Streamlining Data Quality Efforts Using the Dimensions of Data Quality

-Dan Myers (dan@DQMatters.com)

6/14/2017

Professional Profile

Dan Myers is Principal Info Quality Educator at DQMatters- an eLearning organization focused on Information Quality training and consulting.

In previous roles Dan has managed business intelligence teams, and lead architecture reviews of data management (metadata, data quality…etc) tools and implemented associated governance programs. In his role at Farmers Insurance, he authored the Finance led data governance policies for integration/sourcing, metadata, 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.

Agenda

Introduction

- What are the Dimensions of Data Quality?
- Why do I need them?
- Where to use them in the SDLC?
- Brief history
- Which set do I use?
- (Dis)agreement in the Industry about Definition & Scope
- Reasons to Agree Upon a Cross-Industry Standard
- Conformed Dimensions
  - Dimensions Level
  - Underlying Concepts Level

What Are The Dimensions of Data Quality?

Definition: The Conformed Dimensions of Data Quality are categories used to characterize data and it's fitness for use.

Application: These can be applied in any industry to assess, measure, track and communicate information and data quality.

Accuracy

Accessibility

Lineage

Timeliness
Timeliness is a measure of time between when data is expected versus made available.

Currency
Currency measures how quickly data reflects the real-world concept that it represents.

Validity
Validity measures whether a value conforms to a preset standard.

Completeness
measures the degree of population of data values in a data set.

Integrity
Integrity measures the structural or relational quality of data sets.

Believability

Representation

Consistency

Precision

Why do I need the Dimensions of DQ?

[b]Need broader adoption than 38%

Because they add value:

a) Act as quick reference, checklist, and guide to quality standards[a]

b) Can be used as framework to segment DQ efforts across a business unit, or even a company

c) Enable people to communicate current and desired state of data

d) Reuse of existing categories and definitions enables faster implementation times

e) Match dimensions against a business need and prioritize which assessments to complete first[1]

f) Understand what you will (and will not) get from assessing each dimension.[1]

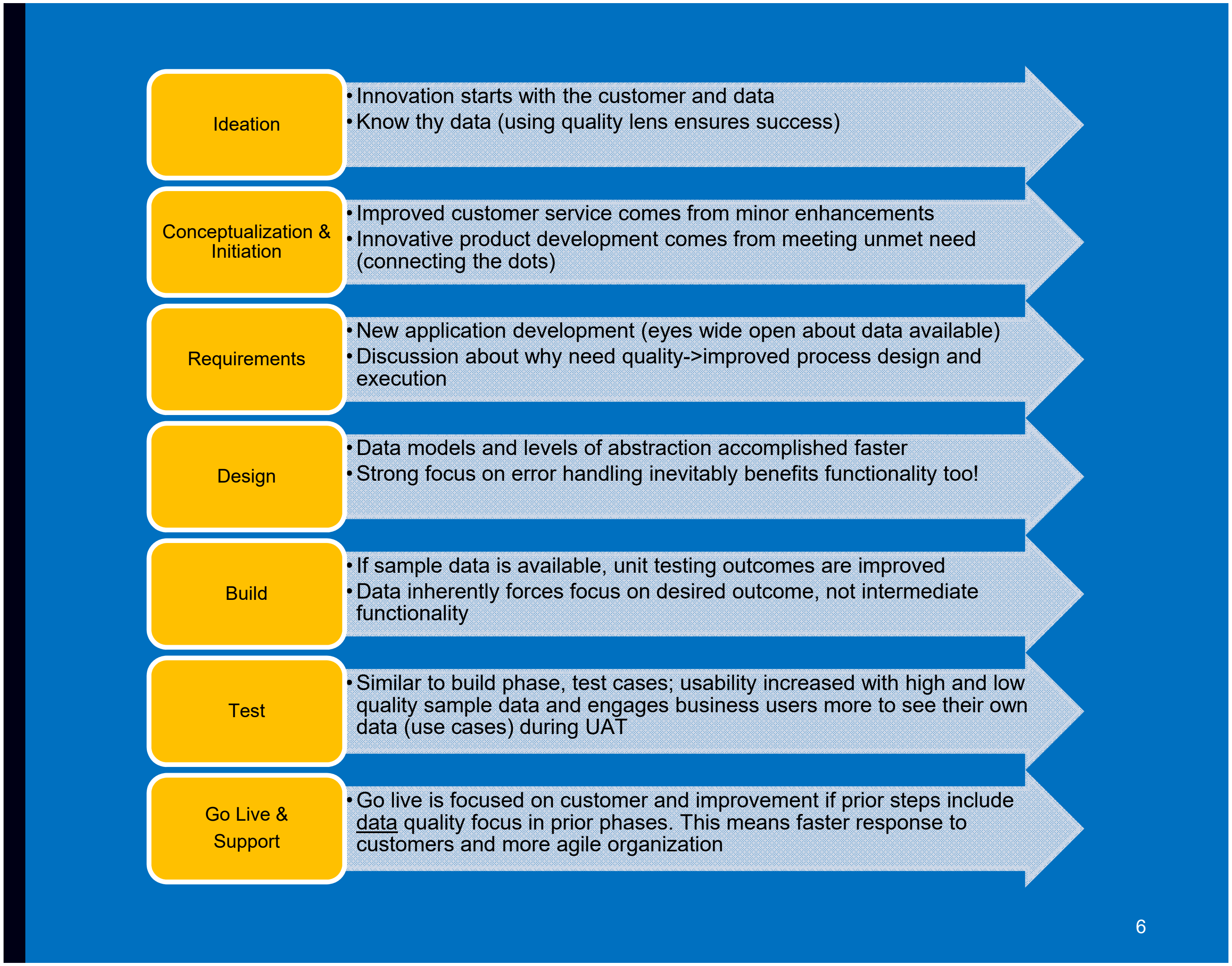
References: 1. McGilvray, 2008 p. 30-31

Where are they used:

- To define DQ measures on scorecards, dashboards
- In conversation
- Embedded in instructions or on forms
- Included in Service Level Agreements
- Throughout the SDLC (next slide)

Where to use them in the SDLC?

Streamline your lifecycle



Ideation

- Innovation starts with the customer and data
- Know thy data (using quality lens ensures success)

Conceptualization & Initiation

- Improved customer service comes from minor enhancements
- Innovative product development comes from meeting unmet need (connecting the dots)

Requirements

- New application development (eyes wide open about data available)
- Discussion about why need quality->improved process design and execution

Design

- Data models and levels of abstraction accomplished faster
- Strong focus on error handling inevitably benefits functionality too!

Build

- If sample data is available, unit testing outcomes are improved
- Data inherently forces focus on desired outcome, not intermediate functionality

Test

- Similar to build phase, test cases; usability increased with high and low quality sample data and engages business users more to see their own data (use cases) during UAT

Go Live & Support

- Go live is focused on customer and improvement if prior steps include data quality focus in prior phases. This means faster response to customers and more agile organization

Brief history of the dimensions

MIT Work:

- Richard Wang and Diane Strong (1996), “Beyond Accuracy: What Data Quality Means to Data Consumers”
- Leo Pipino, Yang Lee, and Richard Wang (2002), “Data Quality Assessment”
- Yang Lee, Leo Pipino, Richard Wang, James Funk (2006), Journey to Data Quality

Pioneer Practitioners:

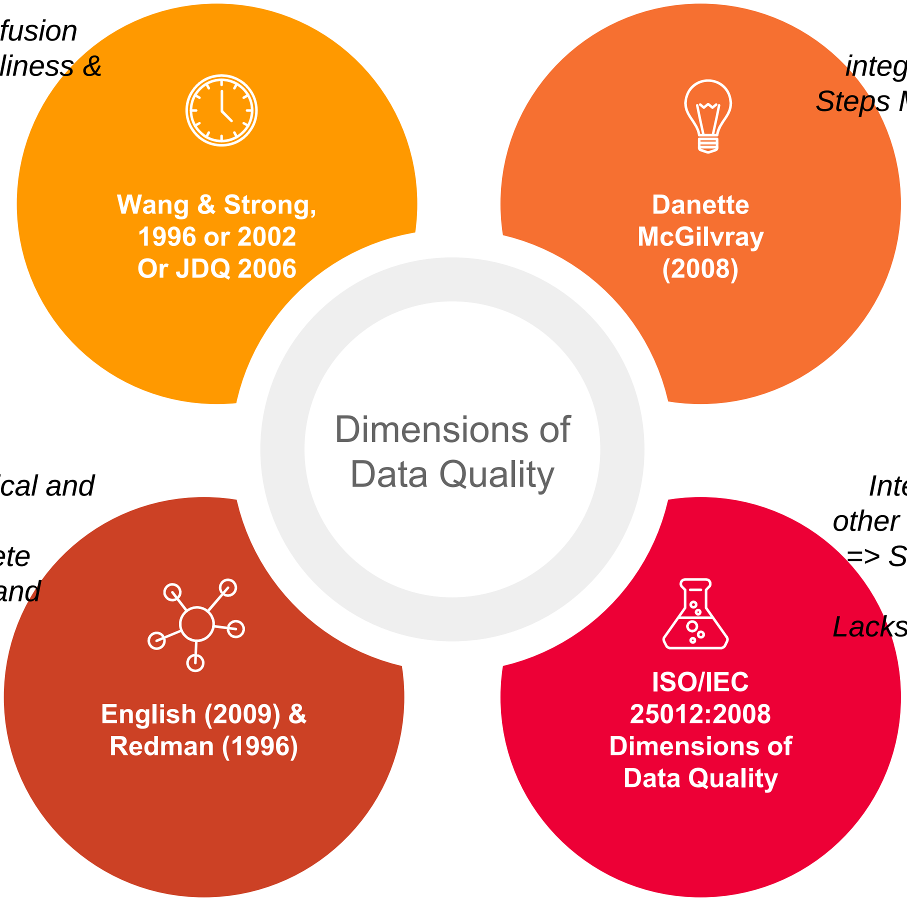
- Tom Redman (1996), Data quality for the information age; (2001) Data quality the field guide
- Larry English (1999), Improving data warehouse and business information quality: Methods for reducing costs and increasing profits

Neo-Practitioners:

- David Loshin (2011), The practitioner's guide to data quality improvement
- Danette McGilvray (2008), Executing data quality projects: Ten steps to quality data and trusted information
- Laura Sebastian-Coleman (2011), Measuring data quality for ongoing improvement: a data quality assessment framework

All source information available at: http://dimensionsofdataquality.com/research and in some cases links to the actual research (if freely available on the Internet)

Which set do I use?



Dimensions of Data Quality

Wang & Strong, 1996 or 2002 Or JDQ 2006

Solid, but Confusion between Timeliness & Currency

Danette McGilvray (2008)

Practical and integrated with Ten Steps Methodology™

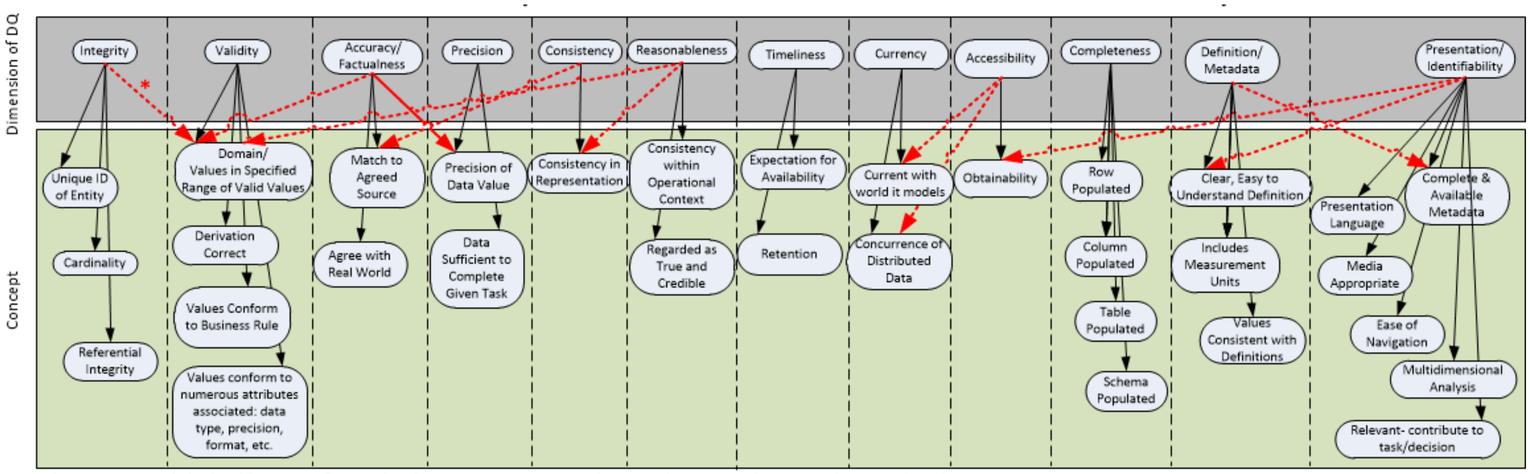
English (2009) & Redman (1996)

Strong technical and logical basis. Lacks complete descriptions and hierarchy

ISO/IEC 25012:2008 Dimensions of Data Quality

Interconnected to other ISO standards => Strong Systems Focus. Lacks hierarchy and overly context sensitive

(Dis)agreement in the Industry about Definition & Scope


Dimension of DQ
Concept
Integrity
Validity
Accuracy/ Factualness
Precision
Consistency
Reasonableness
Timeliness
Currency
Accessibility
Completeness
Definition/ Metadata
Presentation/ Identifiability
*
Unique ID of Entity
Cardinality
Referential Integrity
Domain/ Values in Specified Range of Valid Values
Derivation Correct
Values Conform to Business Rule
Values conform to numerous attributes associated: data type, precision, format, etc.
Match to Agreed Source
Agree with Real World
Precision of Data Value
Data Sufficient to Complete Given Task
Consistency in Representation
Consistency within Operational Context
Regarded as True and Credible
Expectation for Availability
Retention
Current with world it models
Concurrence of Distributed Data
Obtainability
Row Populated
Column Populated
Table Populated
Schema Populated
Clear, Easy to Understand Definition
Includes Measurement Units
Values Consistent with Definitions
Presentation Language
Media Appropriate
Complete & Available Metadata
Ease of Navigation
Multidimensional Analysis
Relevant- contribute to task/decision

http://dqm.mx /imdm2013

So this confusion led me to write a series of articles for Information-Management.com, titled “The Value of Using the Dimensions of Data Quality” (2013)

- Compares six author’s definitions of the dimensions of data quality
- Proposes began the work to define a Conformed set of Dimensions of Data Quality

Reasons to Agree Upon a Cross-Industry Standard

- **Communication-**
  - Provide language to communicate DQ requirements
- **Efficiency-**
  - Enables faster implementation times based on decreased argument between implementation team members (local)
  - Discourages repetitive philosophical arguments on the same topic (global)
- **Measurement-** *if it isn't measured it can't be managed*
  - Consistency between organizations enables comparisons used to benchmark and improve
  - Provides framework to define more detailed measurements associated with sub-concepts
- **Teaching-** Provides a solid framework for teaching

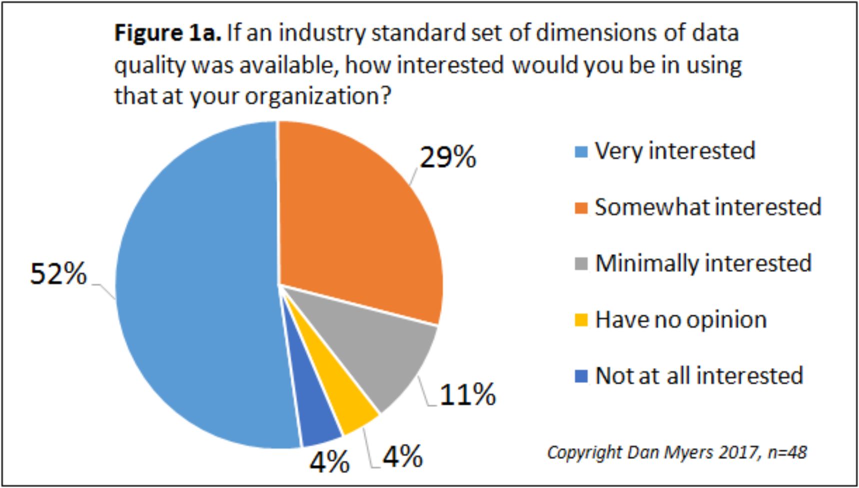


Figure 1a. If an industry standard set of dimensions of data quality was available, how interested would you be in using that at your organization?

Copyright Dan Myers 2017, n=48

- Survey conducted 3 years in a row Shows general consistency of
  - 45-50% Very Interested Respondents
  - 32-35% Somewhat Interested Respondents

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.

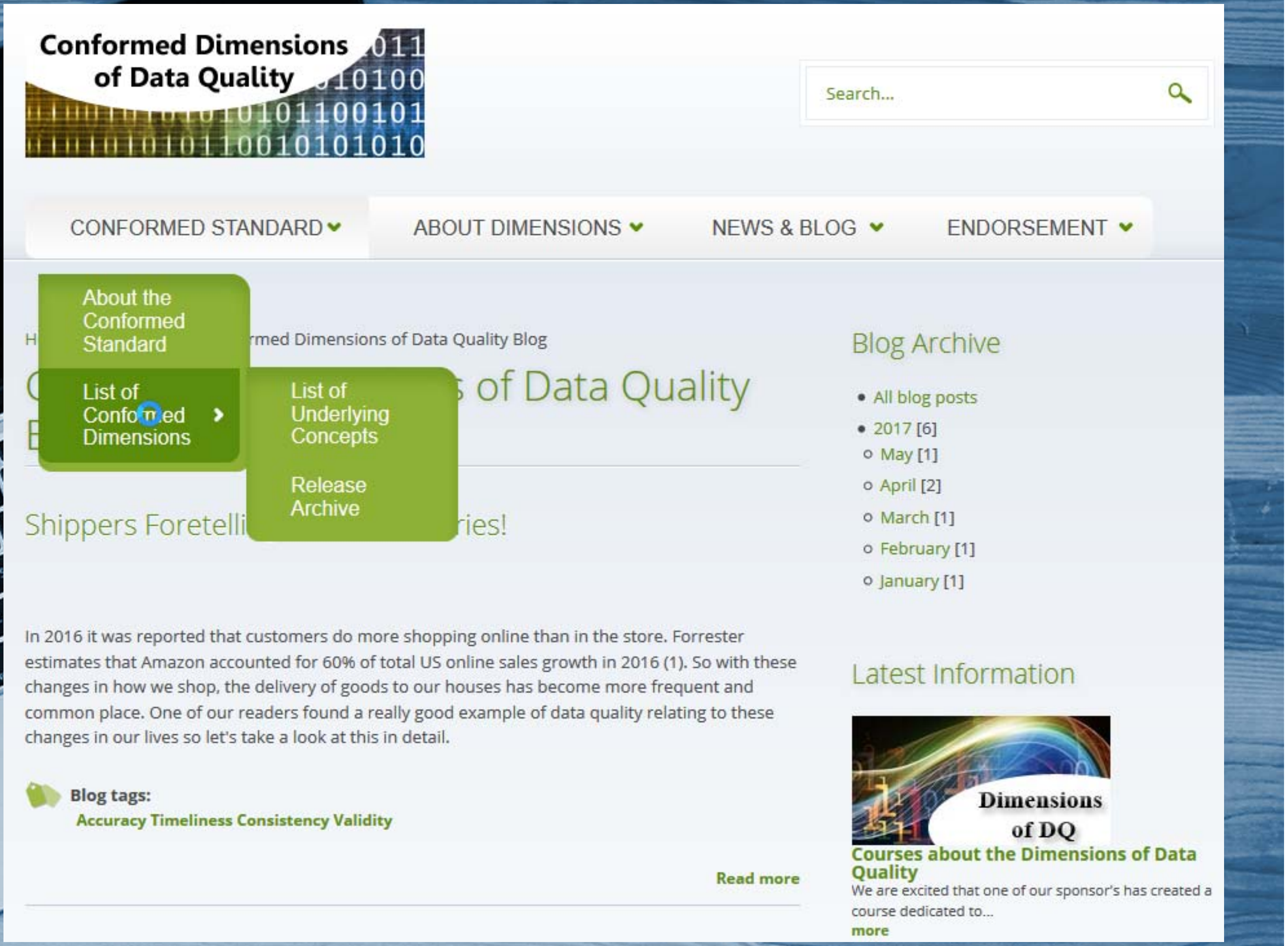
Version 3.4.1. http://dimensionsofdataquality.com/content/list-underlying-concepts

Website

Site:
http:DimensionsOfDataQuality.com

Blog Site:
http:DimensionsOfDataQuality.com/blog

Blog Signup:
http://eepurl.com/cEgkJj


Conformed Dimensions of Data Quality
Search...
CONFORMED STANDARD
ABOUT DIMENSIONS
NEWS & BLOG
ENDORSEMENT
About the Conformed Standard
List of Conformed Dimensions
List of Underlying Concepts
Release Archive
Dimensions of DQ

H rmed Dimensions of Data Quality Blog

C s of Data Quality
E

Shippers Foretelli ries!

In 2016 it was reported that customers do more shopping online than in the store. Forrester estimates that Amazon accounted for 60% of total US online sales growth in 2016 (1). So with these changes in how we shop, the delivery of goods to our houses has become more frequent and common place. One of our readers found a really good example of data quality relating to these changes in our lives so let's take a look at this in detail.

Blog tags:
Accuracy Timeliness Consistency Validity

Read more

Blog Archive

- All blog posts
- 2017 [6]
  - May [1]
  - April [2]
  - March [1]
  - February [1]
  - January [1]

Latest Information

Courses about the Dimensions of Data Quality
We are excited that one of our sponsor's has created a course dedicated to...
more

Q&A