IQ International Journal

VOLUME 14 | ISSUE 1 | DECEMBER 2017

Information Quality Journal

A Publication of
IQ International

iqint.org

iqint.org

Contact Us

IQ International
6920 Brookmill Rd,
Baltimore MD 21215
USA
Email: editor@iqint.org

IQ International Journal Editorial Staff

Graphic Designer

VOLUME 14 | ISSUE 1 | DECEMBER 2017

In this Issue…

John Talburt, IQCP
IQ International Editor-in-Chief

From the Editor

I want to begin by thanking everyone who participated in and supported the MIT International Conference on Information Quality this past October. It was an outstanding event with many high-quality presentations. I especially want to thank Lwanga Yonke for his hard work in putting together the IQ International track at the conference. The industry speakers in the IQ International track provided a good balance to the papers presented in the research tracks. The conference attendees were fortunate to have so many session choices. Look for some of the highly-ranked papers from the conference to appear in future issues of the Journal. The research papers and industry presentations, not only from the 2017 conference, but also from previous conferences, are available online at the UA Little Rock Information Quality Graduate Program's website http://ualr.edu/informationquality/iciq-proceedings/

We were also fortunate to have such outstanding keynote speakers to start each day of the conference. On the first day of the conference, Pieter De Leenheer, Co-Founder and VP for Research and Education at Collibra gave a forward looking perspective on "Data Governance and Data Capitalization in the Big Data Era." This was followed on the second day by Doug Laney, VP and Distinguished Analyst at Gartner, discussing the emerging field of "Infonomics."

Finally, I would like to thank the many sponsors for their support of the conference and UA Little Rock Information Quality Graduate Program. In addition to IQ International, our sponsors included Collibra, Centene Charitable Foundation, PiLog Group, Simmons Bank, USAA, Black Oak Analytics, First Orion, and Fusion Alliance, Information Asset, and the UA Little Rock Institute for Chief Data Officers (iCDO).

Last but not least, I want to thank UA Little Rock staff members Collette Johnson, Lisa Garrett, and Devon Holiman for their assistance, and the many student volunteers who helped as guides, servers, and A/V assistants at the conference. Many of the student volunteers are members of the UA Little Rock Student Chapter of IQ International.

The only downside to the conference was that it did cause a delay to our IQ Journal publication schedule. Hopefully, we will be back on schedule for 2018.

Sincerely,

John Talburt, IQCP
IQ International Journal Editor-in-chief

Mapping the ISO Dimensions of Data Quality to the Conformed Dimensions of Data Quality (CDDQ)

(Based on CDDQ release 3.3.)

by
Dan Myers

Introduction

After attending a number of data management conferences where I heard a number of professionals use terms such as Integrity, Accuracy and Currency/Timeliness to mean different things, I began to look for a detailed and agreed upon standard set of definitions for the dimensions of data quality. What I found was a few great academic studies on the topic that have proposed a list of dimensions[1], a series of author's lists[2], and a few sets of dimensions documented by professional organizations[3]. What I discovered was that no two sets of dimensions were in agreement on what should be included within a set of dimensions and most lacked complete and verbose descriptions.

In June of 2013, I wrote a series of six articles reviewing four authors and two organization's list of the dimensions of data quality. This effort did reveal similarity in some areas, Completeness and Validity and the need for a conformed version that brings together the best of each, called the Conformed Dimensions of Data Quality (CDDQ). In 2016, I enhanced my conclusions made during the prior work and published them at http://dimensionsofdataquality.com/ for others to review, leverage during future research/study, and most importantly use on a day-to-day basis to communicate data quality issues.

Since the release in 2016, I have presented on this topic at conferences in the USA, Japan and individual organizations around the globe. I also conduct an annual survey[4] on the topic of the dimensions of data quality in order to foster further development and understanding. I owe a debt of gratitude to Danette McGilvray, Laura Sebastian-Coleman and Tom Redman for their input on the definitions and explanation of their own works on this topic over the years. My expectation is that this article will stir you to: discuss this topic among your peers, consider writing on this and related topics for the IQ International Journal, and contribute to the Conformed Dimensions going forward.

Initially, before considering publication of this work, the primary purpose for conducting this comparison of the CDDQ to the ISO/IEC 25012:2008 Data Quality Model's Dimensions of Data Quality was to identify how comprehensive the CDDQ is if used as an organizational standard and even an industry standard. After my review, it is clear to me that the CDDQ is robust and should be considered a candidate for use as an organizational standard and consideration for the basis for an industry standard.

Specifically, the purpose of this document is to:

- **Objective 1:** Identify how comprehensive the CDDQ is relative to the ISO/IEC 25012:2008.
- **Objective 2:** Map the two sets of dimensions in order to understand their respective strengths and identify challenges if moving from the use of the ISO standard to the CDDQ.
  a. Identify whether adherence to the CDDQ would satisfy ISO compliance requirements (section 2 of the ISO 25012:2008 allows for an alternative categorization of data quality characteristics, given that a mapping is provided).

  b. Establish a baseline gap analysis of additional measures of quality that could be used by an organization to shore-up their use of the ISO standard, or prepare for the use of additional facets if beginning to use the CDDQ within their organization.

The following are a few general observations:

	ISO/IEC 25012:2008	Conformed Dimensions of Data Quality (CDDQ)
Scope	The ISO standard is a subset of a larger framework, called the Software product Quality Requirements and Evaluation (SQuaRE). As such, it includes pieces that are system focused and intertwined with system management considerations (e.g. Portability, Recovery and to some extent Confidentiality). These aspects may be valuable if one's approach is System focused or requires continuity with other ISO standards.	The scope of the CDDQ is "Independent of System," excluding IT system management considerations. The CDDQ only focuses on data, not operating system or storage layers.
Context	Each of the ISO definitions includes a suffix phrase, "in a specific context of use" apparently limiting the application of each dimension. The ISO standard has defined internally used terms which adds clarity and facility to implementation, which the CDDQ should consider doing.	The CDDQ does not assume to have context, and relies upon the Information Quality professional and organization adopting the standard to identify whether additional context is required. The goal is to use the CDDQ as a cross-industry standard requiring that it remain abstract so that it can be applied universally across systems, companies, and industries. This also enables organizations to extend the framework as needed feeding recommendations back into the CDDQ, akin to Open Source software projects.
Granularity	The ISO standard lists dimensions (named characteristics) in a flat format- without subcategories for each dimension.*	The CDDQ is structured in a hierarchal form allowing for more granular classifications as required (see note below). Four of the CDDQ dimensions (Validity, Lineage, Timeliness and Integrity) can partially be found embedded within the ISO characteristics, but the level of detail and segregation is much more concise within the CDDQ.

* In their 2016 book, Data and Information Quality: Dimensions, Principles and Techniques, Carlo Batini and Monica Scannapieco identify this and three other weaknesses of the ISO standard. Page 19 (43).

Table 1: Observations on ISO Standards.

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. The following table lists these principles.[5]

#	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.

The Conformed Dimensions of Data Quality list a set of high level Dimensions (which correspond to the ISO standard's Characteristics) and Underlying Concepts (sub categories of each Dimension). Both, the ISO and CDDQ provide example metrics, but those were not directly mapped as part of this review.

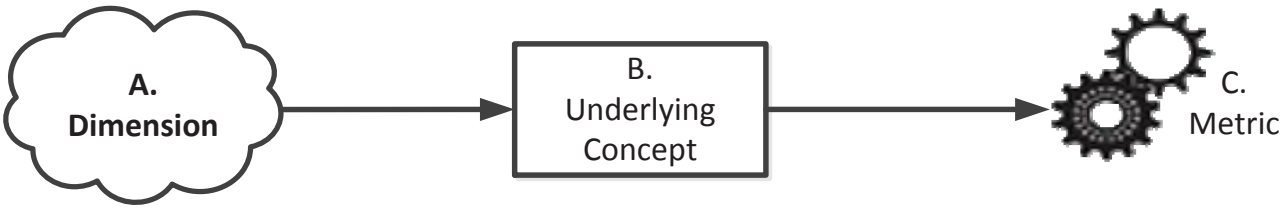


Figure 1: Conformed Dimensions of Data Quality.

Mapping the CDDQ to the ISO Standard

#	Characteristic	Characteristic Definition	Example	Example Data Quality Meausre Name	Conformed Dimension of Data Quality	Dimension Definition	CDDQ Target ID	Concept #	Underlying Concept	Concept Definition
1.	Accuracy.	The degree to which data has attributes that correctly represent the true value of the intended attributes of a concept or event in a specific context of use.	Syntactic Accuracy: A low degree of syntactic accuracy is when the word Mary is stored as Marj.	Record's field syntactic accuracy.	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).	1.1	1	Agree with Realworld.	Degree that data factually represents its associated real-world object, event, or concept.
							1.2	2	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.
			Semantic Accuracy: A low degree of semantic accuracy is when the name John is stored as George. Both names are syntactically accurate, because of the domain of reference in which they reside, but George is a different name related to another person.		Validity.	Validity measures whether a value conforms to a preset standard.	4.3	3	Domain of Predefined Values.	This is a set of permitted values.
2.	Complete-ness.	The degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use.	For an employee data base, completeness lessens, if some employees' records do not contain the data regarding a number where they can be reached in the event of an emergency.	Complete-ness of data within a file.	Complete-ness.	Completeness measures the degree of population of data values in a data set.	2.2	4	Attribute Population.	This measures whether a value is present (not null) for an attribute (column).
							2.1	5	Record Population.	This measures whether a row is present in a data set (table).

Mapping the CDDQ to the ISO Standard

#	Characteristic	Characteristic Definition	Example	Example Data Quality Meausre Name	Conformed Dimension of Data Quality	Dimension Definition	CDDQ Target ID	Concept #	Underlying Concept	Concept Definition
3.	Consistency.	The degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use. It can be either or both among data regarding one entity and across similar data for comparable entities.	An employee's birth date cannot be later than his "recruitment date".	Name Consis-tency of a data file.	Consis-tency.	Consistency measures whether or not data is equivalent across systems or location of storage.	3.1	6	Equiva-lence of Redundant or Distributed Data.	The measure of similarity with other sources of data that represent the same concept.
							3.3	7	Logical Consisten-cy.	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.
4.	Credibility.	The degree to which data has attributes that are regarded as true and believable by users in a specific context of use.	Data certified from an independent and trusted organization should be considered credible.	Credi-bility of data used by a bank for evalua -ting credit risk.	No Direct Mapping, See Explana-tion for Indirect Mapping.			8		
5.	Currentness.	The degree to which data has attributes that are of the right age in a specific context of use.	The timetable of a railway station must be updated with the frequency required to allow passengers to take a train even if the scheduled time or platform change.	Current-ness of a field data value.	Currency.	Currency measures how quickly data reflects the real world concept that it represents.	11.1	9	Current with World it Models.	Data is current if it reflects the present state of the concept it models.
6.	Accessibility.	The degree to which data can be accessed in a specific context of use, particularly by people who need supporting technology or special configuration because of some disability.	Data that should be managed by a screen reader cannot be stored as an image.	Sound data accessi-bility.	Accessi-bility.	Accessibility measures how easy it is to acquire data when needed, how long it is retained, how access is controlled, and whether facts exist as data.	7.1	10	Ease of Obtaining Data.	This measures how easy it is to obtain data.

Mapping the CDDQ to the ISO Standard

#	Characteristic	Characteristic Definition	Example	Example Data Quality Meausre Name	Conformed Dimension of Data Quality	Dimension Definition	CDDQ Target ID	Concept #	Underlying Concept	Concept Definition
7.	Compliance.	The degree to which data has attributes that adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use.	Credit risk data of a bank must comply with specific laws and standards.	1. Privacy law non-conformity: values. 2. Privacy law non-conformity: architecture.	No Direct Mapping, See Explanation for Indirect Mapping.	Validity measures whether a value conforms to a preset standard.	4.2	11		
8.	Confidentiality.	The degree to which data has attributes that ensure that it is only accessible and interpretable by authorized users in a specific context of use.	Data that refers to personal or confidential information like health or profit must be accessed only by authorized users or should be written in secret code.	Encryption usage.	Accessibility.	Accessibility measures how easy it is to acquire data when needed, how long it is retained, how access is controlled, and whether facts exist as data.	7.2	12	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.
				Non vulnerability.						
9.	Efficiency.	The degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use.	Using more space than necessary to store data can cause waste of storage, memory and time.	Numbers stored as strings.	No Direct Mapping, See Explanation for Indirect Mapping.	N/A	4.4	13	Validity: 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.
				Storage wasted space.			4.5	14		
10.	Precision.	The degree to which data has attributes that are exact or that provide discrimination in a specific context of use.	A precision of 5 decimal places allows different functionalities rather than a precision of 2 decimal places.	Precision of data values and Precision of fields of a database.	Precision.	Precision measures the number of decimal places and rounding of a data value or level of aggregation.	8.1	15	Precision of Data Value.	The measure of preciseness of numeric data using decimal places, rounding and truncation.

Mapping the CDDQ to the ISO Standard

#	Characteristic	Characteristic Definition	Example	Example Data Quality Meausre Name	Conformed Dimension of Data Quality	Dimension Definition	CDDQ Target ID	Concept #	Underlying Concept	Concept Definition
11.	Traceability.	The degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use.	Public administrations must keep information about the access executed by users for investigating who read/ wrote confidential data.	Trace-ability of values.	Accessi-bility.	Accessibility measures how easy it is to acquire data when needed, how long it is retained, how access is controlled, and whether facts exist as data.	7.2	16	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.
				Automa-tic trace-ability.	Lineage.	Lineage measures whether factual documen-tation exists about where data came from, how it was transformed, where it went and end-to-end graphical illustration.	9.2	17	Segment Documen-tation.	Segment documentation provides how data is transformed and transported from one location to another.
12.	Understand-ability.	The degree to which data has attributes that enable it to be read and inter-preted by users, and are expressed in appropriate languages, symbols and units in a specific context of use.		Master data un-derstand-ability due to existing metada-ta.	Repre-sentation.	Representation measures ease of understanding data, consistency of presentation, appropriate media choice, and avail-ability of documentation (metadata).	10.1	18	Easy to Read & Interpret.	Illustrations and charts should be self explanatory and presented with appropriate labels, providing context.
				Master data un-derstand-ability due to linked metada-ta.			10.2	19	Presen-tation Language.	Data that is represented well is simple but elegantly formed with good grammar and presented in a standard way.
							10.3	20	Includes Measure-ment Units.	Well represented data includes the scale of measurement, such as weight, height, distance… etc.
							10.4	21	Metadata Availability.	Comprehensive descriptions and other information about the characteristics of the data are provided in plain language.

#	Characteristic	Characteristic Definition	Example	Example Data Quality Meausre Name	Conformed Dimension of Data Quality	Dimension Definition	CDDQ Target ID	Concept #	Underlying Concept	Concept Definition
13.	Availability.	The degree to which data has attributes that enable it to be retrieved by authorized users and/or applications in a specific context of use.	Data should be available also during managing operations like backup.	Data items availabi-lity.	Accessi-bility.	Accessibility measures how easy it is to acquire data when needed, how long it is retained, how access is controlled, and whether facts exist as data.	7.3	22	Retention.	Retention refers to the period of time that data is kept before being removed from a database through purge or archive processing.
14.	Portability.	The degree to which data has attributes that enable it to be installed, replaced or moved from one sys-tem to another preserving the existing quality in a specific context of use.		Data porta-bility.	No Direct Mapping, See Explana-tion for Indirect Mapping.	N/A		23		
15.	Recoverability.	The degree to which data has attributes that enable it to maintain and preserve a specified level of operations and quality, even in the event of fail-ure, in a specific context of use.	When a media device has a failure, data stored in that device should be recoverable.	Recover-ability.	No Direct Mapping, See Explana-tion for Indirect Mapping.	N/A		24		

ISO: Accuracy and CDDQ: Accuracy

Unlike most authors/organizations[6], the ISO standard doesn't refer to Accuracy as relating to the correctness of data compared to the real-world. The, "true value of the intended attributes" is the language used. Within the two named subtypes of Accuracy, both are very closely aligned with CDDQ Validity concepts, where the ISO "Syntactic Accuracy" is most closely aligned with the CDDQ Validity dimension's underlying concepts of "Values Conform to Format" and "Semantic Accuracy" aligns with the concept called "Domain of Predefined Values".

Note that the ISO standard doesn't name a separate dimension for "Validity", likely because those concepts are located here within what they identify as "Accuracy". The author is only aware of two other sources (either author or organizations) that place Validity within the Accuracy[7] dimension.

The Conformed Dimensions proposed standard is inclusive of all of the points covered by the ISO standard, and goes further, adding the two following underlying concepts based on strong industry consensus[8] that these form the bedrock of the Accuracy dimension.

- **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.

ISO: Completeness and CDDQ: Completeness

Both of the standards are very similar regarding Completeness, where both call out completeness of attributes and records/rows. The CDDQ does expand on the subject, offering explanation of three additional concepts that are sometimes overlooked. These are:

- **Truncation** - This measures whether the value contains all characters expected.
- **Existence** - Existence identifies whether a real - life fact has been captured as data.

Some authors place "Truncation" within the Precision dimension given it's similarity to rounding of numeric values, but technically this concept can apply to many other data types. For example, if a file transmission is ended incorrectly it can result in truncation, which end-users refer to as file incompleteness. Sometimes typographical errors, in written documents, such as words, phrases and sentences are truncated. The most difficult versions of this to detect are ***compound*** words such as *lifestyle*, *website*, and *schoolboy*, because their truncated versions are still valid words that don't register as misspellings with spell check software.

Lastly, most people take for granted that the existence of data is an attribute of quality. Later, under the discussion of Accessibility we differentiate between the act of getting the data (accessing it) versus knowing it has been recorded, or exists somewhere. The CDDQ classifies Existence (the act of having recorded an observation as data) under Completeness.

ISO: Consistency and CDDQ: Consistency

The underlying language of the ISO definition of consistency, "free from contradiction and are coherent with other data" fundamentally aligns with the CDDQ definition. Although the ISO standard doesn't name any subtypes, it also notes that, "It can be either or both among data regarding one entity and across similar data for comparable entities." Which aligns with the CDDQ underlying concept called, "Equivalence of Redundant or Distributed Data," defined as, "The measure of similarity with other sources of data that represent the same concept."

It isn't clear whether the ISO definition of consistency means to make a distinction between the equivalence of data for the same characteristics but represented differently (e.g. "Male" and "1", where 1 means male based on a lookup table) or those logically related (e.g. "birthdate" and "hire date"). Based on the ISO example provided, ("birthdate" and "hire date") it appears that they agree that there are two underlying concepts here. The CDDQ explicitly differentiates these as "Equivalence of Redundant or Distributed Data" referring to the former example and "Logical Consistency" as the latter.

- **Equivalence of Redundant or Distributed Data** - The measure of similarity with other sources of data that represent the same concept.

- **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.

Lastly, the CDDQ also identifies one last underlying concept within Consistency, called "Format Consistency". This often is found when attributes are stored inconsistently across tables, referred to as, "Consistent Representation- the extent to which data is presented in the same format".[9]

- **Format Consistency** - This measures the conformity of format of the same data in different places.

ISO: Credibility and CDDQ

According to the ISO definition of this dimension, the data should be "true" and "believable". This aligns with the CDDQ Accuracy dimension to the extent that "true" and "factual" are synonyms. Given that philosophically and in some religious contexts, *Truth* implies much more than *Fact*. The use of this word is discouraged. As a general tenet of the CDDQ, subjective measures of data quality should not be part of the Conformed Dimension standard. Having said that, we understand that there will likely be organizational variations and customizations for any implementation of the CDDQ. As such, these respective organizations are recommended to use *Believability* within the context of observable measures of quality.

For example, after discussions with data consumers, it is found that they frequently require a *Believability* check, then it is worth measuring what that means. This can be done by narrowing down which CDDQ dimensions and sub concepts most closely align with how they define *Believability* (most often this is done through example and case study). If it is said that the data should be all there, look right, and be delivered quickly, then the organizationally defined definition (outside of the CDDQ) may be called *Believability* with the following CDDQ dimensions and sub concepts as a foundation. This allows data management professionals to develop dashboards of DQ based on automated metrics (reused from the CDDQ) that roll up to broader classifications of that the individual organization defines having "Believability".

D. Myers and B. Blake conducted a detailed comparison of Information Quality research definitions of Believability in their paper presented at the 2017 MIT International Conference on Information Quality.[5]

Mapping Everyday Common Language with the Conformed Dimensions

Organizationally Defined Characteristic of Believability	Conformed Dimensions Name	Associated Conformed Dimensions Sub Concept(s)
"Be all there"	Completeness	1. Record Population 2. Attribute Population 3. Truncation
"Look right"	Validity	1. Values in Specified Range 2. Values Conform to Business Rule 3. Domain of Predefined Values 4. Values Conform to Data Type 5. Values Conform to Format
"Delivered quickly"	Timeliness	1. Time Expectation for Availability 2. Concurrence of Distributed Data
	Accessibility	1. Ease of Obtaining Data 2. Access Control

Note: there are more sub concepts associated with each of the Conformed Dimensions standard (e.g. Existence, and Retention). These are only the ones most closely aligned with these example phrases which are often repeated by data consumers.

ISO: Currentness and CDDQ: Currency

The key phrase of the ISO definition of Currentness is "right age" which closely aligns with the CDDQ *Currency* dimension. The CDDQ definition, however, focuses on the aspect of modeling the real-world and the age associated with whether the data represents the real-world concept, rather than the "right" age which implies more of a customer centric, almost "fitness for use" based view.

ISO: Accessibility and CDDQ: Accessibility

Generally speaking the CDDQ "Accessibility" dimension and underlying concept of "Ease of Obtaining Data" maps pretty well. The additional language, saying, "need supporting technology" is a system related constraint and therefore is outside the scope of the Conformed Dimensions.

In terms of Accessibility, the CDDQ goes further, covering additional concepts (below). *Access Control* defined in the CDDQ corresponds most closely to the ISO Dimension called "Traceability" as discussed later.

- **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.
- **Fact Captured as Data** - Before one can access data, real-world observations must be recorded as data facts.

ISO: Compliance and CDDQ: Validity

At a high level, the phrase, "adhere to standards, conventions or regulations" within the ISO definition of "Compliance" maps to the CDDQ *Validity* dimension which also addresses "standards", saying, "Validity measures whether a value conforms to a preset standard." Going beyond this however, the CDDQ underlying concepts would not cover cases of illegality (e.g. collection of race identifiers by automobile insurers for the purpose of pricing/rate making), which seems to be included within the ISO dimension.

At this point it is worth noting that there is no single dimension of the ISO standard that maps directly to the CDDQ's *Validity* dimension. Perhaps future updates to the ISO standard will include such a dimension. The CDDQ's coverage of the Underlying Concepts of data quality within the *Validity* dimension include the following:

- **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.

ISO: Confidentiality and CDDQ: Accessibility

The key phrase of the definition of "Confidentiality" according to the ISO standard, is that the data, "is only accessible and interpretable by authorized users" which directly maps to the CDDQ Accessibility dimension and underlying concept of Access Control (see below). Accessibility is also one of the ISO and CDDQ dimensions discussed later so we'll address further similarity in that section.

- **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.

ISO: Efficiency and CDDQ

The ISO categorizes characteristics by whether they can be evaluated as "inherent" and/or "system dependent" points of view. These aren't defined and there isn't any explanation about how to use these two views should be used. The inclusion of the Efficiency dimension appears to come more from the "system dependent" category given the nature of what it measures. Of the seven or more authors/organizations reviewed[10], there are none that identify an Efficiency dimension as espoused here by the ISO.

Principle number three of the CDDQ states that, "*Independent of System* - The Conformed Dimensions do not contain any storage or system specific constraints." As such, facets like data storage "Efficiency" are purposefully excluded. Having said that, the first example of this dimension, provided by the ISO, says that "Numbers stored as strings" exemplifies an inherent aspect covered by this dimension. The CDDQ would cover such cases under the definition of the *Validity* dimension's sub concepts of *Values Conform to Data Type* or even *Values Conform to Format*.

ISO: Precision and CDDQ: Precision

Of all of the dimensions identified under the ISO and CDDQ, the "Precision" dimension is the most closely aligned. The ISO identifies "attributes that are exact or that provide discrimination" as the fundamental of this dimension and both of their examples are covered by the CDDQ sub concept of *Precision of Data Value* which is coincidentally the same name as the first example provided by the ISO.

The CDDQ also identifies a very important sub concept, called Granularity which is referred to by some as level of aggregation, grain or coverage.

- **Granularity** - The detail or summary of data defines the granularity measured by the number of attributes used to represent a single concept.

ISO: Traceability and CDDQ: Lineage and Accessibility

The ISO dimension of "Traceability" has two primary underlying facets, that an "audit trail of access to the data" is available, and an audit trail for "changes made to the data". The former, regarding whom accesses data, falls under the CDDQ *Accessibility* dimension, and sub concept of *Access Control*. The later facet falls under the CDDQ dimension called *Lineage* and sub concept of *Segment Documentation*. From the beginning of the CDDQ standard, the *Lineage* concept was identified in Loshin (2011)[11]. Some organizations place this as a subset of documentation or metadata, so it's understandable that this doesn't show up as a stand-alone dimension within the ISO standard. The Underlying Concepts of *Lineage* in the CDDQ include:

- **Source Documentation** - Source documentation provides data provenance which describes the origin of the data.
- **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.

ISO: Understandability and CDDQ: Representation

There are three phrases within the ISO definition of "Understandability" that map directly to the CDDQ Representation dimension's underlying concepts.

- The ISO phrase, "enable it to be read and interpreted by users" most closely corresponds to the CDDQ sub concept of *Easy to Read & Interpret*.
- The ISO phrase, "expressed in appropriate languages, symbols" very closely matches the CDDQ sub concept of *Presentation Language*.
- The ISO identification of the importance of "units" is also identified in the CDDQ sub concept of *Includes Measurement Units*.

The ISO also provides two examples highlighting the importance of Metadata used to fully understand Master Data. The CDDQ standard places *Metadata Availability* as an Underlying Concept within the Representation dimension. Each of the concepts within the *Representation* dimension of the CDDQ are as follows:

- **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.
- **Includes Measurement Units** - Well represented data includes the scale of measurement, such as weight, height, distance…etc.

ISO: Availability and CDDQ: Accessibility

The CDDQ identified "Availability" as synonymous with "Accessibility" and chose to label it as a secondary, non-standard name for the dimension. The ISO standard includes both with Accessibility applying to both the *Inherent* and *System Dependent* categories and Availability to only the *System Dependent* category. Fundamentally, the distinction isn't clear based on just reviewing their descriptions.

Given that the CDDQ standard is *Independent of System* as outlined in its Principles, the system-specific constraints listed by the ISO aren't applicable. Having said that, the second note of further explanation by ISO on this topic, saying, "Note 2. Another case of availability is the capability of data to be available for a specific period of time." Aligns well with the CDDQ *Accessibility* dimension's underlying concept of *Retention*, defined as, "Retention refers to the period of time that data is kept before being removed from a data store through purge or archive processing."

ISO: Portability and Recoverability and CDDQ

As stated in the last paragraph, the CDDQ standard is *Independent of System* and the last two dimensions cited in the ISO standard are functions of a system. Logically, data itself is inanimate and has no function to install, replace or move itself (ISO, "Portability"), nor maintain quality, even in system failure (ISO, "Recoverability").

Conclusion

The following conclusions can be made regarding the two objectives identified at the beginning of this paper.

Objective 1:
Identify how comprehensive the CDDQ is relative to the ISO/IEC 25012:2008.

Findings 1:
The Venn diagram and table below illustrate the level of comprehensiveness that the CDDQ standard offers relative to the ISO standard. At the most aggregate level the ISO standard includes 15 dimensions versus the CDDQ which includes only 11 dimensions. So naturally one would think that the ISO standard is more granular, but just the opposite was found to be the case. The ISO includes 25 concrete concepts within the 15 dimensions (though not always identified by the ISO as such). Because the CDDQ is constructed as a hierarchal framework with 11 parent Dimensions and 33 Underlying Concepts it actually provides more specificity and granularity than the ISO standard.

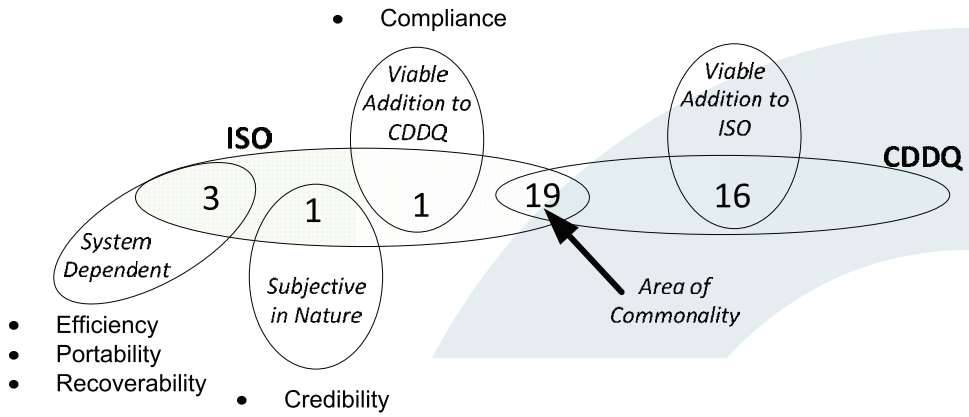
Categorizing the findings:

- Because the CDDQ has a different scope (excluding systems related dimensions), three dimensions should be set aside if trying to directly compare the ISO and CDDQ. These three are: Efficiency, Portability and Recoverability).

- Because the CDDQ doesn't include any subjective measures the Credibility dimension needs to be excluded or developed based on organizational needs as a composite of CDDQ dimensions.

- The area of agreed upon scope includes 20 individual Underlying Concepts, of which the CDDQ includes 19 and is missing one (Compliance).
- The CDDQ additionally, has one partial dimension (Validity) not directly found in ISO including 3 Underlying Concepts (Values in Specified Range, Values Conform to Business Rule, Values Conform to Format). And 13 additional Underlying Concepts (see full list on the following page).

Venn Diagram



Tabular Comparison

	ISO Underlying Concepts	CDDQ Underlying Concepts	CDDQ Count as Percentage of ISO Count
Raw Comparison	24	33	138%
Within Similar Scope	20	19	95%
CDDQ Additional		16	

Summary of Findings (see table above)

- From a raw comparison of the CDDQ to the ISO, the CDDQ is offers more granularity and descriptive detail than the ISO standard.
- When comparing apples to apples, based on similar scope, the ISO includes one Dimension-Underlying Concept (Compliance- Privacy Laws Conformity) greater granularity than the CDDQ. Conversely, the CDDQ offers one Dimension and 16 Underlying Concepts (across various Dimensions) of greater granularity than the ISO.

Objective 2:

Map the two sets of dimensions in order to understand their respective strengths and identify challenges if moving from the use of the ISO standard to the CDDQ.

a. Identify whether adherence to the CDDQ would satisfy ISO compliance requirements (section 2 of the ISO 25012:2008 allows for an alternative categorization of data quality characteristics, given that a mapping is provided).

b. Establish a baseline gap analysis of additional measures of quality that could be used by an organization that is transitioning from the ISO standard to the CDDQ standard.

Findings 2:

a. Based on the findings of objective 1, it can be seen that three additional system-specific dimensions of data quality are required (Efficiency, Portability and Recoverability) and perhaps one subjective Dimension (Credibility) will be required if using the CDDQ and seeking to conform to the ISO standard.

b. Notwithstanding these areas, the CDDQ offers twelve (16) additional Sub Concepts not covered by the ISO standard (16/24=66% more than the ISO standard). ISO compliant organizations are recommended to review these and identify whether the addition of these Sub Concepts is of value going forward. (See section "c" below for the complete list and descriptions)

c. The following is a list of the sub concepts identified by the CDDQ, missing from the ISO standard.

- **Accuracy**
  1. **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.
- **Completeness**
  2. **Truncation** - This measures whether the value contains all characters expected.
  3. **Existence** - Existence identifies whether a real-life fact has been captured as data.
- **Consistency**
  4. **Format Consistency** - This measures the conformity of format of the same data in different places.
- **Validity**
  5. **Values in Specified Range** - Values must be between some lower number and some higher number.
  6. **Values Conform to Business Rule** - Validity measures whether values adhere to some declarative formula.
  7. **Values Conform to Format** - Validity measures whether the data are arranged or composed in a predefined way.
- **Precision**
  8. **Granularity** - The detail or summary of data defines the granularity measured by the number of attributes used to represent a single concept.
- **Lineage**
  9. **Source Documentation** - Source documentation provides data provenance which describes the origin of the data.
  10. **Target Documentation** - Documentation about the target explains where the data moved to and how it is stored.
  11. **End-to-End Graphical Documentation** - End-to-End documentation provides diagrammatic visual representation of how the data flows from beginning to end.

- **Representation**
  12. **Media Appropriate** - The appropriate media (e.g. Web-based, hardcopy, or audio…etc.) are provided.
- **Timeliness**
  13. **Time Expectation for Availability** - The measure of time between when data is expected versus made available.
- **Integrity**
  14. **Referential Integrity** - Integrity measures whether if when a value (foreign key) is used it must reference an existing key (primary key) in the parent table.
  15. **Unique Identifier of Entity** - Integrity measures whether a data set has a unique key for each fact it represents.
  16. **Cardinality** - Cardinality describes the relationship between one table to another, such as one-to-one, one-to-many, or many-to-many.

- The following is a list of the Dimension and Underlying Concepts identified by the ISO, missing from the CDDQ. The inclusion of this Dimension and Underlying Concepts should be considered in future versions of the CDDQ.
  - **Compliance** - The degree to which data has attributes that adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use.
    1. **Privacy law non-conformity (Values)** - number of items that do not conform to privacy law statements due to data content.
    2. **Privacy law non-conformity (Architecture)** - number of items that do not conform to privacy law statements due to technical architecture failures.

.

References

1. 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.
   D. M. Strong, Y. W. Lee, and R. Y. Wang, *"Data Quality in Context,"* Commun. ACM, vol. 40, pp. 103-110, May 1997.
   C. Carlo Batini, Cinzia Cappiello, Chiara Francalanci, Andrea Maurino, *Methodologies for data quality assessment and improvement*, ACM Computing Surveys, Vol. 41, No. 3, Article 16, Publication date: July 2009.
   For a more complete list see http://dimensionsofdataquality.com/research

2. L. English, *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*, Wiley, 1999.
   L. English, *Information Quality Applied*, Wiley Publishing, 2009.
   T. Redman, *Information Age*, ARTECH HOUSE, 1997.
   T. Redman, *The Field Guide*, Digital Press, 2001.
   D. McGilvray, *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*, Morgan Kaufmann, 2008.
   D. Loshin, *The Practitioner's Guide to Data Quality Improvement*, Morgan Kaufmann, 2009.

3. Data Administration and Management Association (DAMA) DMBOK, Technics Publications LLC, 2009.
   Unknown, TDWI- Data Quality Fundamentals, The Data Warehousing Institute (TDWI), 2011.

4. The results of the past two year's surveys are available upon request at:
   http://dimensionsofdataquality.com/dims_survey

5. D. Myers, B. Blake, *"An Evaluation of the Conformed Dimensions of Data Quality in Application to an Existing Information Quality-Privacy-Trust Research Framework"* MIT International Conference on Information Quality, UA Little Rock, October 6-7, 2017.

6. Examples include: Redman (2001), English (2009), TDWI (2011), and the DAMA DMBOK (2009).

7. Loshin (2011) calls it "Domain Definition". See p.136[158] and Batini & Scannapieco (2016) call it "Syntactic Accuracy". See p. 24[47]

8. Redman (2001), English (2009), DAMA DMBOK (2009), Lee/Pipino/Funk/Wang (2006), McGilvray (2008).

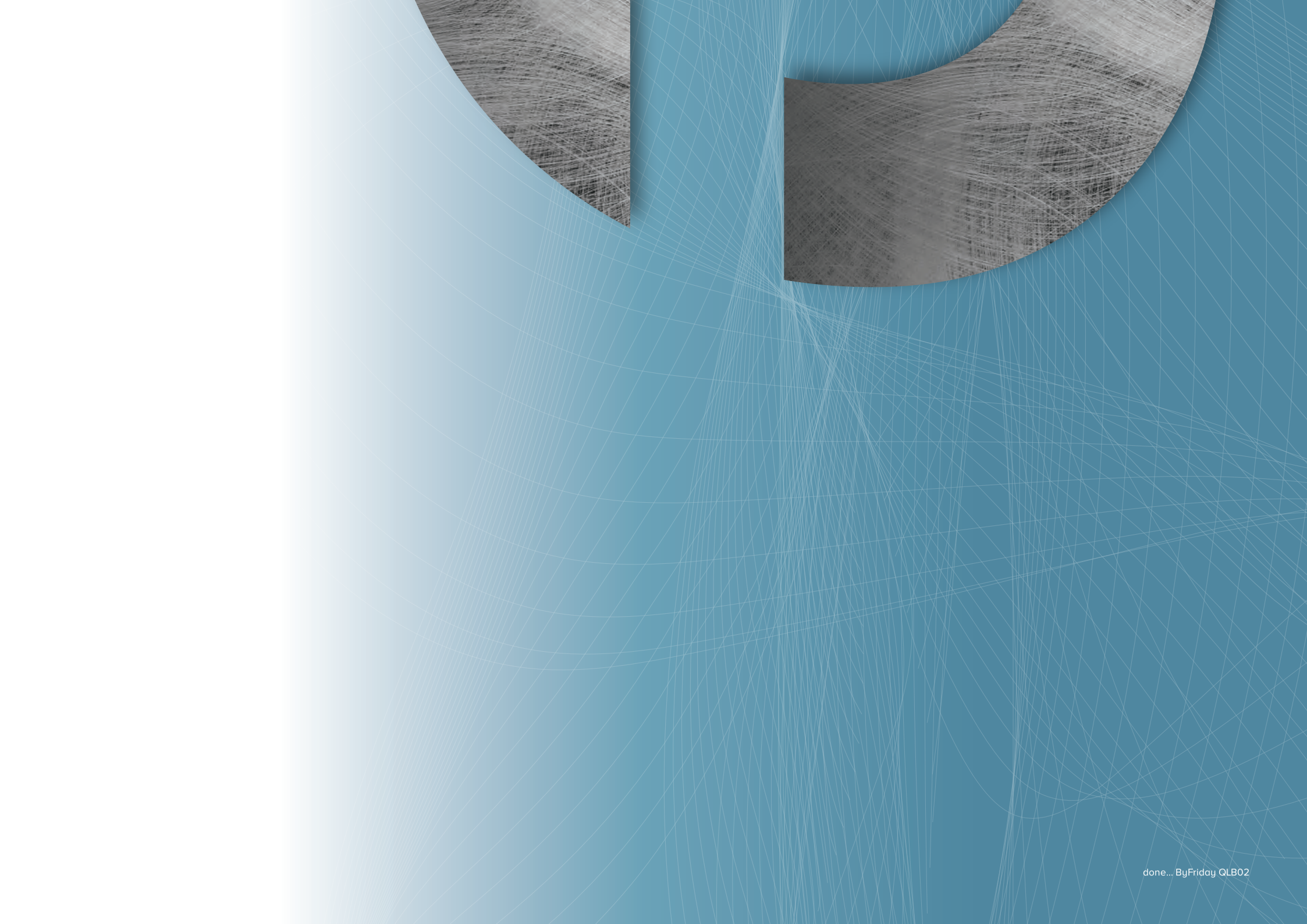
9. L. L. Pipino, Y. W. Lee, and R. Y. Wang, *"Data Quality Assessment,"* Commun. ACM, vol. 45, pp. 211-218, Apr. 2002.

10. The author's review of dimensions to date has included the following: Redman (2001), English (2009), TDWI (2011), and the DAMA DMBOK (2009), Loshin (2001 & 2011), Lee/Pipino/Funk/Wang (2006), McGilvray (2008).

11. D. Loshin, *The Practitioner's Guide to Data Quality Improvement*, Elsevier 2011. Page 136.

About the Authors

Dan Myers (MBA, IQCP) is the Principal Educator at Data Quality Matters (DQMatters.com), a Silicone Valley based eLearning company focused on providing data quality and data management learning material. As a thought leader, Dan conducted a robust comparison of key IQ authors' lists of dimensions of Data Quality and proposed a way to align the Information Quality community by using common definitions. This led to his 2016 proposal to form a standard set of dimensions, called the Conformed Dimensions of Data Quality (CDDQ), and published it at (http://DimensionsOfDataQuality.com). Dan publishes an annual report about the Dimensions of Data Quality (http://dqm.mx/cddq-survey) and a blog on the topic via the CDDQ website. As a practitioner, Dan has worked as an applications developer, data modeler, and manager of Data Governance/Data Quality. Dan's fluency in Japanese enabled him to work in both the public and private sector in Japan, and he speaks abroad on various data management topics.


done... ByFriday QLB02