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

ICIQ2017 – OCTOBER 6TH

DAN MYERS & BRIAN BLAKE

Presentation Outline

- Background
  - Information Quality-Privacy-Trust Research Framework Matrix
  - Conformed Dimensions of Data Quality
- Rationale & Purpose
- Comparative Findings
- Application Findings
- Discussion
  - Subjective versus Objectives Dimensions
  - Defining Subjective Dimensions of Data Quality
- Conclusions
- Questions

Background

INFORMATION QUALITY-PRIVACY-TRUST RESEARCH FRAMEWORK MATRIX

Information Quality-Privacy-Trust Research Framework Matrix

- This research hypothesizes that:
  - H1: The multi-faceted dimensions, aspects, and properties of trust, privacy, and information quality can be effectively overlaid within a series of related matrices.
  - H2: An understanding of intersections of these sub-aspects lends itself to a broader understanding of the relationship of these concepts.
  - H3: An understanding of intersections of these sub-aspects lends itself to specific target areas for future research.

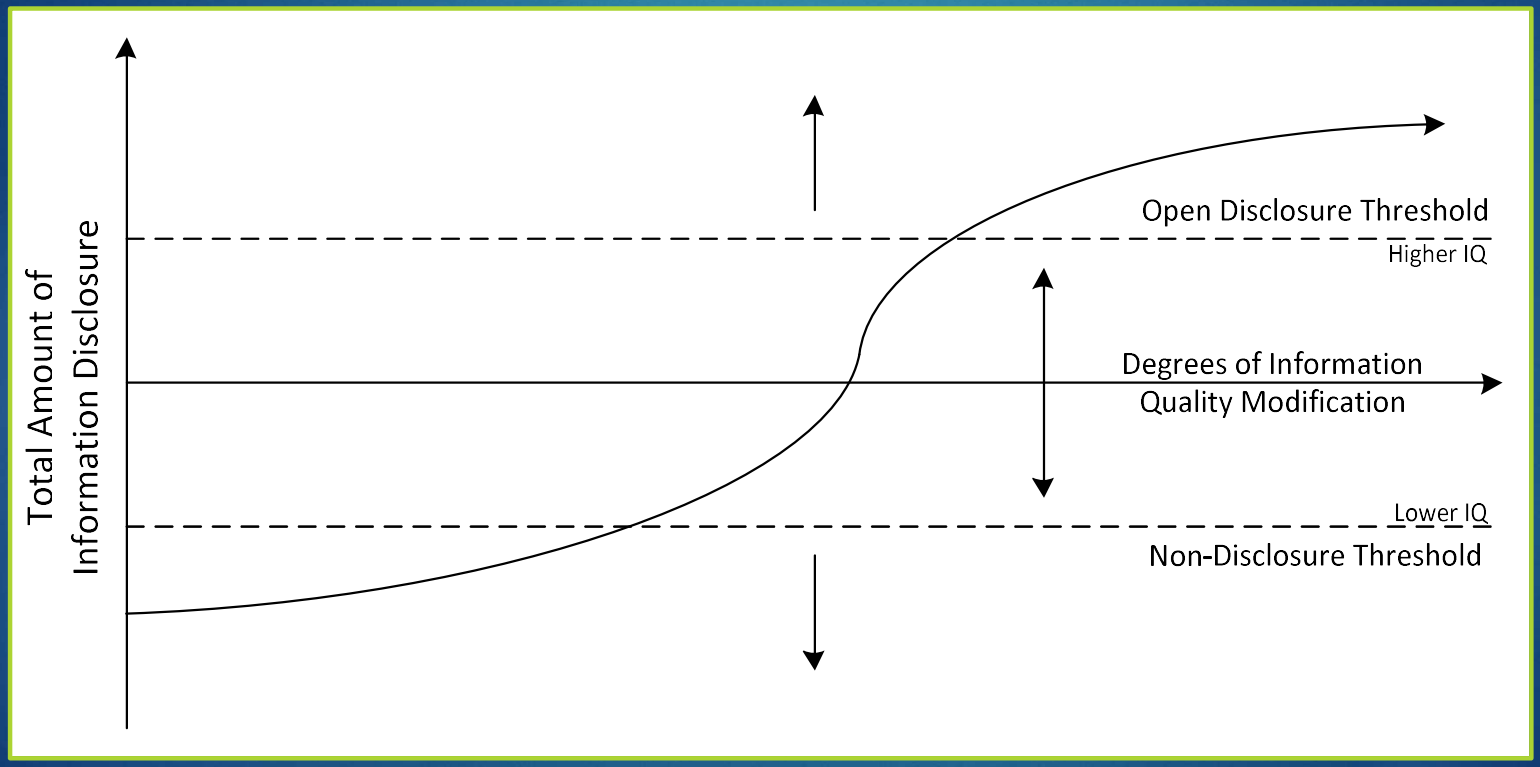
Initial Online Social Network Matrix

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

Initial IQ-Privacy Matrix

	Types of Data Privacy Issues					
	Information Processing					
	Aggregation	Identification	Insecurity	Secondary use	Exclusion	
Information Quality Dimensions	Accuracy Appropriate Amount Relevancy Believability Timeliness	Accuracy Believability Reputation	Security Accessibility	Appropriate Amount Accessibility Security Relevancy Accuracy	Security Accessibility Understandability Interpretability Timeliness	
	Information Dissemination					
	Breach of Confidenciality	Disclosure	Exposure	Increased Accessibility	Appropriation	Distortion
Information Quality Dimensions	Reputation Accuracy Believability Accessibility	Reputation Believability Accuracy Accessibility Appropriate Amount Relevancy	Reputation Believability Accuracy Accessibility Appropriate Amount	Accessibility Security Appropriate Amount	Security Reputation Believability Accuracy	Reputation Believability Accuracy Accessibility
	Invasions					
	Intrustion	Decisional Interference				
Information Quality Dimensions	Security Accessibility Appropriate Amount	Security Accessibility Appropriate Amount				

Information Quality Modification Concept



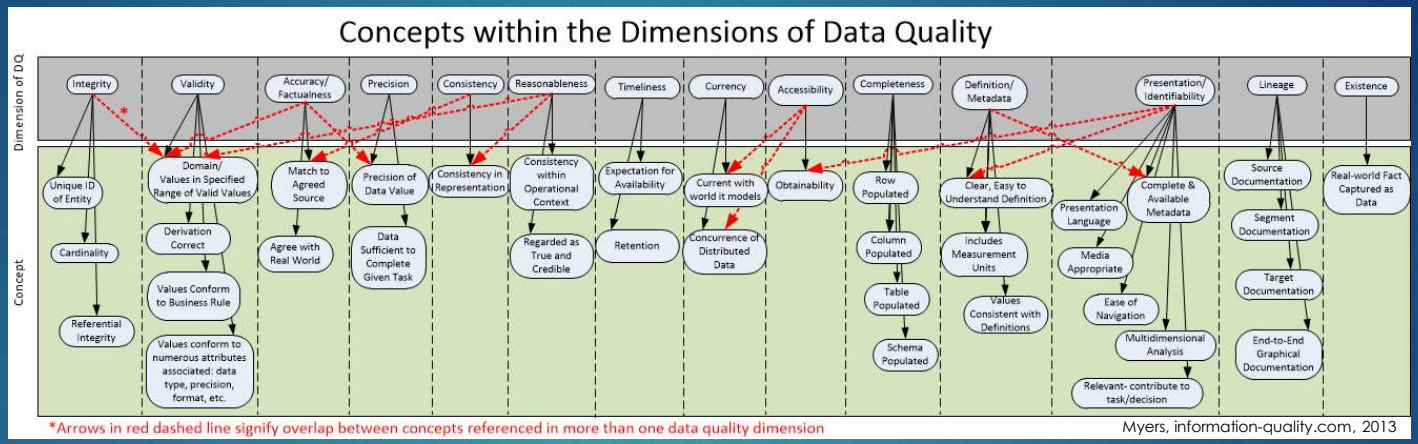
Relationship between IQ-Privacy-Trust Research and CDDQ

- Because the CDDQ provides additional detail and explanation than the original work by Pipinio, Lee and Wang, it may provide a better framework for identifying underlying concepts for measurement needed in a structural equation model developed as part of the IQ-Privacy-Trust research.

Background

CONFORMED DIMENSIONS OF DATA QUALITY

Overview of the Conformed Dimensions of Data Quality (CDDQ)



Rhetorical Questions:

- Why isn't there greater agreement between authors regarding the definitions of the dimensions of data quality?
- Can't we find a standardized set consistently used by researchers, educators, and practitioners?

These questions led to the prior research in 2013 and the formation of a "Conformed" set of dimensions.

2013 Publication

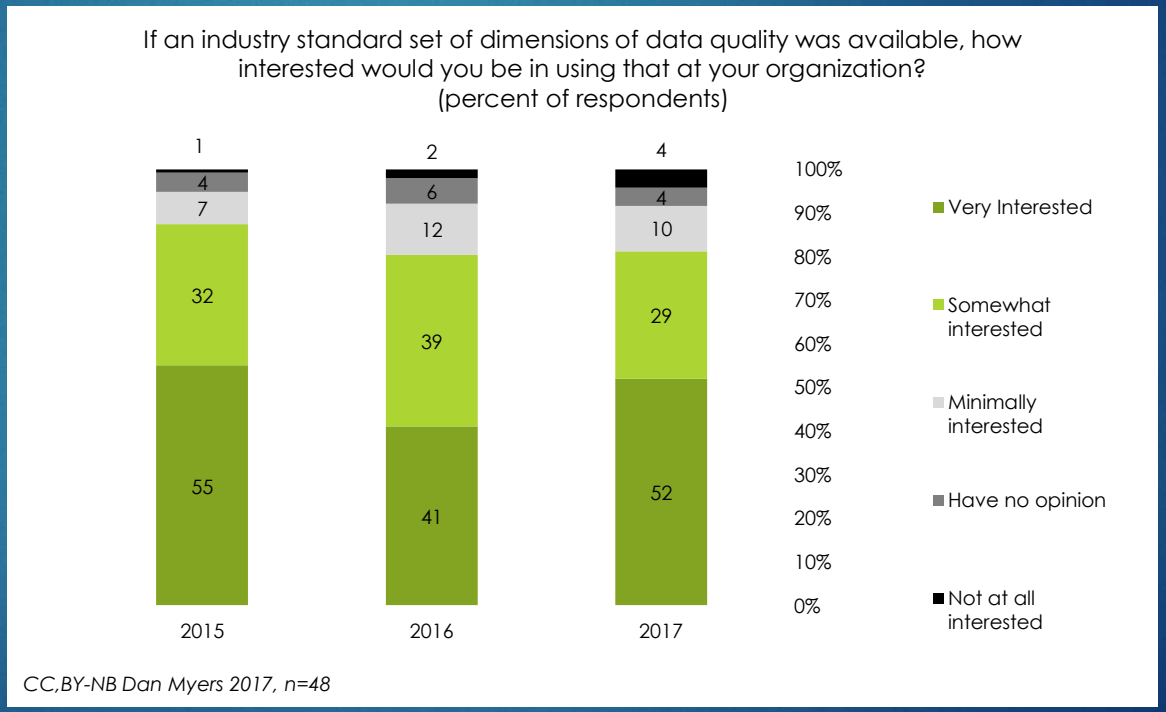
CDDQ History

- 2013
  - Publication of Series of Articles in Information-Management.com, Comparing Six Authors/Organizations' Definitions of Dimensions of Data Quality
- 2015
  - Initial Survey on Dimensions of Data Quality Conducted
- 2016
  - Roll-out of ConformedDimensionsOfDataQuality.com Website
  - 2nd Annual Survey Validated Most of 2015 Findings
- 2017
  - 3rd Annual Survey Refined and Focus on Latent Industries

2013 Publication

Annual Survey

Annual Dimensions of Data Quality Survey- based on CDDQ



CC,BY-NB Dan Myers 2017, n=48

2017 Survey Report

Recorded Webinar

Rationale & Purpose

OF THE CONFORMED DIMENSIONS OF DATA QUALITY

Rationale & Purpose for the CDDQ

- **Communication-**
  - Provide complete and common language to communicate DQ requirement and findings
- **Standardization-**
  - Enables efficiency through faster implementation times based on decreased argument between implementation team members (local); and discourages repetitive philosophical arguments on the same topic (global)
  - Repeatability offered due to standardization enables comparison and benchmarking
- **Measurement- if it isn't measured, it can't be managed**
  - Simplified due to common vocabulary and ease of reuse across systems and tools
  - Provides framework to define more detailed measurements associated with sub-concepts
- **Teaching-**
  - Provides a solid framework for teaching- avoiding labor intensive comparison and rationalization of differing terms and definitions

Rationale & Purpose

- If the dimensions are created with the purpose of 'communicating' the characteristics of data, then it is preferred that we use a single set that doesn't contradict or have overlapping definitions.
- Arguing about what should be in a set of enterprise, or even department level DQ dimensions, wastes time and confuses people who are beginning to learn about DQ.
- With a standard set of dimensions, organizations can skip over the first wave of arguments and can begin using the terminology and concepts to measure data quality from day one.

Comparative Findings

CDDQ VS. LEE, PIPINO, FUNK & WANG DQ DIMENSIONS

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

- Schema Completeness
  - Pipino, Lee, and Wang include a metric called Schema Completeness, the Conformed Dimensions do not include this as an underlying concept for a few reasons
- Accuracy
  - Although the Conformed Dimensions don't include an underlying concept explicitly stating 'Free of Error' this is equivalent to the CDDQ underlying concept of "Agree with the Real-World".

To audience: Do we need these?

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

- Completeness
  - **Truncation-** This measures whether the value contains all characters of the correct value.
  - **Existence-** Existence identifies whether a real-life fact has been captured as data.
- Accuracy
  - **Match to Agreed Source-** Measure of agreement between data and the source of that data. This is used when the data represent intangible objects or transactions that can't be observed visually.
- Consistency
  - **Equivalence of Redundant or Distributed Data-** The measure of similarity with other sources of data that represent the same concept.
- Integrity
  - **Referential Integrity-** Referential integrity measures whether if when a value (foreign key) is used it must reference an existing key (primary key) in the parent table.
  - **Uniqueness-** Uniqueness measures whether each fact is uniquely represented.
  - **Cardinality-** Cardinality describes the relationship between one table to another, such as one-to-one, one-to-many, or many-to-many.

Concepts Cont'd

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

Concepts Cont'd

- Currency
  - **Current with World it Models-** Data is current if it reflects the present state of the concept it models.
- Timeliness
  - **Time Expectation for Availability-** The measure of time between when data is expected versus made available.
  - **Manual Float-** Manual float is a measure of the time from when an observation is made to the point it is recorded in electronic format.
  - **Electronic Float-** Electronic float is a measure of the time from when data is captured in an electronic format until it is accessed by a person.[1]

1. The Electronic Float underlying concept was added to the CDDQ in release 3.4.

Application Findings

INFORMATION QUALITY-PRIVACY-TRUST RESEARCH FRAMEWORK MATRIX

Non-Conformed and Conformed Dimensions in Framework Matrix

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

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

Subjective and System Needs

- The subjective dimensions of Believability, Objectivity, Relevancy, and Reputation are excluded from the Conformed Dimensions by CDDQ Principle #1
- The system related dimensions of Ease of Manipulation, and possibly Security, are excluded from the Conformed Dimensions by CDDQ Principle #2.

Subjective versus Objectives Dimensions

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

Discussion

DEFINING SUBJECTIVE DIMENSIONS OF DATA QUALITY

Defining Subjective Dimensions of Data Quality

- Comparison of definitions of Believability
- Reviewed existing research on subjective dimensions

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

Proposed Definition of Believability Dimension

To audience:
All or nothing?

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

- Therefore if we create a composite definition (function of multiple variables) which is objective, we can use this definition, above, with associated underlying concepts.
- Note that concept 6 (prior slide), "Likely/Possible" was removed from the proposed definition due to its subjective and abstract nature that is difficult to quantify in a standardized way.

Conclusions

LIMITATIONS, FUTURE RESEARCH, AND FINAL THOUGHTS

Limitations

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

Future Research

- Regarding the Information Quality-Privacy-Trust research framework, a validation survey has been developed and implementation for select professionals and topic experts is pending.
- For the next phase of this research, a structural equation model for understanding the trade-offs and influences between data privacy, trust, and information quality in online social networks is being developed.
- Believability has been considered, but other typically subjective measures may need to be defined in terms of the CDDQ over time.
- The question regarding if and how system related dimensions should be approached by the CDDQ needs to be addressed.
- Additional published quality dimension frameworks will be evaluated and user feedback regarding the current CDDQ will be incorporated.

Final Thoughts

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

Questions

Author Contacts

DAN MYERS - DAN@DQMATTERS.COM

BRIAN P. BLAKE - BPBLAKE@UALR.EDU

Conformed Dimension (11)	Conformed Dimension Definition	Underlying Concepts	Non Standard Terminology for Dimension
Completeness	Completeness measures the degree of population of data values in a data set.	Record Population, Attribute Population, Truncation, Existence	Fill Rate, Coverage, Usability, Scope
Accuracy	Accuracy measures the degree to which data factually represents its associated real-world object, event, concept or alternatively matches the agreed upon source(s).	Agree with Real-world, Match to Agreed Source	Consistency
Consistency	Consistency measures whether or not data is equivalent across systems or location of storage.	Equivalence of Redundant or Distributed Data, Format Consistency	Integrity, Concurrence, Coherence
Validity	Validity measures whether a value conforms to a preset standard.	Values in Specified Range, Values Conform to Business Rule, Domain of Predefined Values, Values Conform to Data Type, Values Conform to Format	Accuracy, Integrity, Reasonableness, Compliance
Timeliness	Timeliness is a measure of time between when data is expected versus made available.	Time Expectation for Availability, Manual Float	Currency, Lag Time, Latency, Information Float
Currency	Currency measures how quickly data reflects the real-world concept that it represents.	Current with World it Models	Timeliness
Integrity	Integrity measures the structural or relational quality of data sets.	Referential Integrity, Uniqueness, Cardinality	Validity, Duplication
Accessibility	Accessibility measures how easy it is to acquire data when needed, how long it is retained, and how access is controlled.	Ease of Obtaining Data, Access Control, Retention	Availability
Precision	Precision measures the number of decimal places and rounding of a data value or level of aggregation.	Precision of Data Value, Granularity	Coverage, Detail
Lineage	Lineage measures whether factual documentation exists about where data came from, how it was transformed, where it went and end-to-end graphical illustration.	Source Documentation, Segment Documentation, Target Documentation, End-to-End Graphical Documentation	
Representation	Representation measures ease of understanding data, consistency of presentation, appropriate media choice, and availability of documentation (metadata).	Easy to Read & Interpret, Presentation Language, Media Appropriate, Metadata Availability, Includes Measurement Units	Presentation

Conformed Dimension	Underlying Concepts	Definition of Underlying Concept
Completeness	Record Population	This measures whether a row is present in a data set (table).
	Attribute Population	This measures whether a value is present (not null) for an attribute (column).
	Truncation	This measures whether the value contains all characters of the correct value.
	Existence	Existence identifies whether a real-life fact has been captured as data.
Accuracy	Agree with Real-world	Degree that data factually represents its associated real-world object, event, or concept.
	Match to Agreed Source	Measure of agreement between data and the source of that data. This is used when the data represent intangible objects or transactions that can't be observed visually.
Consistency	Equivalence of Redundant or Distributed Data	The measure of similarity with other sources of data that represent the same concept.
	Format Consistency	This measures the conformity of format of the same data in different places.
	Logical Consistency	Logical consistency measures whether two attributes of related data are conceptually in agreement, even though they may not record the same characteristic of a fact.
Validity	Values in Specified Range	Values must be between some lower number and some higher number.
	Values Conform to Business Rule	Validity measures whether values adhere to some declarative formula.
	Domain of Predefined Values	This is a set of permitted values.
	Values Conform to Data Type	Validity measures whether values have a specific characteristic (e.g. Integer, Character, Boolean). Data types restrict what values can exist, the operations that can be use on it, and the way that the data is stored.
	Values Conform to Format	Validity measures whether the data are arranged or composed in a predefined way.
Timeliness	Time Expectation for Availability	The measure of time between when data is expected versus made available.
	Manual Float	Manual float is a measure of the time from when an observation is made to the point it is recorded in electronic format.
Currency	Current with World it Models	Data is current if it reflects the present state of the concept it models.

Conformed Dimension	Underlying Concepts	Definition of Underlying Concept
Integrity	Referential Integrity	Referential integrity measures whether if when a value (foreign key) is used it must reference an existing key (primary key) in the parent table.
	Uniqueness	Uniqueness measures whether each fact is uniquely represented.
	Cardinality	Cardinality describes the relationship between one data set and another, such as one-to-one, one-to-many, or many-to-many.
Accessibility	Ease of Obtaining Data	This measures how easy it is to obtain data.
	Access Control	Access control includes the identification of a person that wants to access data, authentication of their identity, review and approval to access required data, and lastly auditing the access of that data.
	Retention	Retention refers to the period of time that data is kept before being removed from a database through purge or archive processing.
Precision	Precision of Data Value	The measure of preciseness of numeric data using decimal places, rounding and truncation.
	Granularity	The detail or summary of data defines the granularity measured by the number of attributes used to represent a single concept.
Lineage	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.
Representation	Easy to Read & Interpret	Illustrations and charts should be self-explanatory and presented with appropriate labels, providing context.
	Presentation Language	Data that is represented well is simple but elegantly formed with good grammar and presented in a standard way.
	Media Appropriate	The appropriate media (e.g. Web-based, hardcopy, or audio…etc) are provided.
	Metadata Availability	Comprehensive descriptions and other information about the characteristics of the data are provided in plain language.
	Includes Measurement Units	Well represented data includes the scale of measurement, such as weight, height, distance…etc.

Version 3.4.1. http://dimensionsofdataquality.com/content/list-underlying-concepts