# We are READY for a Standard Set of Dimensions of Data Quality

A Whitepaper

Sponsored by

DQMatters
Data Quality eLearning

4/2015

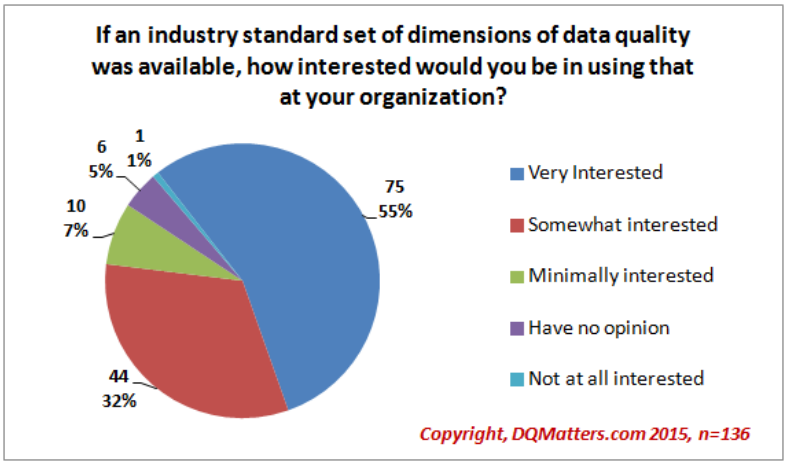
# Executive Summary

## Purpose

A number of academics, consultants, authors and organizations have identified the "Dimensions of Data Quality", as a mechanism to measure data quality. The problem is that there isn't a single cross-industry agreed upon standard definition of the dimensions. All accountants know what belongs on a Balance Sheet, and similarly data professionals should agree upon a core set of data characteristics that communicate fitness for use. The purpose of this survey was to measure how frequently different dimensions of data quality are used and whether data management practitioners would adopt a standard if one existed.

This was a web-based survey distributed via LinkedIn, Twitter, Email, and a paper-based survey version given to Enterprise Data World tutorial attendees. There were 136 complete responses to the survey.

## Summary of Findings




- 35% of respondent's organizations classify data related defects using the dimensions of DQ on an ongoing basis.
- A large proportion (87%) of the respondents are interested in using an Industry Standard at their organizations (55% are very interested and 32% are somewhat interested).
- We acknowledge that there is a level of self-selection bias within the survey, however these numbers still represent a strong market need for a cross-industry agreed upon standard set of dimensions.
- 23% of organizations have one formally defined and governed set of dimensions used for categorizing DQ issues.
- The top 6 dimensions cited were: **Accuracy, Completeness, Consistency, Validity, Timeliness, Integrity**

## Next Steps

Based on the higher than expected use of the dimensions of data quality, and survey-validated desire to have a cross-industry standard set of dimensions of data quality, a consortium of organizations is forming a website to propose a possible ***Conformed Dimensions of Data Quality*** standard. Please take the time to explore that site and our sponsor's site. Volunteer opportunities and organizational sponsorship is encouraged via the website.

**Proposed Standard:**
Conformed Dimensions of Data Quality
http://dimensionsofdataquality.com

**Sponsor Website:**
Data Quality Matters
http://DQMatters.com

# Introduction

The dimensions of data quality have been around for a long time[i] and many areas of the information and data quality domain have matured, but unlike other professions where we see standards formed over time, we haven't seen a standard evolve for the dimensions of data quality. For the reasons outline in this white paper, it is time to form a standard. **The purpose of this survey was to measure how frequently different dimensions of data quality are used and whether data management practitioners would adopt a standard if one existed.**

## Value of Using the Dimensions of Data Quality in General

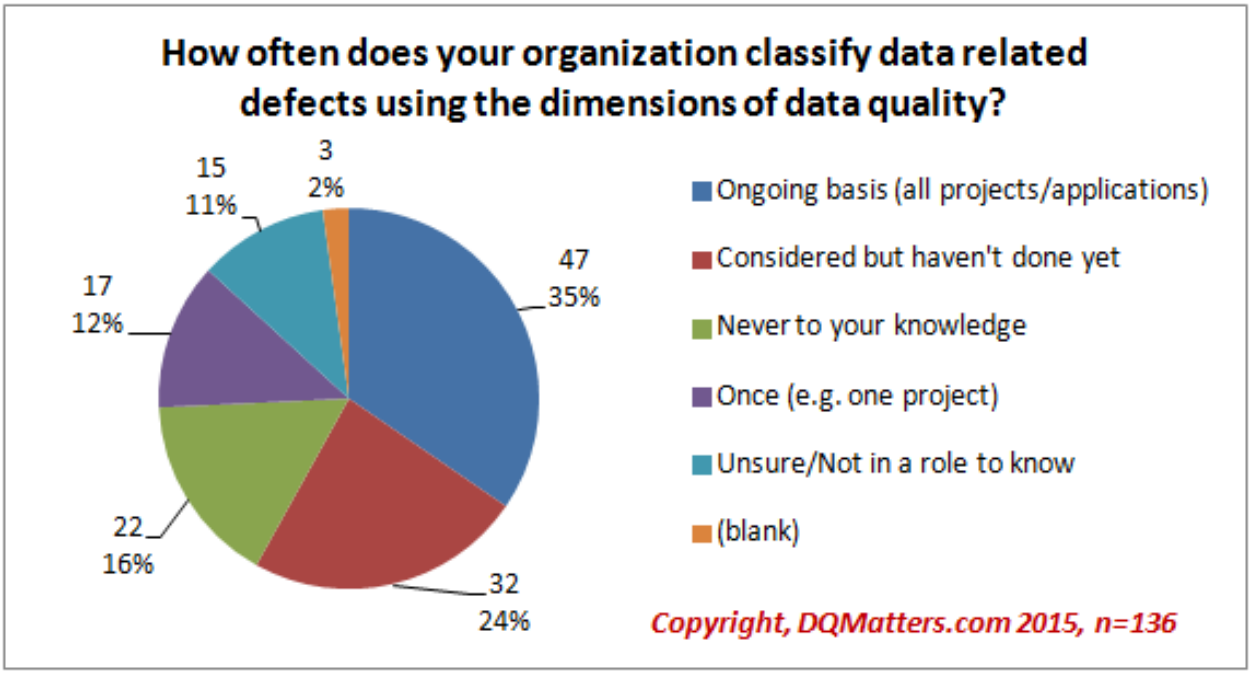
- Act as quick reference, checklist, and guide to quality standards
- Can be used as framework to structure DQ efforts across a business unit, or even a company Enable people to communicate current and desired state of data
- Reuse of existing categories and definitions enables faster implementation times
- Understand what your organization will (and will not) gain by assessing each dimension[ii]
- Match dimensions against a business need and prioritize which assessments to complete first

## Need for a Conformed Standard with Agreed Upon Concepts and Definitions

In a series of articles, addressing the lack of agreement on the Dimensions of Data Quality in Information-Management.com in 2013, Dan Myers proposed a conceptual list of dimensions that agrees with most authors' definitions. Based on that work and discussion with data management industry leaders, we identified the following areas of misunderstanding and disagreement. Generally speaking the survey results affirmed this observation.

| Conformed Data Quality Dimension | Examples of Use of Non-Conformed Terminology | Disagreement about name of dimension |
|---|---|---|
| Accuracy | | Consistency |
| Completeness | Fill Rate, Coverage | Usability |
| Consistency | Concurrence, Coherence | Integrity |
| Validity | | Accuracy, Integrity, Reasonableness |
| Timeliness | | Currency |
| Integrity | Duplication | Validity |
| Accessibility | | Availability |
| Precision | | |
| Lineage | | |
| Currency | Data Decay | Timeliness, Accessibility |
| Representation | Presentation | |

## Usage of the Dimensions

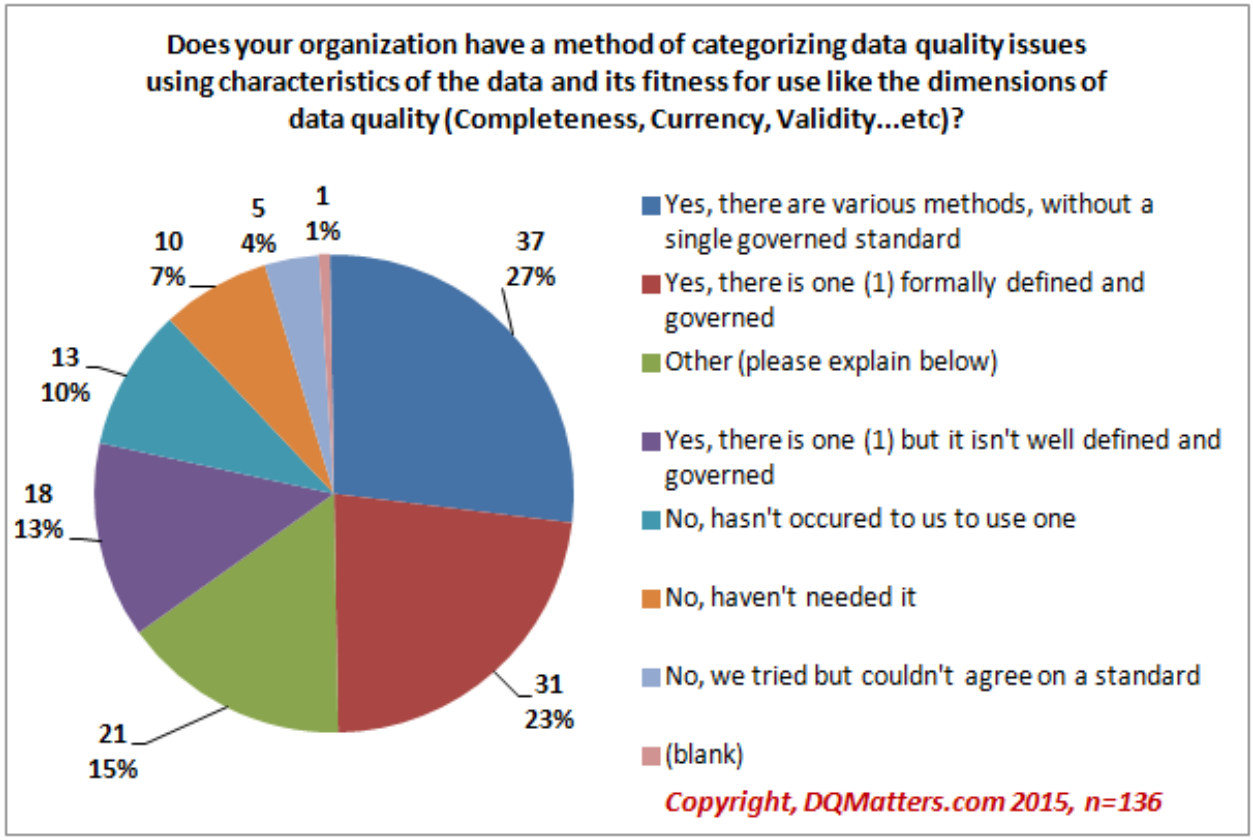



a

Logically the dimensions of data quality offer a lot, but we wanted to gain insight into how widely (percent of all organizations) and frequently they are used. As seen below, we were surprised by the number of organizations that use some classification of data defects using the dimensions of data quality. We were also surprised that such a high proportion of the respondents said that they use them on an ongoing basis (35%).

Two of the groups of respondents who use the dimensions (*Ongoing* & *Once*) made up nearly half (47%) and, if you include those considering future use of the dimensions, the number is much larger (71%). The over-all use of the dimensions of data quality is relatively high so the other responses in the survey regarding how the dimensions are used and the respondent's opinions about the definitions of each dimension were credible.

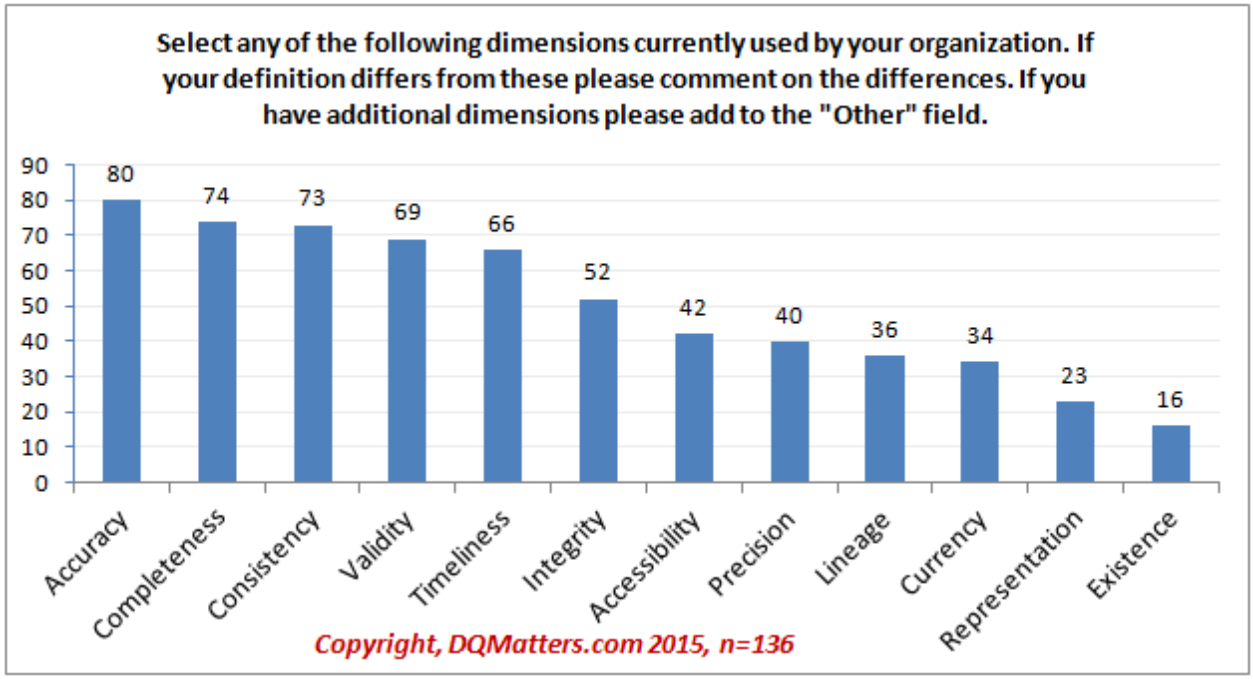
## Governance of Dimensions Used




Of the 47 respondents who stated that they use the dimensions of data quality in an ongoing basis (prior question), most had a formally defined and governed set of dimensions (45%). Some had various methods (30%) or had one that wasn't well defined or governed (13%).

Within those respondents who said they do use the dimensions of data quality, the largest group, said that as we'd expect, there are various methods, without a single governed standard. We believe that these respondents who either have various ungoverned methods (27%), one but poorly defined (13%), or no standard because they can't agree on it (4%), would benefit from a single cross industry standard set of dimensions of data quality. That is 44% in total.

## Popularity of Each Dimension

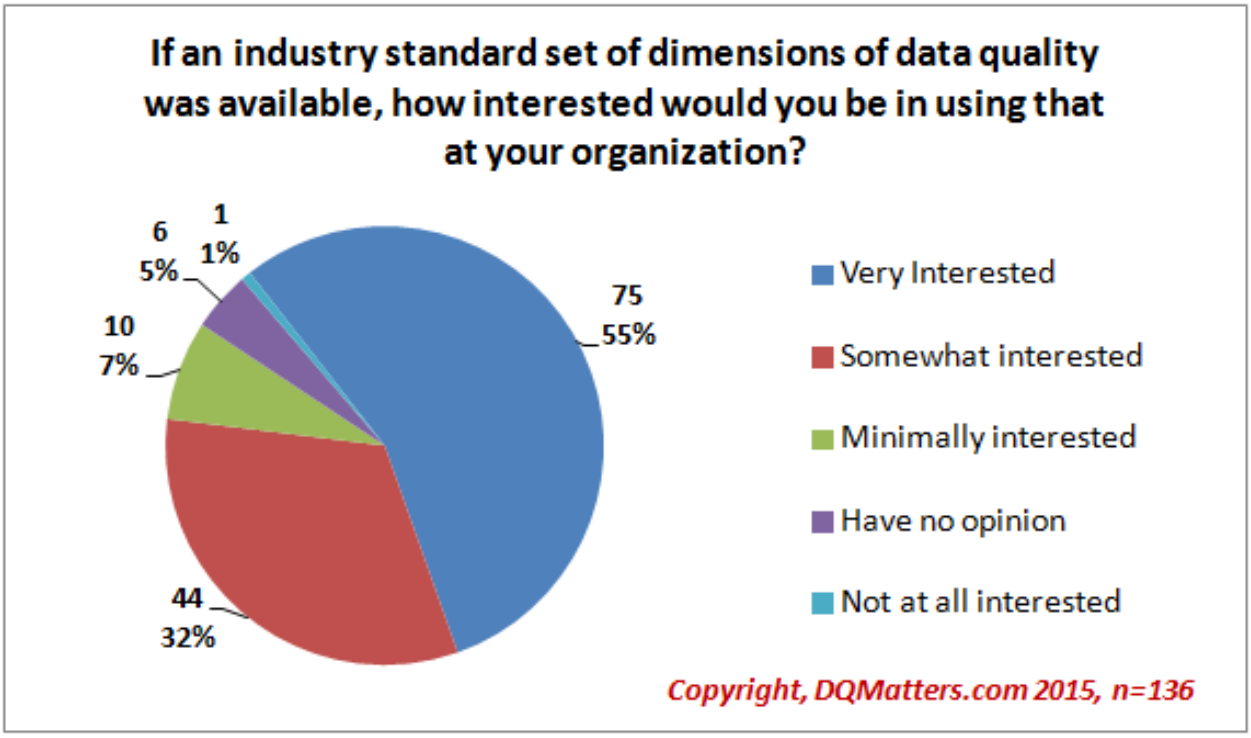



The primary question of the survey was geared to get feedback regarding how organizations **define** each of the dimensions of data quality. We did this by providing a list of dimensions with detailed descriptions and asking respondents whether they use that dimension and if so, how their definitions differs from ours. The list that we provided is a proposed set of Conformed Dimensions of Data Quality that we believe can take the place of all of the fragmented and undefined lists of dimensions used today.

Due to the complexity of unpacking and interpreting the respondent's answers about how their definitions differ from our proposed standard, we will publish articles on that topic at a later time on dimensionsofdataquality.com. As likely expected, the familiar five dimensions that are often discussed in data quality rose to the top: Accuracy, Completeness, Consistency, Validity and Timeliness.

- There is some confusion in the data management industry about whether Precision belongs within Accuracy or not, but based on the response provided in the survey we can see the significant differences in frequency of use of Accuracy and Precision, so we believe that this lends credibility to separating them.
- Similarly, we see that when given the opportunity, DQ professionals have distinctly different levels of use of the two time related dimensions, Timeliness and Currency. This also supports the decision to separate them based on differing underlying concepts.

## Conclusion




One of the goals of the survey was to identify the need and likely demand for a standard set of dimensions of data quality that are robustly defined and universally agreed upon. Two groups that answered positively- meaning they would be interested in using a standard- were well in the majority at 88%.

We believe that, although there is a self-selection bias (people experienced in the dimensions or who care more than average data management professionals about a standard) in the survey sample[1], the results prove that there is enough interest in a standard to actively pursue it. For this reason, our sponsor, DQMatters.com is funding the creation of a new website to support the pursuit of an open and freely available standard. Check it out for yourself at: dimensionsofdataquality.com.

[1] Sample meaning that the all of the responses of the survey make up only a sample of the complete population of organizations eligible to have taken the survey.

# Appendix

## General Survey Information

| | |
|---|---|
| **Count of Full Responses:** | 136 |
| **Dates Survey was Open:** | 3/10/2015 to 4/9/2015 |

## Research Methodology & Future Opportunities

- Because there is somewhat of a self-selection bias due to the fact that the people who opted to take the survey on "categories of data quality" are orientated to the topic and may even have been the ones to implement such dimensions at their organizations. Future surveys will need to control for this through documentation of respondent's role and other factors likely to bias.
- In addition to using the dimensions to classify defects, requirements gathering can leverage the dimensions to communicate desired levels of data quality at the beginning of the data life-cycle. Future surveys will also need to assess how often organizations are using the dimensions at other points of the Software Development Life-Cycle (SDLC).

## Source of Survey Responses

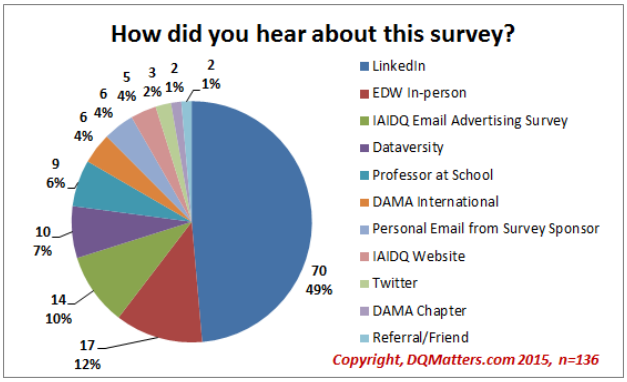
The survey was advertised in a number of online locations and offered in web-based format. Additionally, attendees of Dan Myers' 3 hour tutorial on this topic at Enterprise Data World were given the opportunity to take a paper-based survey in the class. The primary Web-survey respondents were referred by announcements in: Various LinkedIn groups (49%), IAIDQ E-mail (10%), Dataversity (7%), School professor (6%), DAMA International (4%). The in-person tutorial attendee responses composed an additional 12% of the responses. (See appendix, item #2 for additional detail).

## Industries Represented by Respondents

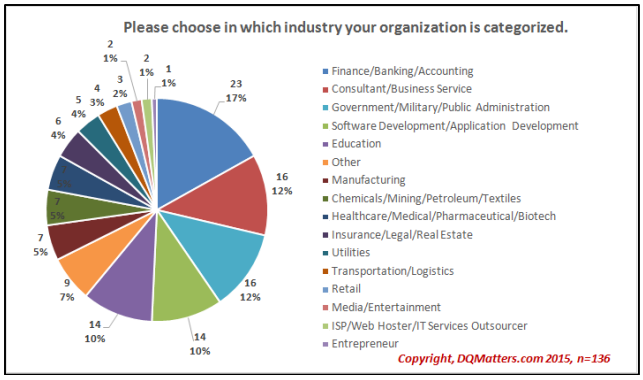
The top five industries represented by the responses were as follows: (See appendix, item #3 for additional detail).
17%, Finance/Banking/Accounting
12%, Consultant/Business Service
12%, Government/Military/Public Administration
10%, Software Development/Application Development
10%, Education

## Additional Questions Included in the Survey

How did you hear about this survey




Please choose in which industry your organization is categorized




---

END NOTES

[i] The earliest published work in this area that we are aware of was by Professors Richard Wang and Diane Strong in their 1996 paper titled Beyond Accuracy: What Data Quality Means to Data Consumers, http://courses.washington.edu/geog482/resource/14_Beyond_Accuracy.pdf.

[ii] Danette McGilvray, Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information, Morgan Kaufmann, 2008 p. 30-31