Past instances of a similar data set	Based on comparison to benchmark data		
Prat and Madnick [25]	Believability	Originates from trustworthy sources	[Composite]	Consistency over time	Based on proximity of transaction time to valid times	Consistency over sources	Likely/ Possible
Loshin [31]	Reasonable-ness	Agreements-governing data provider performance	[Composite]	Temporal reasonability		Multi-value consistency	Data meet rational expectations
Batini & Scannapieco [35]	Trust	Info derives from an authoritative source					
ISO/IEC 25012:2008 [38]	Credibility	Truthfulness of origins, attributions, commitments					

Table 13: Comparison of Underlying Concepts within Believability and Reasonability

As stated earlier during the explanation of CDDQ principle number 1 (as shown in Table 2), the CDDQ only contains objective, quantifiable, criteria that enable repeatability through standardization. To the extent that each of the underlying concepts above are objective, in that they can be explicitly defined and quantified without human judgement, then an organization may optionally include the Believability dimension as defined below. Note that concept 6, Likely/Possible was removed from the proposed definition due to its subjective and abstract nature that is difficult to quantify in a standardized way.

Dimension	Definition of Dimension	Underlying Concept	Definition of Underlying Concepts
Believability	Believability defines whether the data are from an authorized source; have temporal validity and display consistency between sources.	From Authoritative Source	Data originates from a trustworthy source defined as the system of record.
		Temporal Consistency	The extent to which a data value is consistent with other values of the same data over time.
		Temporal Validity	The extent to which a data value falls within a set of valid times.
		Consistency between sources	The extent that the data is equivalent across different providers.

Table 14: Proposed Definition of Believability Dimension

Unlike the other objective dimensions within the CDDQ, as an inherently subjective dimension, Believability is based on the principal that the dimension itself is a composite of other measures. This means that Believability is best calculated when all its underlying inputs are available. So, we may refer to *Attribute Population* (an underlying concept of Completeness) as a metric, but we would likely not do the same by using only *From Authoritative Source* as the basis for Believability. Rather, only if data is sourced from an authoritative system *and* has temporal consistency *and* temporal validity *and* is consistent between sources, would we say that it is Believable.

Ongoing research is also underway regarding the additional four subjective dimensions proposed by Pipino, Lee, and Wang [15] in 2002, but the authors expect diminishing returns from a practical perspective if those have fewer underlying concepts in agreement between practitioners and researchers in the field. Appendix E includes this information for the Relevance dimension.

LIMITATIONS

The primary limitation of this paper is that it presents research-in-progress. It is meaningful to both research efforts to perform this comparison. Key benefits that move discussions within the information quality field forward are highlighted, but our findings will have more weight and broader application as increased usage of CDDQ is documented and future research efforts formally validate the Information Quality-Privacy-Trust research framework.

FUTURE RESEARCH

Regarding the Conformed Dimensions of Data Quality, we have proposed an additional dimension, Believability, which though most often referred to in its subjective context, can be defined in a composite manner based on objective inputs from existing underlying concepts found in the CDDQ. Other typically subjective measures may need to be defined in terms of the CDDQ over time. Second, the question regarding if and how system related dimensions should be approached by the CDDQ needs to be addressed. [21] Having said this, the robustness of the Conformed Dimensions of Data Quality will also continue to increase as additional published quality dimension frameworks are evaluated and user feedback regarding the current Conformed Dimensions is incorporated.

Regarding the Information Quality-Privacy-Trust research framework, this dissertation research is ongoing. To date, relationship matrices for data privacy, online social networks, information quality, and trust as a research framework have been developed and presented [1][2]. A validation survey for the research framework has been developed and implementation for select professionals and topic experts is pending. For the next phase of this research, a structural equation model for understanding the trade-offs and influences between data privacy, trust, and information quality in online social networks is being developed. A survey will also be undertaken to validate the model. Future research application is likely to include expanded validation of different areas of overlap within framework matrices. It would be of interest to explore application of this research beyond the current focus of user-controlled aspects such as Disclosed, Entrusted, and Incidental data to include Service, Behavioral, and Derived data within online social networks and third-party venders.

Further, the current research has shown several benefits in the utilization of the CDDQ in terms of standardized definitions and underlying concepts. Additional research is being considered into how to best apply these benefits to the Information Quality-Privacy-Trust research framework going forward.

CONCLUSIONS

This research confirms that the Pipino, Lee, and Wang [15] data quality dimensions map well to the Conformed Dimensions of Data Quality in both direct comparison and when evaluated in application. We found, though, in applications that utilize more subjective dimension of data quality, the CDDQ requires users to define composite subjective dimensions from the underlying objective dimensions of data quality available in the framework. As an example of this, we present a proposed extension to the CDDQ using the Believability dimension. We also consider that there may be a need to address information system level quality attributes within the CDDQ and propose future research to better understand this issue.

REFERENCES

[1] B. Blake and N. Agarwal, "Modeling User-Based Modifications to Information Quality to Address Privacy and Trust Related Concerns in Online Social Networks," International Journal On Advances in Security, Sec17v10n12, 2017. Online.

[2] B. Blake and N. Agarwal, "Understanding User-Based Modifications to Information Quality in Response to Privacy and Trust Related Concerns in Online Social Networks," The Sixth International Conference on Social Media Technologies, Communication, and Informatics (SOTICS), pp.18-28, 2016.

[3] D. Myers, "Conformed Dimensions of Data Quality," DQMatters, 2017. [Online]. Available from: http://dimensionsofdataquality.com 2017.05.29

[4] D. Myers, "The Value of Using the Dimensions of Data Quality." Information Management, Aug. 2013. [Online]. Available from: https://www.information-management.com/news/the-value-of-using-the-dimensions-of-data-quality 2017.05.29

[5] D. Myers, "We are READY for a Standard Set of Dimensions of Data Quality" Conformed Dimensions of Data Quality Website, April 2015. [Online]. Available from: http://dqm.mx/blakemyers2017-5 2017.06.08

[6] D. Myers, "2016 Annual Report on the Dimensions of Data Quality Year-two: General Usage Improves but Confusion Remains" Conformed Dimensions of Data Quality Website, 2016. [Online]. Available from: http://dqm.mx/blakemyers2017-6 2017.06.08

[7] Y.W. Lee, L. L. Pipino, R. Y. Wang, and J. D. Funk, Journey to Data Quality, MIT Press, 2006

[8] P. Bertini, "Trust Me! Explaining the Relationship Between Privacy and Data Quality," Information Technology and Innovation Trend in Organization, 2010. [Online]. Available from: http://www.cersi.it/itais2010/. 2017.05.29.

[9] D. J. Solove, Understanding Privacy. Cambridge, MA: Harvard University Press, 2008

[10] B. Schneier, "A Taxonomy of Social Networking Data," IEEE Security & Privacy Magazine, vol. 8, no. 4, p. 88, 2010, doi: 10.1109/MSP.2010.118

[11] S. E. Madnick, R. Y. Wang, Y. W. Lee, and H. Zhu, "Overview and Framework for Data and Information Quality Research," Journal of Data and Information Quality, vol. 1, pp. 2:1-2:22, 2009.

[12] C. Fisher, E. Lauria, S. Chengalur-Smith, R. Wang, Introduction to Information Quality, M.I.T. Information Quality Program, 2006.

[13] R. Y. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," Journal of Management Information Systems, vol. 12, no. 4, pp. 5-33, 1996.

[14] D. M. Strong, Y. W. Lee, and R. Y. Wang, "Data Quality in Context," Commun. ACM, vol. 40, pp. 103-110, May 1997.

[15] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data Quality Assessment," Commun. ACM, vol. 45, pp. 211-218, Apr. 2002.

[16] S. P. Marsh, Formalising Trust as a Computational Concept, unpublished doctoral dissertation, University of Stirling, 1994. [Online]. Available from: https://dspace.stir.ac.uk/ 2017.05.29.

[17] D. Gefen, "Reflections on the Dimensions of Trust and Trustworthiness Among Online Consumers," SIGMIS Database, vol. 33, pp. 38-53, 2002.

[18] T. Kosa, "Vampire Bats: Trust in Privacy," Eighth Annual International Conference on Privacy Security and Trust (PST), 2010, pp. 96-102, doi: 10.1109/PST.2010.5593227.

[19] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An Integrative Model of Organizational Trust," The Academy of Management Review, vol. 20, no. 3, pp. 709-734, 1995.

[20] G. Hogben (Ed.), ENISA Position Paper No. 1: Security Issues and Recommendations for Online Social Networks, European Network and Information Security Agency, Nov. 2007. [Online]. Available from: https://www.enisa.europa.eu/publications/archive/security-issues-and-recommendations-for-online-social-networks 2017.05.29

[21] D. Myers, Mapping the ISO Dimensions of Data Quality to the Conformed Dimensions of Data Quality (CDDQ), Manuscript in preparation, 2017.

[22] R. Price and G. Shanks, "A Semiotic Information Quality Framework: Development And Comparative Analysis", Journal of Information Technology, 20.2, pp. 88-102, 2005.

[23] R. Price and G. Shanks, "Empirical Refinement of a Semiotic Information Quality Framework, in Proceedings of Hawaii International Conference on System Sciences (HICSS38), Silver Spring, MD: IEEE Computer Society Press, pp.1–10, 2005.

[24] A. AbuHalimeh and M.E. Tudoreanu, "Subjective Information Quality in Data Integration: Evaluation and Principles", Information Quality and Governance for Business Intelligence, IGI Global, pp.44-65, 2013

[25] N. Prat and S. Madnick, Measuring Data Believability: A Provenance Approach, MIT Sloan Research Paper No. 4672-07. 2007, [Online]. Available from: http://dx.doi.org/10.2139/ssrn.1075723. 2017.06.09.

[26] C. Batini and M. Scannapieco, Data and Information Quality- Dimensions, Principles and Techniques, Springer International, Switzerland, 2016.

[27] T. Redman, The Field Guide, Digital Press, 2001.

[28] L. English, Information Quality Applied, Wiley Publishing, 2009.

[29] Data Quality Fundamentals, The Data Warehousing Institute (TDWI), 2011.

[30] DAMA International, The DAMA Guide to The Data Management Body of Knowledge (DMBOK), Technics Publications, 2009.

[31] D. Loshin, The Practitioner's Guide to Data Quality Improvement, Elsevier 2011.

[32] L. English, Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits, Wiley, 1999.

[33] Wikipedia.com, "Matching Principle", [Online]. Available from: https://en.wikipedia.org/wiki/Matching_principle 2017.08.10

[34] DAMA International, The DAMA Guide to The Data Management Body of Knowledge (DMBOK) 2, Technics Publications, 2017.

[35] C. Batini, M.Cannapieco, Data and Information Quality- Dimensions, Principles and Techniques, Springer, 2016.

[36] T. Redman, Information Age, ARTECH HOUSE, 1997.

[37] H. Miller, The Multiple Dimensions of Information Quality, Information Systems Management (Journal), Vol.13, Issue 2, 1996.

[38] ISO/IEC 25012:2008, Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model, 2008.

Appendix A

#	Dimension	Definition	Metrics	Metric Definition	Metric Definition	Metric Definition	Metric Definition	Metric Definition	Metric Definition
1	Accessibility * †	[Accessibility is] the extent to which data is available or easily and quickly retrievable. *	Is available	<Authors do not specifically define>	Accessibility rating = {max[(1-(Interval of time from request by user to delivery to user / Interval of time from request to time at which no longer of any use),0]}[x]	Completeness or Accessibility or Timeliness	Accessibility measures how easy it is to acquire data when needed, how long it is retained, and how access is controlled.	Dimension = Completeness and Underlying Concept = Existence	Existence identifies whether a real-life fact has been captured as data.
			Easily retrievable	<Authors do not specifically define>				Dimension = Accessibility and Underlying Concept = Ease of Obtaining Data	This measures how easy it is to obtain data.
			Quickly retrievable	<Authors do not specifically define>				Dimension = Timeliness and Underlying Concept = Time Expectation of Availability	The measure of time between when data is expected versus made available.
2	Appropriate Amount of Data * †	[Appropriate Amount of Data is] the extent to which the volume of data is appropriate for the task at hand. *	Appropriate Amount of Data	<Authors do not specifically define>	Rating of appropriate amount of data = min [(Number of data units provided / Number of data units needed) , (Number of data units needed / Number of data units provided)]	Completeness	Completeness measures the degree of population of data values in a data set.	Existence	Existence identifies whether a real-life fact has been captured as data.
3	Believability * †	[Believability is] the extent to which data is regarded as true and credible. *	Survey	Subjective rating obtained as part of the IQA survey described in chapter 3.	Believability = min(Believability of source, Believability when compared to internal commonsense standard, Believability based on age of data)	**None, see CDDQ Principle #1. Objective Focus**			
			Function of Multiple Variables	Alternatively, one might wish to define believability as a function of multiple variables.					
4	Completeness * †	[Completeness is] the extent to which data is not missing and is of sufficient breadth and depth for the task at hand.*	Schema Completeness	By schema completeness, we mean the degree to which entities and attributes are not missing from the schema.	Completeness rating = 1- (Number of incomplete items / Total number of items)	Completeness	Completeness measures the degree of population of data values in a data set.	**Note:** Because entities (tables) are composed of attributes (columns) and records (rows), and both of those are addressed with Attribute Population and Record Population (see below) there is no need to identify whether entities themselves are missing from a schema. Additionally, in modern non-relational contexts databases are often schema-less.	
			Column Completeness	By column completeness, we mean the degree to which there exist missing values in a column of a table.				Attribute Population	This measures whether a value is present (not null) for an attribute (column).
			Population Completeness	By population completeness, we mean the degree to which members of the population that should be present are not present. For example, if a column should contain at least one occurrence of all 50 states, but the column contains only 43 states, then the population is incomplete.				Record Population	This measures whether a row is present in a data set (table).
5	Concise Representation *	[Concise Representation is] the extent to which data is compactly represented. *	Compactly Represented	<Authors do not specifically define>	<Authors do not specifically define>	**None, see CDDQ Principle #2. Independent of System**			
6	Consistency † Consistent Representation *	[Consistent Representation is] the extent to which data is presented in the same format. *	Referential Integrity Constraint	[Referential]...consistency of redundant data in one table or in multiple tables. Codd's referential integrity constraint is an instantiation of this type of consistency.	Consistency rating = 1- (Number of instances violating specific consistency type / Total number of consistency checks performed)	Consistency	Consistency measures whether or not data is equivalent across systems or location of storage.	Dimension=Integicy and Underlying Concept= Referential Integrity	Referential integrity measures whether if when a value (foreign key) is used it must reference an existing key (primary key) in the parent table.
			Two Related Data Elements	[Logical]...consistency between two related data elements. For example, the name of the city and the postal code should be consistent.				Logical Consistency	Logical consistency measures whether two attributes of related data are conceptually in agreement, even though they may not record the same characteristic of a fact.
			Consistency of Format	[Format]...consistency of format for the same data element used in different tables.				Format Consistency	This measures the conformity of format of the same data in different places.
7	Ease of Manipulation *	[Ease of Manipulation is] the extent to which data is easy to manipulate and apply to different tasks. *	Easy to Manipulate	<Authors do not specifically define>	<Authors do not specifically define>	**None, see CDDQ Principle #2. Independent of System**			
8	Free of Error * †	[Free of Error is] the extent to which data is correct and reliable. *	Free-of-error Rating	Dimension that represents whether the data is correct. †	Free-of-error rating = 1 - (Number of data units in error/ Total number of data units)	Accuracy	Accuracy measures the degree to which data factually represents its associated real-world object, event, concept or alternatively matches the agreed upon source(s).	Agree with Real-world	Degree that data factually represents its associated real-world object, event, or concept.
9	Interpretability *	[Interpretability is] the extent to which data is in appropriate languages, symbols, and units, and the definitions are clear.	Appropriate Language and Symbols	<Authors do not specifically define>	<Authors do not specifically define>	Representation	Representation measures ease of understanding data, consistency of presentation, appropriate media choice, and availability of documentation (metadata).	Presentation Language	Data that is represented well is simple but elegantly formed with good grammar and presented in a standard way.
			Appropriate Units					Includes Measurement Units	Well represented data includes the scale of measurement, such as weight, height, distance…etc.
			Definitions are Clear					Metadata Availability; Easy to Read & Interpret	**Metadata Availability**- Comprehensive descriptions and other information about the characteristics of the data are provided in plain language. **Easy to Read & Interpret**- Illustrations and charts should be self-explanatory and presented with appropriate labels, providing context.
10	Objectivity *	[Objectivity is] the extent to which data is unbiased, unprejudiced, and impartial.	Unbiased, Unprejudiced and Impartial	<Authors do not specifically define>	<Authors do not specifically define>	**None, see CDDQ Principle #1. Objective Focus**			
11	Relevancy *	[Relevancy is] the extent to which data is applicable and helpful for the task at hand.	Applicable and Helpful	<Authors do not specifically define>	<Authors do not specifically define>	**None, see CDDQ Principle #1. Objective Focus**			
12	Reputation *	[Reputation is] the extent to which data is highly regarded in terms of its source or content.	Highly Regarded in Source and Content	<Authors do not specifically define>	<Authors do not specifically define>	**None, see CDDQ Principle #1. Objective Focus**			
13	Security *	[Security is] the extent to which access to data is restricted appropriately to maintain its security.	Access is Restricted	Extent to which access is restricted appropriately.	<Authors do not specifically define>	Accessibility	Accessibility measures how easy it is to acquire data when needed, how long it is retained, and how access is controlled.	Access Control	Access control includes the identification of a person that wants to access data, authentication of their identity, review and approval to access required data, and lastly auditing the access of that data.
14	Timeliness * †	[Timeliness is] the extent to which the data is sufficiently up-to-date for the task at hand. *	Timeliness	How up-to-date the data is with respect to the task for which it is used.	Timeliness rating = { max [(1- (Currency / Volatility),0)][x] from Ballou et al. (1998)	Currency	Currency measures how quickly data reflects the real-world concept that it represents.	Current with World it Models	Data is current if it reflects the present state of the concept it models.
15	Understandability *	[Understandability is] the extent to which data is easily comprehended.	Ease of Comprehension	<Authors do not specifically define>	<Authors do not specifically define>	Representation	Representation measures ease of understanding data, consistency of presentation, appropriate media choice, and availability of documentation (metadata).	Easy to Read & Interpret	Illustrations and charts should be self-explanatory and presented with appropriate labels, providing context.
16	Value-Added *	[Value-added is] the extent to which data is beneficial and provides advantages from its use.	Provides Advantages from Use	<Authors do not specifically define>	<Authors do not specifically define>	**None, see CDDQ Principle #1. Objective Focus**			
17	Integrity †	Codd integrity constraints consist of entity integrity, referential integrity, domain integrity, and column integrity. Codd introduced a fifth, all-purpose category that he labeled business rules. †	Entity	Entity integrity requires that no primary key field value in a table be null.	Degree of adherence to entity integrity =1 — (Number of null primary keys/Total number of rows)	Integrity	Integrity measures the structural or relational quality of data sets.	Uniqueness	Uniqueness measures whether each fact is uniquely represented.
			Referential	Rule states that the value of a foreign key in a table must match a value of a primary key in a designated related table, or the value of the foreign key must be null.	Degree of adherence to reference integrity = 1— (Number of nonmatching values excluding nulls in the dependent tableITotal rows in the dependent table)			Referential Integrity	Referential integrity measures whether if when a value (foreign key) is used it must reference an existing key (primary key) in the parent table.
			Domain or Column	Column integrity requires that the values in the column be drawn from the set of permissible values.	Degree of adherence to column integrity =1— (Number of invalid column values/Number of rows in table)	Validity	Validity measures whether a value conforms to a preset standard.	Domain of Predefined Values	This is a set of permitted values.

† references from Yang W. Lee, Leo L. Pipino, James D. Funk, Richard Y. Wang. Journey to Data Quality, MIT Press 2006

* references from: Leo L. Pipino, Yang W. Lee, and Richard Y. Wang, "Data Quality Assessment" Communications of the ACM, April 2002, Vol. 45. No. 4ve

Dimensions referenced by both the 2002 article and 2006 book are identified by the presense of both * and †

Appendix B

	Types of Social Networking Data					
	Service Data	Disclosed Data	Entrusted Data	Incidental Data	Behavioral Data	Derived Data
	Data you give the social network site in order to use it	What you post on your own pages	What you post on other people's pages	What other people post about you	Data the site collects about your habits by recording what you do and who you do it with	Data about you that is derived from all other data
Data Privacy Issues	Insecurity	Increased Accessibility	Increased Accessibility	Identification	Aggregation	Aggregation
	Secondary use	Insecurity	Secondary use	Exclusion	Insecurity	Insecurity
	Breach of Confidentiality	Appropriation	Identification	Breach of Confidentiality	Secondary Use	Secondary Use
		Secondary Use	Exclusion	Disclosure	Breach of Confidentiality	Breach of Confidentiality
			Breach of Confidentiality	Exposure	Identification	Identification
			Disclosure	Distortion	Exclusion	Exclusion
			Exposure	Intrusion (onto your pages)		
			Distortion	Increased Accessibility		
			Intrusion (onto their pages)	Secondary use		
Information Quality Dimensions	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy
	Accessibility (Ease of Obtaining)	Accessibility (Ease of Obtaining)	Accessibility (Ease of Obtaining)	Accessibility (Ease of Obtaining)	Completeness (Appropriate Amount / Existence)	Completeness (Appropriate Amount / Existence)
	Accessibility (Access Control / Security)	Accessibility (Access Control / Security)	Accessibility (Access Control / Security)	Accessibility (Access Control / Security)	Completeness (Attribute/Record)	Completeness (Attribute/Record)
	Currency	Completeness (Appropriate Amount / Existence)	Completeness (Appropriate Amount / Existence)	Completeness (Appropriate Amount / Existence)	Accessibility (Ease of Obtaining)	Accessibility (Ease of Obtaining)
	Completeness (Appropriate Amount / Existence)	Representation (Concise Representation)	Representation (Concise Representation)	Representation (Concise Representation)	Accessibility (Access Control / Security)	Accessibility (Access Control / Security)
	Completeness (Attribute)	Representation (Understandability)	Representation (Understandability)	Representation (Understandability)	Representation (Concise Representation)	Representation (Concise Representation)
	Representation (Concise Representation)	Timeliness	Timeliness	Timeliness	Representation (Understandability)	Representation (Understandability)
	Representation (Understandability)	Currency	Currency	Currency	Representation (Interpretability)	Representation (Interpretability)
	Consistency (Consistent Representation)	Believability	Believability	Believability	Timeliness	Timeliness
	Validity	Objectivity	Objectivity	Objectivity	Currency	Currency
	Relevancy	Relevancy	Relevancy	Relevancy	Validity	Validity
	Believability	Reputation	Reputation	Reputation	Integrity	Integrity
					Lineage	Lineage
					Believability	Believability
					Objectivity	Objectivity
					Relevancy	Relevancy
					Ease of Manipulation	Ease of Manipulation
Trust	Ability	Benevolence	Benevolence	Benevolence	Ability	Ability
	Benevolence	Integrity	Integrity	Integrity	Benevolence	Benevolence
	Integrity	Ability	Ability	Ability	Integrity	Integrity

Comparison Legend	
Direct CDDQ to Pipino, Lee, & Wang Mapping	In CDDQ, Not in Pipino, Lee, & Wang
Subjective Attribute (Excluded from CDDQ)	System Attribute (Excluded from CDDQ)

Appendix C, (CDDQ Release 3.3)

Full List of Underlying Concepts for CDDQ		
Conformed Dimension	**Underlying Concepts**	**Definition of Underlying Concept**
Completeness	Record Population	This measures whether a row is present in a data set (table).
	Attribute Population	This measures whether a value is present (not null) for an attribute (column).
	Truncation	This measures whether the value contains all characters of the correct value.
	Existence	Existence identifies whether a real-life fact has been captured as data.
Accuracy	Agree with Real-world	Degree that data factually represents its associated real-world object, event, or concept.
	Match to Agreed Source	Measure of agreement between data and the source of that data. This is used when the data represent intangible objects or transactions that can't be observed visually.
Consistency	Equivalence of Redundant or Distributed Data	The measure of similarity with other sources of data that represent the same concept.
	Format Consistency	This measures the conformity of format of the same data in different places.
	Logical Consistency	Logical consistency measures whether two attributes of related data are conceptually in agreement, even though they may not record the same characteristic of a fact.
Validity	Values in Specified Range	Values must be between some lower number and some higher number.
	Values Conform to Business Rule	Validity measures whether values adhere to some declarative formula.
	Domain of Predefined Values	This is a set of permitted values.
	Values Conform to Data Type	Validity measures whether values have a specific characteristic (e.g. Integer, Character, Boolean). Data types restrict what values can exist, the operations that can be use on it, and the way that the data is stored.
	Values Conform to Format	Validity measures whether the data are arranged or composed in a predefined way.
Timeliness	Time Expectation for Availability	The measure of time between when data is expected versus made available.
	Manual Float	Manual float is a measure of the time from when an observation is made to the point it is recorded in electronic format.
Currency	Current with World it Models	Data is current if it reflects the present state of the concept it models.
Integrity	Referential Integrity	Referential integrity measures whether if when a value (foreign key) is used it must reference an existing key (primary key) in the parent table.
	Uniqueness	Uniqueness measures whether each fact is uniquely represented.
	Cardinality	Cardinality describes the relationship between one table to another, such as one-to-one, one-to-many, or many-to-many.
Accessibility	Ease of Obtaining Data	This measures how easy it is to obtain data.
Accessibility Precision	Access Control	Access control includes the identification of a person that wants to access data, authentication of their identity, review, and approval to access required data, and lastly auditing the access of that data.
	Retention	Retention refers to the period of time that data is kept before being removed from a database through purge or archive processing.
	Precision of Data Value	The measure of preciseness of numeric data using decimal places, rounding and truncation.
Precision Lineage	Granularity	The detail or summary of data defines the granularity measured by the number of attributes used to represent a single concept.
	Source Documentation	Source documentation provides data provenance which describes the origin of the data.
Lineage Representation	Segment Documentation	Segment documentation provides how data is transformed and transported from one location to another.
	Target Documentation	Documentation about the target explains where the data moved to and how it is stored.
	End-to-End Graphical Documentation	End-to-End documentation provides diagrammatic visual representation of how the data flows from beginning to end.
Representation	Easy to Read & Interpret	Illustrations and charts should be self-explanatory and presented with appropriate labels, providing context.
	Presentation Language	Data that is represented well is simple but elegantly formed with good grammar and presented in a standard way.
	Media Appropriate	The appropriate media (e.g. Web-based, hardcopy, or audio, etc.) are provided.
	Metadata Availability	Comprehensive descriptions and other information about the characteristics of the data are provided in plain language.

Appendix D, Authors Citing Believability and Related Dimensions

Author [Reference]	Dimension	Dimension Definition	Metric	Metric Definition
Pipino et al [15]	Believability	[Believability is] the extent to which data is regarded as true and credible.	<authors don't define>	
Lee et al [7]	Believability	Believability is the extent to which data is regarded as true and credible.	Survey	Subjective rating obtained as part of the IQA survey described in chapter 3.
			Function of Multiple Variables	Alternatively, one might wish to define believability as a function of multiple variables.
Prat and Madnick [25]	Believability-Trustworthiness of Source	The extent to which a data value ordinates from trustworthy sources.		
	Believability-Reasonableness of Data	The extent to which a data value is reasonable (likely).	Possibility	The extent to which a data value is possible
			Consistency	**Definition:** The extent to which a data value is consistent with other values of the same data. **Consistency over sources:** a data value is possible **Consistency over time:** the data value is consistent with past data values
	Believability-Temporality of Data	The extent to which a data value is credible based on transaction and valid times.	Transaction and valid times closeness	The extent to which a data value is credible based on proximity of transaction time to valid times.
			Valid times overlap	The extent to which a data value is derived from data values with overlapping valid times.
Loshin [31]	Reasonableness	General statements associated with expectations of consistency or reasonability of values, either in the context of existing data or over a time series, are included in this dimension.	Multi-value consistency	The value of one set of attributes is consistent with the values of another set of attributes.
			Temporal reasonability	New values are consistent with expectations based on previous values.
			Agreements	Service level agreements (SLA), security agreements, and other authoritative documents governing data provider performance will be defined.
			Reasonableness	The data meet rational expectations.
			Data correction	When possible, poor data quality will be improved by implementing data correction processes.
DMBOK 1 [30]	Reasonableness		Consistency expectations	
DMBOK 2 [34]	Reasonability	Reasonability asks whether a data pattern meets expectations.	Comparison to benchmark data	
			[Comparison to] past instances of a similar data set	
			[Composite of subjective]	Some ideas about reasonability may be perceived as subjective. If this is the case, work with data consumers to articulate the basis of their expectations of data to formulate objective comparisons.
Batini & Scannapieco [35]	Trust	Including believability, reliability, and reputation, catching how much information derives from an authoritative source.		
ISO/IEC 25012:2008 [38]	Credibility	The degree to which data has attributes that are regarded as true and believable by users in a specific context of use.	Authenticity	The truthfulness of origins, attributions, commitments

Appendix E, Authors Citing Relevance as Dimension and Associated Metrics

Author [Reference]	Dimension	Dimension Definition	Metric	Metric Definition
Pipino et al [15]	Relevancy	[Relevancy is] the extent to which data is applicable and helpful for the task at hand.	Applicable Helpful	<authors don't define>
	Value-Added	[Value-Added is] the extent to which data is beneficial and provides advantages from its use.	<authors don't define>	
Redman [36]	Relevance	The view should provide data needed by the application.	<author doesn't define>	
English [32]	"Rightness" or Fact Completeness	The characteristic of having the right kind of data with the right quality to support a given process, such as to perform a process or support a decision.	Rightness	The measure rightness is an assessment of the percent of fact types, weighted, available out of the total fact types required to support a specific process.
Holmes Miller [38]	Relevance	The key component for information quality is whether the information addresses its customer's needs.	<author doesn't define>	
Kumar and Jakhar (Date unavailable)	Relevance	Relevancy is the degree of match between information being supplied and information required for making a decision.	<author doesn't define>	
Batini & Scannapieco [35]	Usefulness	Related to the advantage the user gains from the use of information.	<author doesn't define>	

Appendix F, Illustration for Table 1.

File Illustration:

Row-level Count = 3 (excluding header)

ITEM_ID	ITEM_NAME	SALE_ITEM	DIRECT_SHIP	ITEM_PRICE
001	T-SHIRT	Y		15.25
002	DIAPERS			33.12
003	PANTS	N	Y	24.95

Attribute/Column-level Count or Amount

ITEM_ID	ITEM_NAME	SALE_ITEM	DIRECT_SHIP	ITEM_PRICE
3	3	2	1	73.32