(http://www.information-management.com)

By Dan Myers
JUN 20, 2013 3:16am ET

FEATURE

Dimensions of Data Quality Under the Microscope

By Dan Myers
JUN 20, 2013 3:16am ET

As information management and data quality roles and responsibilities become more mainstream in large organizations there has been a call to agree upon standard categories of data quality. Malcolm Chisholm's recent Information Management article suggests that there is no consensus regarding the dimensions of data quality. This complaint is not new (http://web.mit.edu/tdqm/www/tdqmpub/WandWangCACMNov96.pdf). Yair Wand and Richard Y. Wang further argue that the expected value of dimensions of quality hasn't been seen and even create a distraction.

Whether you call these categories "dimensions" [of data quality] or something else is a discussion for another time. I think Malcolm Chisholm's proposal to call these "properties" makes a lot of sense, and I appreciate his ability to cut to the chase.

Having said that, I believe we'd be throwing out the baby with the bathwater if we dismissed the writing of multiple authors on the dimensions of data quality just because there isn't a current consensus. Now is the right time for the data quality industry to finalize a set of standards, much like the accounting field has done with the Generally Accepted Accounting Principles. Every organization needs to have a defined set of measures of quality, which should be composed of industry standard dimensions. Each organization should then identify its unique needs for measurement. In this series of articles, I will document the level of consistency between six authors' definitions of each of the dimensions of quality. The first of these, "accuracy," is covered in this article.

In the capstone article, I will propose a conformed set of dimensions that incorporates the six authors' definitions and my own experience.

It is my expectation that this article will speed up the industry's rationalization of the dimensions of quality. This series of articles will answer two of the three challenges identified by Malcolm Chisholm.

1. Define the concepts that compose the dimensions of quality and propose an alignment of the major contributor's works, which is the first step to defining the dimensions themselves.
2. Compile a thorough list of the underlying concepts of the dimensions of data quality, with the expectation that this work will cover the majority of all concepts.

When discussing the level of agreement on the dimensions of quality, consensus of definition should be measured within its intended scope. Dimensions of quality are most often implemented as a part of a broader data quality/governance effort and, as such, are determined and maintained within a given unit of authority, like the data governance board of an organization. There is authority given to them by the leadership of that organization and consensus is only required within that group (or within the data management roles across the company). This limits the scope of consensus building, making it feasible, compared to requiring consensus among all employees, companies, industries, etc. In this context, the dimensions may be considered principles to organize and direct change, rather than fixed laws, which would require stronger controls and global consensus.

The first challenge is to collect each author's definition, and I have done so for six mainstream authors. I realize that every single contributor or author can't be reviewed for this article, but, as Danette McGilvray pointed out, some authors (including herself) established dimensions of data quality not for the purpose of identifying the root concept and associated dimension, but rather established dimensions by type of method/technique of remediation. In the process, it is also helpful to reference these dimensions in context of the groupings they were explained by each respective author. Here are a few schemes for grouping all of the dimensions by author. Unfortunately I don't have room within this article to compare each.

Tom Redman: Dimensions can be grouped by those having to do with a data model, data value or data presentation.[1]
Larry English: Dimensions can be grouped by information content or information presentation.[2]
David Loshin: Dimensions can be grouped as intrinsic, a measurement associated with data values themselves; contextual, in terms of relationship between records; qualitative, a synthesis of measures associated with intrinsic and contextual; or classifying.
Yang W. Lee, Leo L. Pipino, James D. Funk and Richard Y. Wang: Intrinsic IQ - accuracy, objectivity, believability and reputation; Accessibility IQ - accessibility and security; Contextual IQ - relevancy, value added, timeliness, completeness and amount of information[3;] Representational IQ - interpretability, ease of understanding, concise representation and consistent representation.[4]

Please note that there are other data quality subject matter experts that could be added to this list, including but not limited to: Arkady Maydanchik, Danette McGilvray, Jack Olson, Carlo Batini and Monica Scannapieco.

Table 1: Six Data Quality Authors' Definitions of the "Accuracy" Dimension

Author/ Source	Dimension	Definition
Tom Redman	Accurate	Almost always one of the most important dimensions of data quality. Defined as a measure of the degree of agreement between a data value or collection of data values and a source agreed to be correct. Informally, customers say they need "data to be accurate" and mean they require that data values agree with the real world values.
Larry English	Accuracy (reality, surrogate source)	Accuracy to surrogate source: The data agrees with an original, corroborative source record of data, such as a notarized birth certificate, document, or unaltered electronic data received from a party outside the control of the organization that is demonstrated to be a reliable source. Accuracy to reality: The data correctly reflects the Characteristics of a Real-World Object or Event being described. Accuracy and Precision represent the highest degree of inherent information quality possible.
TDWI	Accuracy	Data is accurate when it matches reality and is a representation of the truth.
DMBOK	Accuracy	Data accuracy refers to the degree that data correctly represents the real-life entities they model. In many cases, you can measure accuracy by how the values agree with an identified reference source of correct information, such as comparing values against a database of record or a similar corroborative set of data values from another table, checking against dynamically computed values, or perhaps applying a manual process to check value accuracy.
David Loshin	Accuracy	The characteristics of accuracy include the definition of a system of record (or a set of values of record), the precision of data values, value acceptance and domain definition.
Lee, Pipino, Funk, Wang	Free of Error	It is a common practice to use the term accuracy when referring to whether the data is correct. The dimension of accuracy itself, however, can consist of one or more variables, only one of which is whether the data are correct. We use "free of error" to label the dimension that represents whether the data is correct.

Rather than pick at semantic differences between each of the definitions listed in Table 1, let's look at the conceptual similarities, which have been underlined. In Table 2, the three primary concepts that encompass accuracy have been identified with key quotes extracted from each author's definition.

Table 2: Concepts Within Accuracy Dimension

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

*Grey cells represent a lack of coverage of a concept by author.

Concept similarity:

Five out of the six authors explicitly cite "agreement with the real world" as a component of accuracy.
Four of the six say that data should "Match To Agreed Source."

Two authors include precision (the exactness of data, like the number of digits a number must include or if rounding is allowed).

If our goal is to identify consensus and disaggregate the concepts that overload this dimension, we could separate out “precision of the data” as its own dimension (as we see in Table 3 that three authors have done).

Table 3: Accuracy and Precision Dimension Similarity by Author

	Tom Redman	Larry English	TDWI	DMBOK	David Loshin	Lee, Pipino, Funk, Wang
Accuracy (Factualness)	Accurate	Accuracy (reality, surrogate source)	Accuracy	Accuracy	Accuracy	Free of Error
Precision	Format Precision	Precision	Precision	Precision	(included in Loshin's definition of Accuracy)	

*Green cells represent strong similarity/agreement between authors for a given dimension. Yellow cells represent are as of weak agreement or difference in categorization. Grey cells represent a lack of coverage of a concept by author.

The goal of disaggregation is to make communication more precise and remove assumptions. So if a dimension isn't broadly known to include a particular concept, I suggest that, in theory, it is easier to remove this concept without changing the known meaning. As you can see above, if we break Precision out of Accuracy, then five out of six sources would agree with regard to a Precision dimension with the primary concept of “Precision of Data Value” (number of decimal places and rounding).

According to Tom Redman, the two apparent concepts unique to Accuracy are “Agree with Real World” and “Match to Agreed Source” because the former only works when there are physical objects/phenomena to observe, but in the case of events, an agreed upon source of record is usually needed. English puts it this way: “To measure Information Process Quality, you compare the sampled data to the Characteristics of the Real-World Object or Event that the data represents.”

Though a number of authorities cite correct sourcing as a component of data quality, not all cite it as a part of Accuracy, but rather as the primary concept within the Consistency dimension. One large insurance company effectively identified “sourcing” as a standalone dimension of data quality, which may work for your organization as well.

Table 4: Concepts Within Consistency Dimension

Author/ Source	Free of Conflict with Other Sources	Consistency in Representation
Tom Redman		
Larry English		
TDWI	"Free of conflict" [with other sources]	
DMBOK	"One set consistent with another" [at equivalent level, row, column, time, etc.]	
David Loshin		"Consistency in representation"
Lee, Pipino, Funk, Wang		"Same element in different tables"

*Grey cells represent a lack of coverage of a concept by associated author.

Although both Redman and English cite the correct source concept inside of accuracy, TDWI and DMBOK cite it within Consistency. I propose that we move this concept from the dimension of Accuracy and place it within Consistency, as the primary concept. This would give a majority (four out of six) agreement within this dimension as shown in Table 5.

It may be helpful at this point to note that by saying “correct source,” I mean the correct data system or file/table. Based on my review, I didn’t come to the conclusion that existence in reality is a source, but rather a separate concept. This implies that either I compare my data to real-world observation *or* to a data source — they are not the same thing, even though the data source may agree with the real-world observation.

English has another dimension titled “Source Quality & Security Warranties or Certifications,” composed of the following metrics.

1. Guarantees Quality-: Guarantees the quality of information it provides with remedies for non-compliance.

2. Documents Certification: Documents its certification in its Information Quality Management capabilities to capture, maintain and deliver Quality Information.

3. Provides Measures: Provides objective and verifiable measures of the Quality of Information it provides in agreed-upon Quality Characteristics.

4. Guarantees Unauthorized Access: Guarantees that the Information has been protected from unauthorized access or modification.

Because this proposed dimension isn’t another concept, but rather detail for the consistency/sourcing concept, I’d move the first three of these measures into Consistency. The last one is a concept covered in the next article in this series, on the Accessibility dimension.

As is the case with a couple of the concepts within the dimensions of quality, data security and access controls, falls on the line or well within a discipline other than data quality. Information Security is a well-established domain with many more written works and established organizations, certifications, conferences, laws, standards and training curriculums than data quality. For this reason, many people acknowledge this data security concept, but in terms of areas of responsibility, elect to have separate dedicated IT security departments handle these aspects.

As seen in Table 5, although the last two authors (Loshin and Lee, et al.) don’t include the sourcing concept, they do bring additional value through their insights into the concept of Consistency in Representation:

• Referential: Refers to the consistency of redundant data in one table or in multiple tables.

• Logical/Structural: Consistency between two related data elements (e.g., city name and postal code).

• Format: Consistency of format for the same data element used in different tables.

• Semantic: Consistency of definitions among attributes within a data model.

Table 5: Accuracy and Consistency Dimension Similarity by Author

	Tom Redman	Larry English	TDWI	DMBOK	David Loshin	Lee, Pipino, Funk, Wang
Accuracy (Factualness)	Accurate	Accuracy (reality, surrogate source)	Accuracy	Accuracy	Accuracy	Free of Error
Consistency	(included in Redman's definition of Accuracy)	(included in English's definition of Accuracy)	Consistency & Dependency	Consistency	Consistency (Structural, Semantic)	Consistency (Referential, Logical, Format)

*Green cells represent strong similarity/agreement between authors for a given dimension. Yellow cells represent areas of weak agreement or difference in categorization.

The word “accuracy” regarding data quality is often used too broadly. For instance, if we receive a bill for services and it is understated by $5 for parts that were purchased in addition to the services, one might say that the bill is not “accurate.” From a data quality perspective though, this concept is referred to as completeness, where all the data needed for its intended use is not available. By using a word other than “accuracy” for this dimension, we avoid ambiguity and more effectively diagnose the problem.

Until now, we have clarified the term “accuracy” to mean “Agreement with the real-world,” but, because this word is so universally used – almost to the extent that it is synonymous with “quality.” After discussing this with Danette McGilvray, we agree that the industry should use something more distinctive that doesn’t mean so many things to everyone. Personally I find the word "Factualness" to represent the concept well.

(Editor’s Note: For part 2 in the series, on reasonability, click here. For part 3, on completeness, click here.)

***References**:*

Redman, Tom. "Data Quality: The Field Guide," Digital Press 2001.
English, Larry. "Information Quality Applied," Wiley Publishing, 2009.
TDWI. "Data Quality Fundamentals," The Data Warehousing Institute, 2011.
DAMA International. "The DAMA Guide to The Data Management Body of Knowledge" (DAMA-DMBOK Guide) Technics Publications, LLC, 2009.
Loshin, David. "The Practitioner's Guide to Data Quality Improvement," Elsevier 2011.
Yang W. Lee, Leo L. Pipino, James D. Funk, Richard Y. Wang. "Journey to Data Quality," MIT Press 2006.
McGilvray, Danette. "Executing Data Quality Projects- Ten Steps to Quality Data and Trusted Information," Morgan Kaufmann, 2008.

Dan Myers currently manages enterprise data management initiatives for Farmers Insurance. At Farmers he has also managed data and functional B.I. testing. Dan conducted an extensive metadata software review implemented Farmers first enterprise-wide metadata repository. Dan led a committee to review and select data quality tools for Farmers and wrote a comprehensive report comparing the industry's top DQ tools. He authored the Farmers' governance policy for integration/sourcing, metadata management, and data quality. Previously Dan has worked as an independent Oracle Certified Professional consultant in both front and back-end development capacities. Dan's fluency in Japanese enabled him to work in both the public and private sector in Japan. Dan received his MBA from the U.S.C. Marshall School of Business in 2009.

(http://www.information-management.com)

By Dan Myers
JUN 26, 2013 8:51am ET

FEATURE

Examining Dimensions of Data Quality: Reasonability, Time and Access

By Dan Myers
JUN 26, 2013 8:51am ET

The first article in this series clarified what areas of agreement exist for three of the dimensions of data quality (accuracy, precision, consistency) between six of the DQ industry's authorities. This article addresses the reasonability, time and access aspects of data quality.

Because the last article discussed Consistency, a natural place to continue is the related area of Reasonableness, which is often confused with Consistency. The following authors espouse Reasonableness or Believability.

Table 1: "Reasonableness" Dimension Similarities by Author

	Tom Redman	Larry English	TDWI	DMBOK	David Loshin	Lee, Pipino, Funk, Wang
Reasonableness				Reasonableness	Reasonableness	Believability

*Green cells represent strong similarity/agreement between authors for a given dimension. Grey cells represent a lack of coverage of a concept by author.

When we look closer, however, only two authors (Loshin and Lee et al.) identify a new concept not already covered. The DMBOK and Loshin identify consistency of values, which we covered in the last article. Loshin's identification of the time-related aspect of reasonability is spot on, but I'd classify that as simply a domain of acceptable values (though date constrained), which we will cover in the Validity dimension in the fourth article in this series. Rational expectations, which are labeled "reasonable," can also be documented as validity ranges, minimums, maximums and other basic business rules. By documenting these requirements as rules used during profiling, the properties of the data can be measured and managed in an unbiased way.

Table 2: Concepts within "Reasonableness"

Author/ Source	Consistency within Operational Context	Temporal Reasonability	Data Meets Rational Expectations	Regarded as True and Credible
Tom Redman				
Larry English				
TDWI				
DMBOK	Consistency within Operational Context			
David Loshin	Consistency or Reasonability of Values	Temporal Reasonability	Data Met Rational Expectations	
Lee, Pipino, Funk, Wang				Regarded as True and Credible

Lee et al. cite believability as “regarded as true and credible,” but that is very subjective and not a property of the data as much as an opinion of its fitness for use by consumers.

As discussed at the beginning of this series, dimensions are properties of the data relative to its fitness, and we’ve either placed these three concepts (Temporal Reasonability, Meets Rational Expectations or Regarded as True and Credible) in other dimensions or dismissed them as not meetings the criteria of a dimension because it isn’t a property of the data. That being said, surveying end users’ opinions of data desirability is valuable in the context of bigger data quality improvement, but doesn’t fit within the scope of the dimensions of data quality because they are attributes of the customer’s need, not inherent attributes of the data.

There is much more agreement regarding the next dimensions that we’ll cover. All of the authors espouse the Timeliness dimension.

Table 3: “Timeliness,” "Currency" and “Accessibility” Dimension Similarities by Author

	Tom Redman	Larry English	TDWI	DMBOK	David Loshin	Lee, Pipino, Funk, Wang
Timeliness	Timely	(included in English's definition of Accessibility Timeliness)	Timeliness	Timeliness	Timeliness	Timeliness
Currency		- Currency - Concurrence of Redundant or Distributed Data		Currency	Currency	
Accessibility	Accessibility /Delivery	Accessibility/ Availability	(included in TDWI's Privacy & Representation dimensions)	(included in DMBOK's Timeliness and Privacy dimensions)	(included in Loshin's Timeliness dimension)	Accessibility, Appropriate Amount of Data

*Green cells represent strong similarity/agreement between authors for a given dimension. Yellow cells represent weak agreement or difference in categorization. Grey cells represent a lack of coverage of a concept by author.

At first glance one may think that Timeliness and Currency are the same concept, but that isn’t the case. Currency focuses on how up-to-date or how “fresh” data is, reflecting the real-world concept. Timeliness is related to how quickly a stakeholder can gain access to the data needed. An example of this might be when a data mart is loaded with daily granularity sales data once a month, meaning that users can create daily purchase reports but there is a one-month lag between the day that the report represents and the earliest day it can be viewed in the data mart.

Lee et al. call out the Appropriate Amount of Data as well, but that is only a volume metric within the Accessibility concept. In addition to Currency, some authors cite the “Concurrence of Distributed Data” concept, as seen in Table 4.

Table 4: Concepts Within Currency Dimension		
Author/ Source	**Current with World it Models**	**Concurrence of Distributed Data**
Tom Redman	"Current if Up-To-date"	
Larry English	"Age is Correct for Purpose"	"Lag time between when data in system (a) is queried in system (b)"
TDWI	"Age or Freshness of Data"	
DMBOK	"Info. current with World it Models"	
David Loshin	"Age/Freshness"	"Synchronization/Replication"
Lee, Pipino, Funk, Wang	"How Up-To-Date the Data is"	

*Grey cells represent a lack of coverage of a concept by associated author.

Within Timeliness, there is an additional concept of Retention that only the TDWI references. This is especially important to records coordinators within compliance and legal functions who require that documents are properly disposed of after a set period of time.

Table 5: Concepts Within Timeliness Dimension		
Author/ Source	**Time Expectation for Availability**	**Retention**
Tom Redman	"Degree process completed in pre-specified time"	
Larry English		
TDWI		Retention
DMBOK	"Time Expectation for Accessibility and Availability"	
David Loshin	"Time Expectation for Availability"	
Lee, Pipino, Funk, Wang		

*Grey cells represent a lack of coverage of a concept by associated author.

The next article in this series looks at Completeness, which I believe is the most fundamental place to start a data quality effort.

***References**:*

Redman, Tom. "Data Quality: The Field Guide," Digital Press 2001.
English, Larry. "Information Quality Applied," Wiley Publishing, 2009.
TDWI. "Data Quality Fundamentals," The Data Warehousing Institute, 2011.
DAMA International. "The DAMA Guide to The Data Management Body of Knowledge" (DAMA-DMBOK Guide) Technics Publications, LLC, 2009.
Loshin, David. "The Practitioner's Guide to Data Quality Improvement," Elsevier 2011.
Yang W. Lee, Leo L. Pipino, James D. Funk, Richard Y. Wang. "Journey to Data Quality," MIT Press 2006.
McGilvray, Danette. "Executing Data Quality Projects- Ten Steps to Quality Data and Trusted Information," Morgan Kaufmann, 2008.

Dan Myers currently manages enterprise data management initiatives for Farmers Insurance. At Farmers he has also managed data and functional B.I. testing. Dan conducted an extensive metadata software review implemented Farmers first enterprise-wide metadata repository. Dan led a committee to review and select data quality tools for Farmers and wrote a comprehensive report comparing the industry's top DQ tools. He authored the Farmers' governance policy for integration/sourcing, metadata management, and data quality. Previously Dan has worked as an independent Oracle Certified Professional consultant in both front and back-end development capacities. Dan's fluency in Japanese enabled him to work in both the public and private sector in Japan. Dan received his MBA from the U.S.C. Marshall School of Business in 2009.

(http://www.information-management.com)

By Dan Myers
JUL 2, 2013 10:52am ET

FEATURE

Examining Dimensions of Data Quality: Completeness

By Dan Myers
JUL 2, 2013 10:52am ET

The last article in this series looked at six authors' definitions of two related dimensions of data quality (Timeliness and Accessibility). In this article, we'll look at one of the most foundational dimensions, Completeness.

At a high level, Completeness is intuitive. The key to measuring Completeness (or anything in this world, for that matter) is to identify the data's characteristics and then compare those known attributes at a later time to test whether they have changed, in this case whether they have changed from NULL to NOT NULL or vice versa.

The following illustration of a delimited file transmitted from one system to another shows the two primary types of completeness: row-level and column-level.

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

As you can see in the illustration, the file has four physical rows composed of three data rows and a header. Although you can calculate the number of rows as the literal four rows, we usually exclude the header from any counts or amounts to avoid confusion. If I included the header in the count/amount for each column in this file, how would I include 'ITEM_PRICE' in the sum of amounts ($15.25, $33.12, $24.95)?

When data is moved from one location to another, Completeness is a concern. In order to identify the loss of data, we need two measures of completeness for a two dimensional data set (e.g., table of data).

1. **Row-level Completeness**: First and most importantly, to validate completeness we measure the count of physical rows in the file and then

measure that again after we move the data to make sure we have all the observations.

2. **Column-level Completeness**: Next, we count the values of each column or aggregate the values (if logically able), as seen in the ITEM_PRICE column in the previous illustration. This ensures that, on the whole, the data is the same.

Unlike the other dimensions covered so far in this series of articles, there is complete agreement with the primary concept of this dimension: column-level population.

Table 1: "Completeness" Dimension Similarity by Author

	Tom Redman	Larry English	TDWI	DMBOK	David Loshin	Lee, Pipino, Funk, Wang
Completeness	Complete	· Completeness (value, fact) · Existence (record, value)	Completeness	Completeness	Completeness	Completeness (schema, column, population)

All six authors include column-level population within their definitions. This is one of the key measures of data profiling tools (e.g., Null counts and percentages). Other common examples of mechanisms to control completeness are database constraints that enforce null-ability and form validation implemented within the application, typically using JavaScript.

Table 2: Concepts within Completeness

Author/ Source	Table Population	Column Population	Row Population	Schema Population
Tom Redman		"Required Attributes"	"Required Records"	
Larry English		"Required Element"	"Required Facts"	
TDWI	"Entity Occurrence"	"Element Available"	"Transaction Available"	
DMBOK	"Dataset"	"Certain Attribute"	"Appropriate Rows"	
David Loshin		"Mandatory or Optional Attribute"		
Lee, Pipino, Funk, Wang		"Column Completeness"		"Schema Population"

*Grey cells represent a lack of coverage of a concept by associated author.

In the financial services industry, row completeness is also very important (e.g., if a transaction was not recorded/moved, all associated revenues/premiums are likely understated). A consolidated list of concepts within the Completeness dimension must have the column and row levels, but in my experience, schema and table levels aren't frequently measured.

(Editor's note: Look for part four of this series next Thursday. For part 2 on reasonability, click here. For the introductory article, click here.)

***References**:*

Redman, Tom. "Data Quality: The Field Guide," Digital Press 2001.
English, Larry. "Information Quality Applied," Wiley Publishing, 2009.
TDWI. "Data Quality Fundamentals," The Data Warehousing Institute, 2011.
DAMA International. "The DAMA Guide to The Data Management Body of Knowledge" (DAMA-DMBOK Guide) Technics Publications, LLC, 2009.

Loshin, David. "The Practitioner's Guide to Data Quality Improvement," Elsevier 2011.
Yang W. Lee, Leo L. Pipino, James D. Funk, Richard Y. Wang. "Journey to Data Quality," MIT Press 2006.
McGilvray, Danette. "Executing Data Quality Projects- Ten Steps to Quality Data and Trusted Information," Morgan Kaufmann, 2008.

Dan Myers currently manages enterprise data management initiatives for Farmers Insurance. At Farmers he has also managed data and functional B.I. testing. Dan conducted an extensive metadata software review implemented Farmers first enterprise-wide metadata repository. Dan led a committee to review and select data quality tools for Farmers and wrote a comprehensive report comparing the industry's top DQ tools. He authored the Farmers' governance policy for integration/sourcing, metadata management, and data quality. Previously Dan has worked as an independent Oracle Certified Professional consultant in both front and back-end development capacities. Dan's fluency in Japanese enabled him to work in both the public and private sector in Japan. Dan received his MBA from the U.S.C. Marshall School of Business in 2009.

(http://www.information-management.com)

By Dan Myers
JUL 11, 2013 4:18am ET

FEATURE

Examining Dimensions of Data Quality: Validity and Integrity

By Dan Myers
JUL 11, 2013 4:18am ET

Have you ever heard someone say that a statistic is valid, but inaccurate? Or perhaps they adamantly argue with the IT department that, although it isn't in the list of valid values (and fails an error check), the value is accurate (factual). In this article, we'll build on what we discussed in prior articles in this series regarding the dimensions of data quality and look more closely at Validity and Integrity. Below is a comparison of six data quality authors' agreement with the Validity dimension.

Table 1: Validity Dimension Similarity by Author

	Tom Redman	Larry English	TDWI	DMBOK	David Loshin	Lee, Pipino, Funk, Wang
Validity	(Included in Redman's definition of Consistency)	Validity (value, business rule, derivation)	(Included in TDWI's definition of Integrity)	Validity	(Included in Loshin's definition of Accuracy)	(Included as component of author's definition of Column Integrity)

*Green cells represent strong similarity/agreement between authors for a given dimension. Yellow cells represent areas of weak agreement or other difference in categorization. Grey cells represent a lack of coverage by author.

As discussed in the first article, there is relative agreement on the Accuracy dimension, but there is some confusion around the Validity dimension, which is distinctly different. Although people often use the words valid or invalid when they are expressing whether data is factual or not, the words hold different implications when considered in data management/quality context.

The question at the beginning of this article referred to situations where a value can be valid (within a set of predefined accepted values), like "CA" within the list of U.S. state abbreviations, but inaccurate (not factual). One example may be a piece of mail that is intended for a destination in Alaska, but is mistakenly addressed with "AL" (Alabama).

Conversely, many advanced systems now check that a value is within a set of specified valid values and report errors (or even automatically correct the mistake based on some default logic). In this scenario, a factual value may be rejected if the system doesn't have that value within its list of expected values. An example of this may be an insurance policy processing system that rejects a homeowner's address in a state that the insurer didn't

(until very recently) conduct business. Once the system's list of valid states has been updated with the new state value, the entry would be factual and recognized as valid.

As shown in Table 1, there is some consensus on the concepts in this dimension, with some focus on "Values in Specified Range of Valid Values" concept. Loshin places this within Accuracy, and Lee et al. place it within Integrity (see Table 1), but Loshin's April 1, 2011 blog post (http://www.b-eye-network.com/blogs/loshin/) , implies his agreement that data validity and data correctness are different concepts.

Table 2: Concepts within Validity

Author/ Source	Values in Specified Range of Valid Values	Values Conform to Business Rule	Conform to Other Attribute Types
Tom Redman			
Larry English	"Values in specified range of valid values"	"Values conform to business rule" and "derivation correct"	
TDWI			
DMBOK	"Consistent with domain of Values"		"Values conform to numerous attributes associated: data type, precision, format, etc."
David Loshin			
Lee, Pipino, Funk, Wang			

*Grey cells represent a lack of coverage of a concept by associated author.

The process of doing the research and writing this article has been rewarding for me because, as I suspected, knowledge and agreement improve as authors discuss the concepts and consider the best way to express concepts. In discussion with Tom Redman prior to publishing this article, he observed that the concepts of Validity were named Consistency in his book, but now he prefers the term Validity.

So now that we've discussed how we might normalize Validity, let's turn to Integrity (Table 3). Coming from a data modeling background, I find this dimension the most straightforward and common-sense orientated. I have found that IT departments are better equipped to measure and remedy these Integrity concepts, unlike valid value/reference data management often required of business subject matter experts done during validity activities.

Table 3: "Integrity" Dimension Similarity by Author

	Tom Redman	Larry English	TDWI	DMBOK	David Loshin	Lee, Pipino, Funk, Wang
Integrity			Integrity/ Uniqueness	Referential Integrity	(Somewhat included in Loshin's Structure component of Consistency; also somewhat included in Identifiability)	Codd Integrity Constraints · Entity · Referential · Domain · Column

*Green cells represent strong similarity/agreement between authors for a given dimension. Yellow cells represent areas of weak agreement or other difference in categorization. Grey cells represent a lack of coverage by author.

As seen in Table 4, moving the "Values in Specified Range of Valid Values" (Domain) concept into the Validity dimension allows us to focus on relational concepts pure to Integrity. Four of the six authors reference "Referential Integrity," with some going further into similar components (basically the tenets of E.F. Codd's database normalization).

I am unsure why there isn't greater agreement among authors relative to the concepts that are included within Integrity. I suspect that there is a general assumption that these are done through the data modeling process and, therefore, aren't explicitly called out here. Most data profiling tools offer functionality to ensure these concepts. If you are in the market to purchase a profiler, I recommend you validate that the vendor solution sufficiently provides this capability.

Table 4: Concepts within Integrity

Author/ Source	Referential Integrity	Unique Identifier of Entity	Cardinality	Domain
Tom Redman		Identifiability		
Larry English				
TDWI	"Referential integrity"	"Uniquely identifiable occurrences of every entity"	"Cardinality"	"Domain"
DMBOK	"One column to another column"	"Each entity occurs once"		
David Loshin	"Consistency in representation"			
Lee, Pipino, Funk, Wang	"Foreign key match primary key value"	"Entity, no primary key be null"		"Column integrity"

At this point, it should be noted that many authors call out Unwanted Duplication as a separate dimension. The equivalent concept covered by these six authors is named "Unique Identifier of Entity" in Table 4. I believe that because all of Codd's tenets of normalization can be identified within one dimension named Integrity, we don't need a distinct dimension for Duplication. Furthermore duplication, as a concept, isn't always a data quality problem because sometimes data solutions intentionally allow for a level of Intended Duplication, but which still have unique identifiers (surrogate keys) to improve query performance.

(Editor's note: Look for part 5 of this series next Thursday.)

***References**:*

Redman, Tom. "Data Quality: The Field Guide," Digital Press 2001.
English, Larry. "Information Quality Applied," Wiley Publishing, 2009.
TDWI. "Data Quality Fundamentals," The Data Warehousing Institute, 2011.
DAMA International. "The DAMA Guide to The Data Management Body of Knowledge" (DAMA-DMBOK Guide) Technics Publications, LLC, 2009.
Loshin, David. "The Practitioner's Guide to Data Quality Improvement," Elsevier 2011.
Yang W. Lee, Leo L. Pipino, James D. Funk, Richard Y. Wang. "Journey to Data Quality," MIT Press 2006.
McGilvray, Danette. "Executing Data Quality Projects- Ten Steps to Quality Data and Trusted Information," Morgan Kaufmann, 2008.

Dan Myers currently manages enterprise data management initiatives for Farmers Insurance. At Farmers he has also managed data and functional B.I. testing. Dan conducted an extensive metadata software review implemented Farmers first enterprise-wide metadata repository. Dan led a committee to review and select data quality tools for

Farmers and wrote a comprehensive report comparing the industry's top DQ tools. He authored the Farmers' governance policy for integration/sourcing, metadata management, and data quality. Previously Dan has worked as an independent Oracle Certified Professional consultant in both front and back-end development capacities. Dan's fluency in Japanese enabled him to work in both the public and private sector in Japan. Dan received his MBA from the U.S.C. Marshall School of Business in 2009.

(http://www.information-management.com)

By Dan Myers
JUL 17, 2013 2:45pm ET

FEATURE

Examining Dimensions of Data Quality: Definition and Representation

By Dan Myers
JUL 17, 2013 2:45pm ET

The previous articles in this series covered all of the standard dimensions of quality with the exception of two that are sometimes forgotten: Definition and Representation. Personally these are two of my favorite dimensions. (It should be noted that within this article the term Representation is synonymous with Presentation.)

Often data quality issues are not about the transformation of the data, but rather the awkward or misleading definitions. More often than not there are no descriptions or report captions at all.

As one of my co-workers pointed out, the challenge isn't so much about data quality as it is about educating people regarding what the data means and how to use it — concepts tightly related to Definition and Representation. Sometimes a new training program is the solution to removing the impression that data is of poor quality/fitness.

These dimensions present a challenge because of the similarity between the two, so let's first review the concepts presented by the six data quality authors discussed in this series for each area and then normalize what we find. Three authors identify the "Definition" dimension, but because we already covered the "Values Consistent with Definition" concept proposed by Loshin and English in the Consistency dimension, we only need to deal with two concepts provided by Redman:

1. Clear, easy to understand definition.
2. Includes measurement units.

The two concepts identified by Redman seem to fit well within the Representation dimension because these definitions are logically a subcategory of Representation. Table 1 outlines the concepts within the "Representation" dimension cited by three authors.

Table 1: Concepts within Representation

Author/ Source	Easy to Read and Interpret	Presentation Language	Media Appropriate	Accessibility	Complete and Available Metadata
Tom Redman	"Easy to read and interpret"	"Clear and tech terms fully defined"			
Larry English	"Presented in consistent, standard way"	"Signage accuracy and clarity"	"Media: Online, hardcopy, audio"		
TDWI	"Understanding of data"	"Presentation and visualization"		"Accessibility of data"	"Complete and available metadata"
DMBOK					
David Loshin					
Lee, Pipino, Funk, Wang					

*It should be noted that TDWI also called out two additional components "Ease of navigation" and "Multidimensional access and analysis," but since these appear to be tool/usability requirements they were left off of this table of concepts.

Adding Redman's Definition concepts to the Representation dimension and removing the Accessibility dimension cited by TDWI (which we already addressed in the Accessibility dimension in part 2 of this series) provides us with a comprehensive new Representation dimension as seen in Table 2. Redman also goes further, introducing the dimension titled Relevance: "Data are relevant to a particular task or decision if they contribute to the completion of that task or making of the decision" (Redman, 226). One may propose including this concept within Representation, but this dimension is outside the scope of DQ, defined as "fitness for use," because if data isn't meant for use then it will not be relevant.

Another authority in the data quality space, Danette McGilvray, also adds a "Data Specifications" dimension (defined as the measure of the existence, completeness, quality and documentation of data standards, data models, business rules, metadata and reference data) that, "…provides the standard against which to compare data quality assessment results. They also provide instruction for manually entering data, designing data load programs, updating information, and developing applications" (McGilvray, 31).The definition (Representation), format and derivation (Validity), data load and data model (Integrity) have been covered in previous articles in this series. Having said that, integration of data quality standards within IT requirements is absolutely critical to success, but it has to be done per stakeholder group because fitness for use may differ by consumer or even business process.

Table 2: Conformed Representation Dimension by Author

	Tom Redman	Larry English	TDWI	DMBOK	David Loshin	Lee, Pipino, Funk, Wang
Representation	Clear definition	Definition conformance	(included in TDWI's Usability dimension as "Availability and completeness of metadata")		Semantic	

*Green cells represent strong similarity/agreement between authors for a given dimension. Yellow cells represent areas of weak agreement or difference in categorization. Grey cells represent a lack of coverage by author.

I suspect that the DMBOK authors opted not to called "Representation" out explicitly as a dimension of quality because a whole chapter of the DMBOK is devoted to metadata management. Also, Lee and Wang cite a category of information quality, "Representational IQ" in their 1997 paper titled "10 Potholes in the Road to Information Quality" where they

list “Interpretability, Ease of Understanding, Concise Representation, Consistent Representation,” so there is more agreement than may appear.

Loshin added another dimension that we haven’t covered, called Lineage. He defines it as the “Originating data source,” with the additional clarification: “All data elements will include an attribute identifying its original source and date. All updated data elements will include an identifier for the source of the update and a date. Audit trails of all provenance data will be kept and archived” (Loshin, 136).

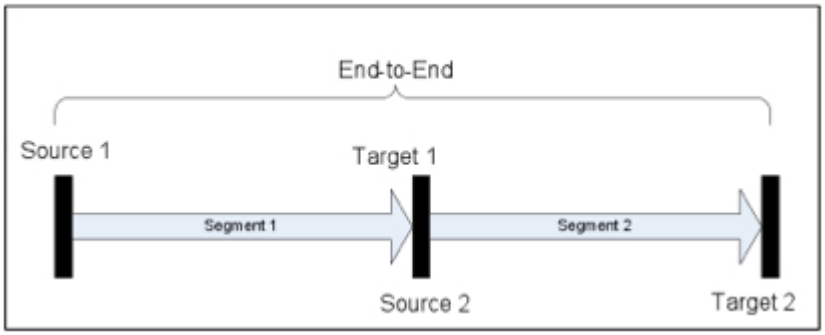
In a broader sense, lineage can imply the collection of all metadata about where data came from and how it was transformed along the way. Until now, I have normalized dimensions recommended by each author, primarily taking a majority-consensus approach. Regarding lineage, where only Loshin calls it out, I agree that even though only one author has identified it, we should include it in an industry set of DQ dimensions. In an interview with Tom Redman, he also agrees that this is a unique contribution that Loshin brought to the field.

Lineage is a valid dimension of data quality because:

1. Using the concepts of lineage identify risk not included in other dimensions of data quality.(For instance, a large number of segments/transformations increases the risk that the data was incorrectly changed. This helps practitioners measure and prioritizes DQ issues.)
2. Using the concepts of lineage identify cost not included in other dimensions of data quality.(As an example, various stakeholders may consume the same sales data, deriving it in tens of unique ways, reducing consistency and increasing complexity, which are IT cost drivers.)

Lineage, like other dimensions of quality, can be used in conjunction with other dimensions to add value. For example, by providing a lineage of the data from end to end with embedded completeness measures for each segment, one can evaluate the total completeness, inclusive of all movement and transformation.

Figure 1



In order to flesh out this dimension, I have outlined the concepts found and a few attributes, as noted below. Figure 1 will look very familiar to those familiar with ETL processes used in data warehousing, illustrating the beginning to end data flow.

Concepts:

1) **Segment**: Movement or transformation of data having a beginning point, called a source, and an end-point, called a target.

Attributes:
a) Derivation code (e.g., SQL, RegEx … etc.)
b) Derivation description (typically pseudo-code/simple fragment plain English)
c) Derivation type (e.g., pass through versus derived)

2) **Source**: Beginning point of data movement or transformation.

Attributes:
a) Source level [e.g., primary source (1st), secondary source (2nd) … nth]
b) Source system name
c) Source type or technology

3) **Target**: End point of data movement or transformation.

Attributes:
a) Target level (see source level)
b) Target system name
c) Target type or technology

4) **End-to-end**: The multisegment definition of data movement or transformation, inclusive of all intermediate segments to provide data.

Attributes:
a) Total number of segments
b) Average {dimension of DQ} (e.g., Average Completeness for three segments)
c) Certification (measure of how thoroughly systems integration testing has been conducted)

In conclusion, we can normalize Definition into the Representation dimension listing the five concepts (1. Easy to Read and Interpret, 2. Presentation Language, 3. Media Appropriate, 4. Includes Measurement Units, and 5. Complete and Available Metadata). In addition, I added reasons why Loshin's Lineage dimension should be included within the industry list of DQ dimensions, providing concepts and example attributes. The next article (the last in this series) will compile each of my recommendations into a single industry-standard list with basic definitions and concepts.

All references to authors' works come from the following sources:

Redman, Tom. "Data Quality: The Field Guide," Digital Press 2001.
English, Larry. "Information Quality Applied," Wiley Publishing, 2009.
TDWI. "Data Quality Fundamentals," The Data Warehousing Institute, 2011.
DAMA International. "The DAMA Guide to The Data Management Body of Knowledge" (DAMA-DMBOK Guide) Technics Publications, LLC, 2009.
Loshin, David. "The Practitioner's Guide to Data Quality Improvement," Elsevier 2011.

Yang W. Lee, Leo L. Pipino, James D. Funk, Richard Y. Wang. "Journey to Data Quality," MIT Press 2006.
McGilvray, Danette. "Executing Data Quality Projects- Ten Steps to Quality Data and Trusted Information," Morgan Kaufmann, 2008.

Dan Myers currently manages enterprise data management initiatives for Farmers Insurance. At Farmers he has also managed data and functional B.I. testing. Dan conducted an extensive metadata software review implemented Farmers first enterprise-wide metadata repository. Dan led a committee to review and select data quality tools for Farmers and wrote a comprehensive report comparing the industry's top DQ tools. He authored the Farmers' governance policy for integration/sourcing, metadata management, and data quality. Previously Dan has worked as an independent Oracle Certified Professional consultant in both front and back-end development capacities. Dan's fluency in Japanese enabled him to work in both the public and private sector in Japan. Dan received his MBA from the U.S.C. Marshall School of Business in 2009.

(http://www.information-management.com)

By Dan Myers
AUG 1, 2013 1:04pm ET

FEATURE

The Value of Using the Dimensions of Data Quality

By Dan Myers
AUG 1, 2013 1:04pm ET

The first five articles in this series contrasted the dimensions of data quality defined by six renowned authors - including valuable points from additional authors as applicable. Clearly there are many valuable aspects to the dimensions of data quality:

- The categorization of data by quality properties allows prospective consumers to evaluate whether the data meets their needs in terms of its current properties (completeness, precision, etc.).
- The categorization of data by quality properties provides a mechanism to prioritize data quality cleanup, process changes and implement data stewardship/governance.
- Dimensions (and, more specifically, the underlying concepts with the associated metrics) provide a method of measuring quality over time.
- The categorization of data by quality properties allows practitioners to predict business impact based on known behavior of each dimension of quality (e.g., lack of completeness yields understated financials, invalid values can lead to miscategorization or aggregation).

The purpose for having an industry-accepted set of dimensions with associated concepts is to allow organizations to effectively communicate internally and externally. In a more networked society, where there are more external demands on our data, such as governmental regulation, legal, security, corporate partnerships and corporate valuation, agreed-upon standards are a must.

In a recent discussion on this topic with data quality author Danette McGilvray, she pointed out that from an internal perspective, the quicker an organization can establish and start using these foundational dimensions, the sooner they will see the benefits. Why not get a jump-start using the industry standard and then add custom categories and concepts as needed?

Bringing it All Together

In this capstone article, I've compiled the proposed list of dimensions Figure 1 lists the dimensions identified by the data quality authors and associated concepts before standardization. Note the red arrows crossing the vertical dashed lines indicate where

authors cited concepts within other dimensions. Using this charting method, the optimal relationship would have dimensions with underlying concepts only within each individual column — no red dashed arrows. (Click here to open Figure 1 (http://cdn.information-management.com/media/newspics/ConceptsWithInTheDimensionsOfDataQuality.jpg) .)

Figure 1 lists concepts, independent of author. Table 1 provides a side-by-side comparison of the dimensions between authors, as covered in articles one through five of this series. (Click here to open Table 1 (http://cdn.information-management.com/media/newspics/SideBySideComparisonByAuthor.jpg) .)

Someone will likely disagree with the way these have been conformed, but as everyone who participates in data governance knows, there has to be some compromise in order to create a standard. I think the following is palatable to most of the authors cited and true to the underlying reasons for each concept.

It should be noted, though, that this work has not taken into account the direct impact of unstructured data quality (e.g., textual documents, video, audio, etc.), and over time we'd expect that the number of concepts documented under these dimensions would grow and other dimensions will likely be introduced. The industry standard will likely be a living cannon of the agreed-upon dimensions.

The consolidated list of dimensions of data quality and underlying concepts, based on the consolidation in articles one through five, are listed in Table 2. (Click here to open Table 2 (http://cdn.information-management.com/media/newspics/ConformedDimensionsOfDataQuality.jpg) .)

It should be noted that this is not a list of definitions of the dimensions, which would require an extensive review, negotiation and compromise effort among industry thought leadership. Rather, this is a conformed list of the underlying concepts for each dimension. (I am presenting this topic at the International Association for Information and Data Quality (http://iaidq.org/) Conferences called IDQ 2013 (http://iaidq.org/idq2013/) in Little Rock, AR this November. I hope to see you there and discuss this topic further.)

In conclusion, I stress that although many of the dimensions put forth by data quality authors are good mechanisms to ensure quality information management work products, they aren't specific to the quality of data and its intended use.

This is where we should go back to the standard definition for data quality: "Fitness for Use," which is a misnomer. It should be "Fitness for *intended* use." After all, we wouldn't say that a Ferrari is of poor quality when used off-roading, would we? Rather it is of exceptional quality for its purpose (aesthetic beauty, acceleration, high-speed maneuvering on flat surfaces, etc.). In terms of creating standards, the presumption has to be that the data is for a given purpose/audience, and *then* within that scope we can define whether it meets our needs or not.

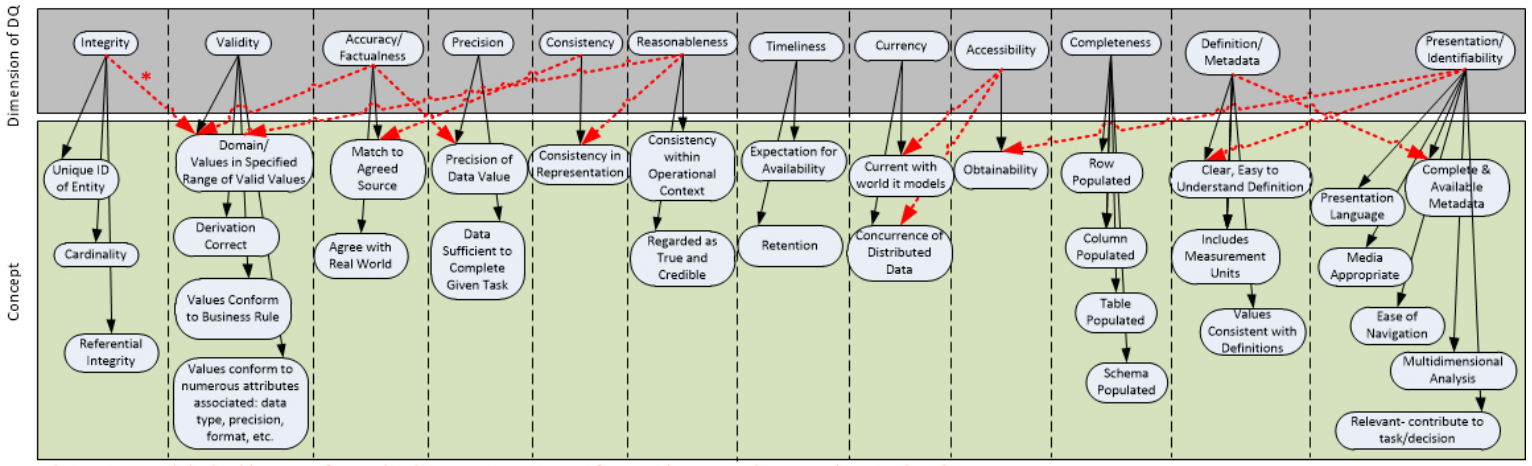
Read the rest of this series:

All references to authors' works in this series come from the following sources:

Redman, Tom. Data Quality: The Field Guide, Digital Press 2001.
English, Larry. Information Quality Applied, Wiley Publishing, 2009.
TDWI, Data Quality Fundamentals, The Data Warehousing Institute, 2011.
DAMA International. The DAMA Guide to The Data Management Body of Knowledge (DAMA-DMBOK Guide) Technics Publications, LLC, 2009.
Loshin, David. The Practitioner's Guide to Data Quality Improvement, Elsevier 2011.
Yang W. Lee, Leo L. Pipino, James D. Funk, Richard Y. Wang. Journey to Data Quality, MIT Press 2006.
McGilvray, Danette. Executing Data Quality Projects- Ten Steps to Quality Data and Trusted Information, Morgan Kaufmann, 2008.

Dan Myers currently manages enterprise data management initiatives for Farmers Insurance. At Farmers he has also managed data and functional B.I. testing. Dan conducted an extensive metadata software review implemented Farmers first enterprise-wide metadata repository. Dan led a committee to review and select data quality tools for Farmers and wrote a comprehensive report comparing the industry's top DQ tools. He authored the Farmers' governance policy for integration/sourcing, metadata management, and data quality. Previously Dan has worked as an independent Oracle Certified Professional consultant in both front and back-end development capacities. Dan's fluency in Japanese enabled him to work in both the public and private sector in Japan. Dan received his MBA from the U.S.C. Marshall School of Business in 2009.

Concepts within the Dimensions of Data Quality



*Arrows in red dashed line signify overlap between concepts referenced in more than one data quality dimension

Table 1. Side-by-side comparison of dimensions of quality by author.

Tom Redman	Larry English	The Data Warehousing Institute (TDWI)	Data Management Body of Knowledge (DMBOK)	David Loshin	Lee, Pipino, Funk, Wang
Complete	· Completeness (value, fact) · Existence (record, value)	Completeness	Completeness	Completeness	Completeness (schema, column, population)
Accurate	Accuracy (reality, surrogate source)	Accuracy	Accuracy	Accuracy	Free of Error
Timely	(included in English's definition of Accessibility Timeliness)	Timeliness	Timeliness	Timeliness	Timeliness
	· Currency · Concurrence of Redundant or Distributed Data		Currency	Currency	
(included in Redman's definition of Accuracy)	(included in English's definition of Accuracy)	Consistency & Dependency	Consistency	Consistency (Structural, Semantic)	Consistency (Referential, Logical, Format)
Accessibility/Delivery	Accessibility / Availability	(included in TDWI's Privacy and Representation dimensions)	(included in DMBOK's Timeliness and Privacy dimensions)	(included in Loshin's Timeliness dimension)	Accessibility, Appropriate Amount of Data
Proper Level Of Detail	Non-duplication	Granularity	Uniqueness		
Format Precision	Precision	Precision	Precision	(included in Loshin's definition of Accuracy)	
		Integrity / Uniqueness	Referential Integrity	(Somewhat included in Loshin's "Structure" component of Consistency Also somewhat included in "Identifiability".)	Codd Integrity Constraings · Entity · Referential · Domain · Column
Clear Definition	Definition conformance	(included in TDWI's Usability dimension as "Availability and completeness of metadata")		Semantic	
	Validity (value, business rule, derivation)	(included in TDWI's definition of Integrity)	Validity	(included in Loshin's definition of Accuracy)	(included as component of author's definition of Column Integrity)
			Reasonableness	Reasonableness	Believability
Format Easy To Interpret	Presentation Standardization	Usability			

Table 2. Conformed Dimensions of Data Quality	
Data Quality Dimension	**Underlying Concepts**
Accuracy (Factualness)	Agree with Real-world, Match to Agreed Source
Consistency	Free of Conflict with Other Source, Consistency in Representation
Precision	Precision of Data Value (number of decimal places and rounding)
Timeliness	Time Expectation for Availability, Concurrence of Distributed Data
Accessibility	Ease of Attainability of Data, Need for Access Control
Completeness	Row Population, Column Population
Validity	Values in Specified Range of Valid Values, Values Conform to Business Rule, Values Conform to Other Attribute Types & Format
Integrity	Referential Integrity (primary-key/foreign-key match), Unique Identifier of Entity, Cardinality
Representation	Easy to Read & Interpret, Presentation Language, Media Appropriate, Complete & Available Metadata, Includes Measurement Units
Lineage	Source Documentation, Segment Documentation, Target Documentation, End-to-End Documentation