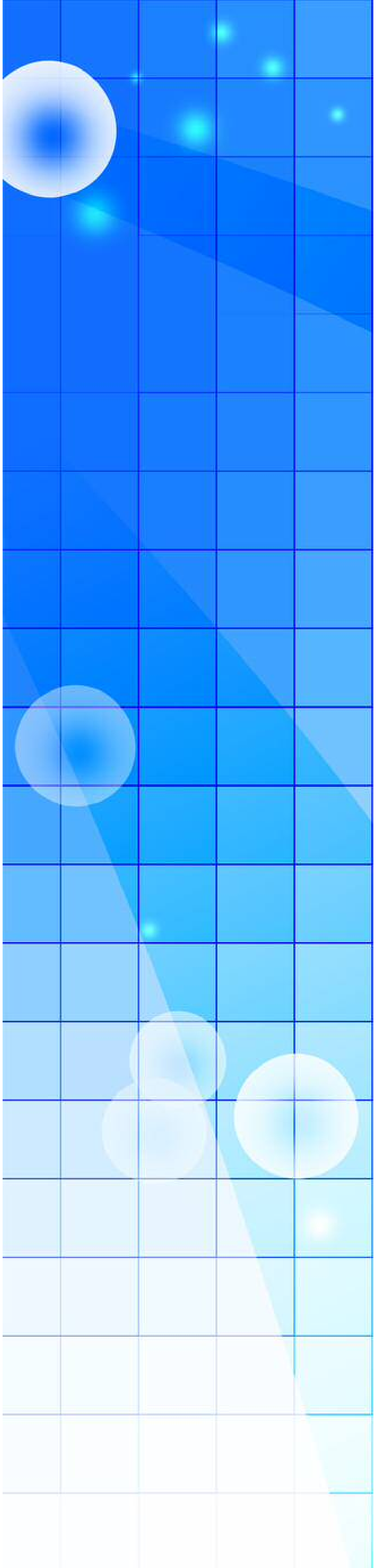
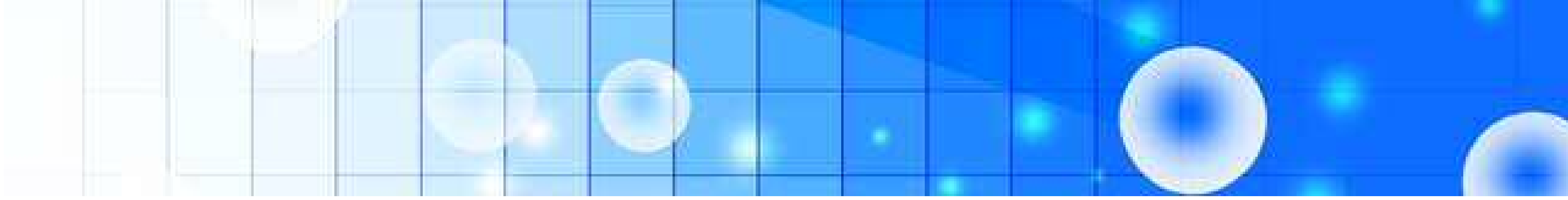


2016 Annual Report on the Dimensions of Data Quality

Year-two:
General Usage Improves but Confusion Remains

A Whitepaper
Sponsored by

DQMatters
Data Quality eLearning



Executive Summary

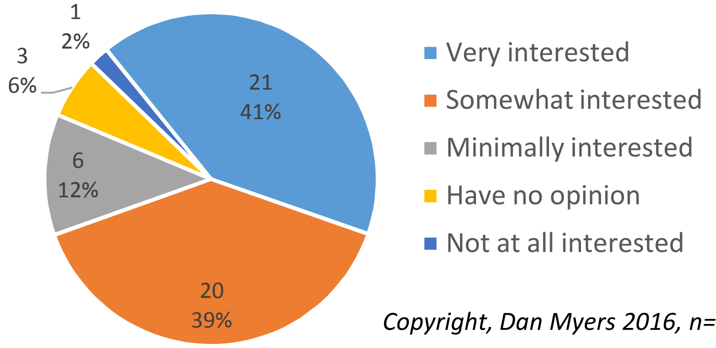
Purpose

The term “Dimensions of Data Quality” is nearly ubiquitous within data management, analytics and research communities, and yet a cross-industry standard has not been developed for how to properly communicate the characteristics of data quality. The purpose of this whitepaper is to measure organizational use of the dimensions of data quality and whether data management practitioners would adopt a standard version of the dimensions of data quality if one existed.

The content of this paper is based on a web-based survey distributed via LinkedIn, Twitter, Email, and in person at a session during Enterprise Data World 2016. There were 51 complete responses to the survey.

Summary of Findings

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 2016, n=51

- 80% of the respondents are interested in using an Industry Standard at their organizations (41% are very interested and 39% are somewhat interested). See figure 1a at right.
- 45% of respondent’s organizations classify data related defects using the dimensions of DQ on an ongoing basis. This number increased by 10% from our findings in 2015.
- 28% of organizations have a single formally defined and governed set of dimensions used for categorizing DQ issues.
- The top 6 most popularly used Dimensions were: **Completeness, Accuracy, Validity, Timeliness, Consistency, and Integrity**

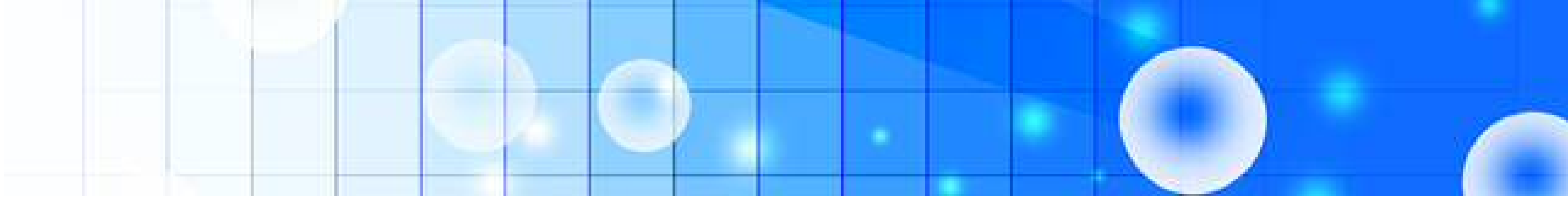
Next Steps

This year’s introduction of named respondents has opened the door to more contact with organizations that are actively implementing and maturing their data management efforts using the dimensions of data quality. We are already planning a number of case-studies about implementation of the dimensions and specifically pros and cons around implementation of the ***Conformed Dimensions of Data Quality*** standard discussed in this report.

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

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

Contact: Dan Myers



Introduction

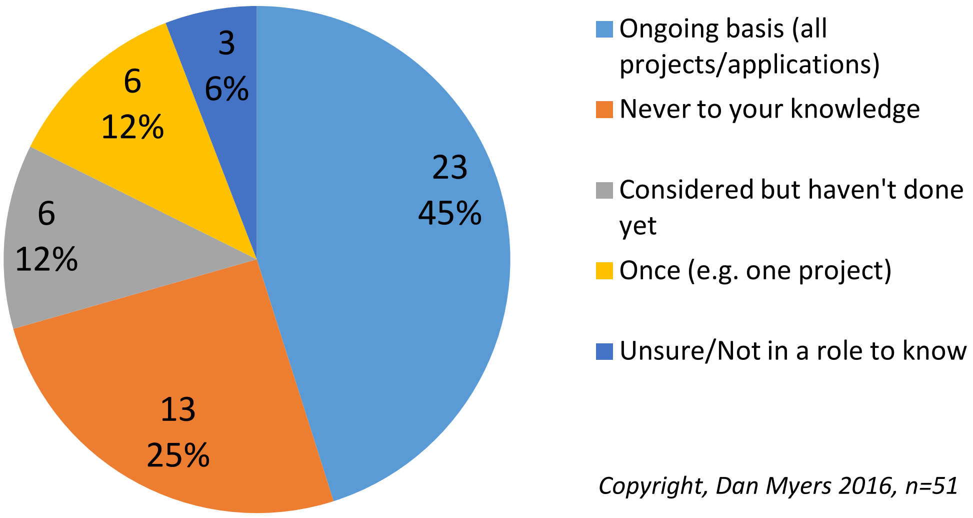
Last year we published the first annual survey regarding organizational use of the "Dimensions of Data Quality" (e.g. Accuracy, Completeness, Validity…etc.) As far as we can tell, the "Dimensions of Data Quality", as we'll refer to them in this paper, have been used since the late 1990s[i]. Many areas of the information and data quality domain have matured since then, but unlike other professions- where standards form over time, we haven't seen a standard evolve for the dimensions of data quality. For the reasons outline in this white paper, and on the website, 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

- Provide a standardized common language to describe data quality
- 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
  - Consistency between projects enables aggregation and comparison of results
  - Reduced tool configuration and customization
- 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

In 2016 the survey shows that nearly a majority of respondent's organizations use some form of the Dimensions of Data Quality in an "Ongoing Basis" (45%). More about the year-over-year growth on page 5.

Figure 2a. How often does your organization classify data related defects using the dimensions of data quality?



Copyright, Dan Myers 2016, n=51

A Little History Helps

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, Dan Myers and a few technical reviewers have identified the following areas of misunderstanding and disagreement. Generally speaking, the survey results affirmed this observation.

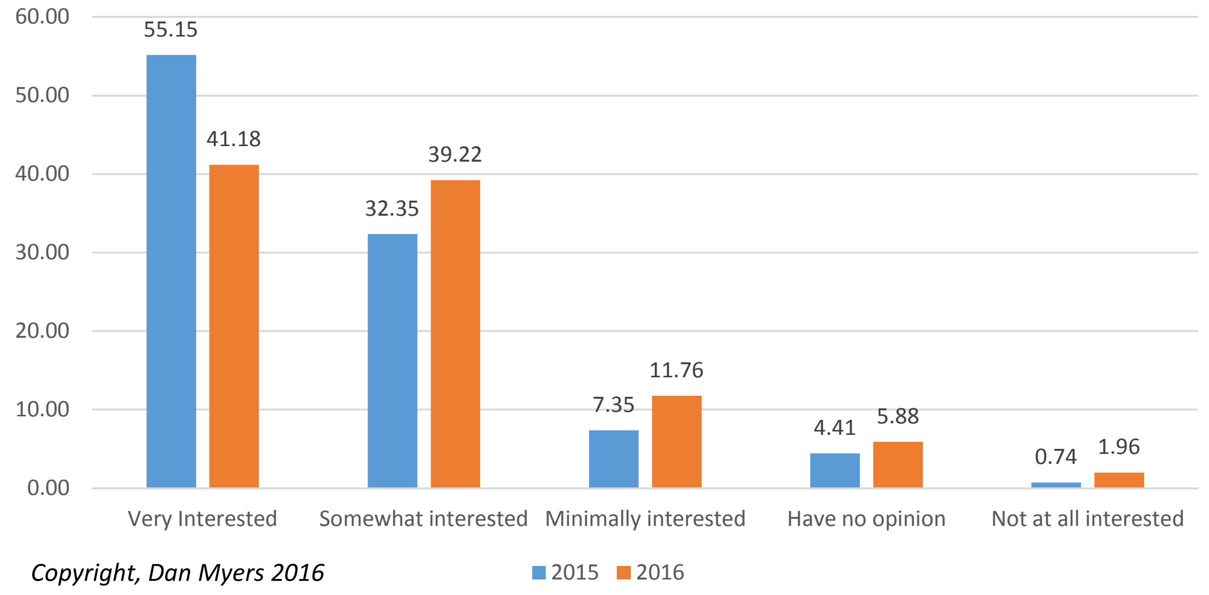
How to read the following table that illustrates current confusion regarding the Dimensions:

The Conformed Dimensions of Data Quality is the proposed standard set of dimensions we have defined (left column). These are often described using non-standard words (middle column) that differ from the standard terminology. Said another way, if you are looking for a data concept that you know and refer to that isn't listed in the Conformed dimensions you may want to look for key words that describe it in this middle column. When comparing the dimensions of data quality espoused by individual authors and organizations who have prescribed sets of dimensions of data quality in the past, we found some terms they used (usually inconsistently and sometimes inaccurately) which we included in the last column.

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

Year to Year Comparison of Interest in Standard

Figure 1b. If an industry standard set of dimensions of data quality was available, how interested would you be in using that at your organization? (percentage of responses for year)



Copyright, Dan Myers 2016

In 2016, there was a decrease in the proportion of respondents self-identified as “Very Interested” in using an industry standard set of dimensions of data quality, but an increase in the “Somewhat Interested” or “Minimally Interested” categories (see Figure 1b, above). Assuming that the sample size is representative, there has been a moderation in the interest in adopting an industry standard set of dimensions of data quality. We don’t interpret this to mean a decrease in value of the concept of the Conformed Dimensions, but rather clarity that a valid sample size and sample characteristics is important going forward and that even an industry standard set of dimensions, will not meet every organization’s needs.

At this point we should call out the fact that there were only 51 respondents this year compared to 136 last year (38% of 2015). We interpret this to mean two things: first, we need a greater number of responses each year in order to ensure we have properly characterized the industry as a whole, and secondly, it’s too much work to acquire new respondents each year (see note in box at right).

A Note On Survey Response Size

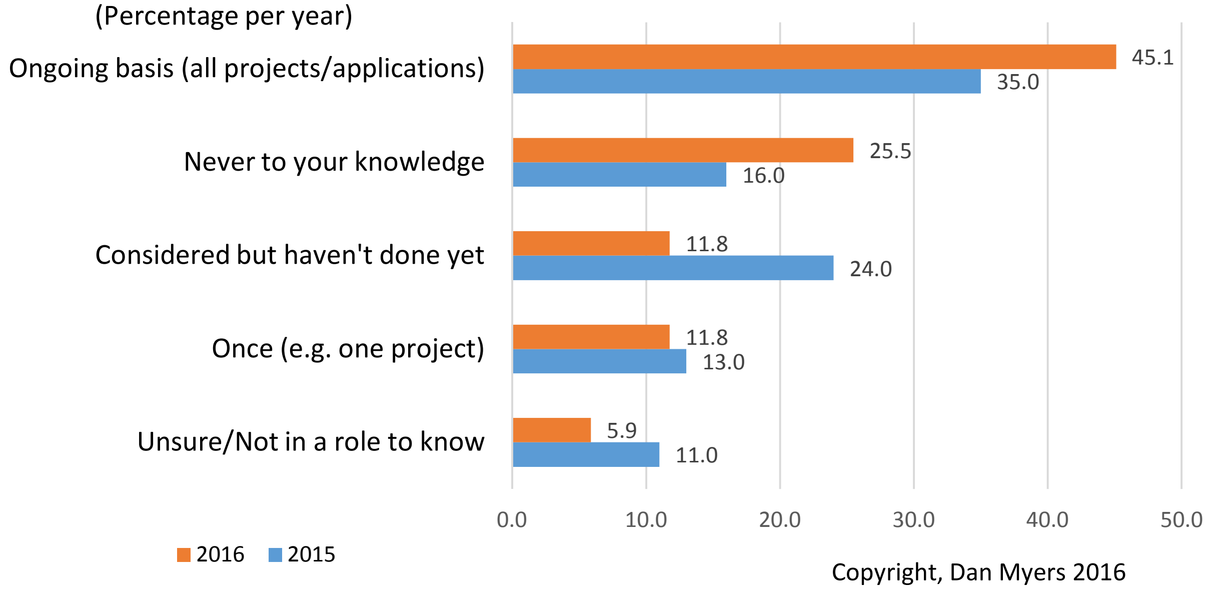
Even in 2015, (the first year that this survey was conducted) we identified the significant effort to acquire survey responses was going to be a hurdle for additional research on this topic. One way we are seeking to mitigate that is to enroll long-term respondents who are willing to take this survey yearly. Thirty percent (30%) of the 2016 respondents opted in, committing to participate in the survey each year going forward.

Perhaps you would like to contribute as a yearly survey participant? Provide your contact information at the following URL in order to sign up for next year’s survey.

http://dqm.mx/surveyopt-in

Usage of the Dimensions

Figure 2b. How often does your organization classify data related defects using the dimensions of data quality?
(Percentage per year)

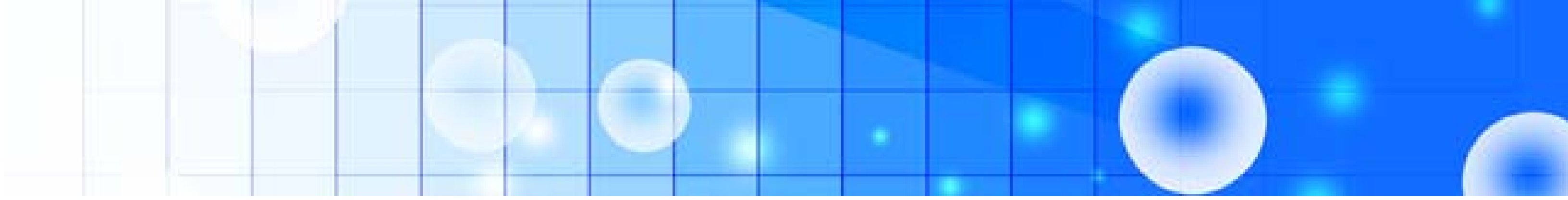


Copyright, Dan Myers 2016

Prior to understanding which dimensions organizations are typically using, we wanted to understand if they use dimensions to classify data issues currently. Last year, we were pleasantly surprised by the large proportion of organizations answering that they do use the dimensions, and for that matter, use them on an ongoing basis (35% in 2015). As you can see that number has jumped in 2016 and now is even higher at 45%.

	2015	2016	Increase/Decrease
Ongoing basis (all projects/applications)	35.0 %	45.1 %	**Increase of 10%**
Once (e.g. one project)	13.0 %	11.8 %	~same
Considered but haven't done yet	24.0 %	11.8 %	**Decreased by 12%**
Never to your knowledge	16.0 %	25.5 %	**Increase of 9-10%**
Unsure/Not in a role to know	11.0 %	5.9 %	~same

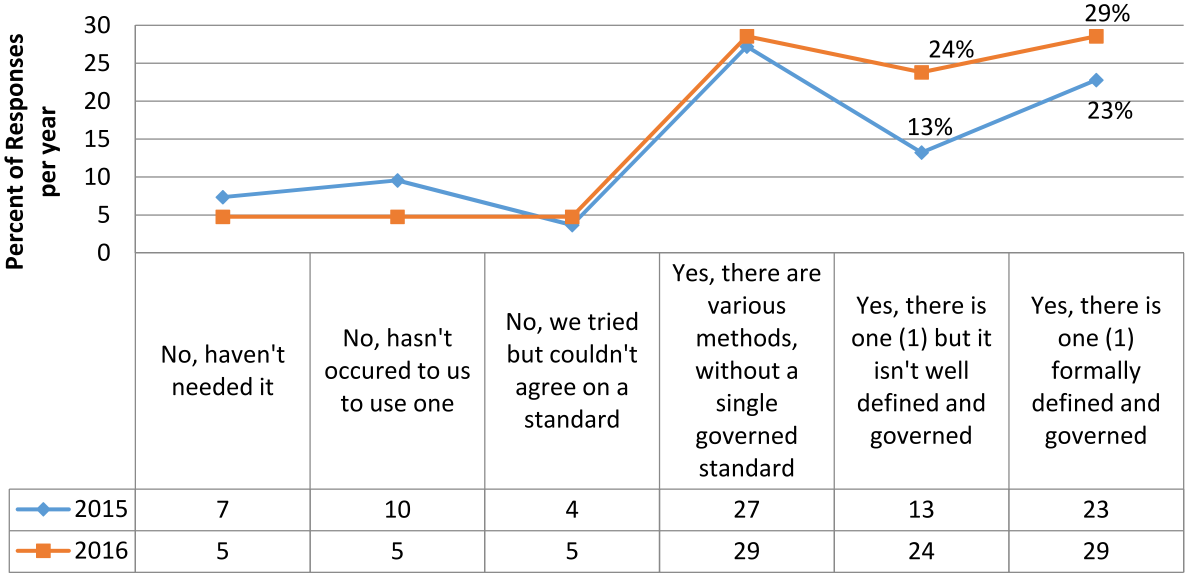
As seen in the table above, those organizations who answered that they use the Dimensions on an "Ongoing Basis" has increased, where as those answering that they have only "Considered" it has decreased. This seemingly implies a growth in maturity. It would appear that it has become less of a fad and rather a common practice for organizations. We are unsure how to interpret the "Never to your knowledge" growth of 9-10% this year, unless the respondents this year happened to be less educated about the broader organizational use or included a segment of smaller organizations that truly have not yet been exposed to them until just recently.



Governance of Dimensions Used

When asked whether responding organizations have a method of categorizing data quality issues using characteristics of data and fitness for use, like the Dimensions of Data Quality, we found some interesting trends. Generally speaking, the categories provided for this question can be ordered in terms of maturity (where in the table below, the left most option represents the least mature organizational response and the right most represents the most mature.

Figure 3c. Organizational maturity change over time through use Dimensions of Data Quality (DQ) as method to categorize DQ issues



	No, haven't needed it	No, hasn't occured to us to use one	No, we tried but couldn't agree on a standard	Yes, there are various methods, without a single governed standard	Yes, there is one (1) but it isn't well defined and governed	Yes, there is one (1) formally defined and governed
2015	7	10	4	27	13	23
2016	5	5	5	29	24	29

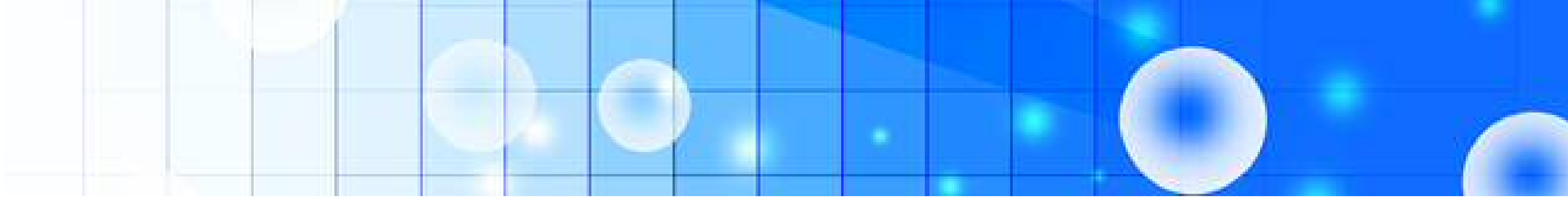
Generally speaking, if the data collected this year (2016) is consistent with last year (2015), we'd expect the curves, represented here in a line-chart format[1], to be reasonably similar. The noted change may be a shift in individual organizational maturity, and hopefully representative of the shift in maturity of the industry itself over time. We interpret the general similarity between responses year-over-year to mean that there still are a number of organizations that in similar proportions each year, as change is slow.

Meaningful Observations this Year:

- It seems that fewer organizations are completely unaware of the Dimensions (those answering, "No, hasn't occurred to us to use one" has dropped from 9.6% to 4.8%).
- We also notice that more organizations at least have a defined corporate standard.
  - About 24% this year, versus 13% last year, have one single, formally defined and governed set of dimensions.
  - About 29% this year, versus 23% last year, have at least identified a corporate set of dimensions of data quality even though it may not be comprehensively governed.

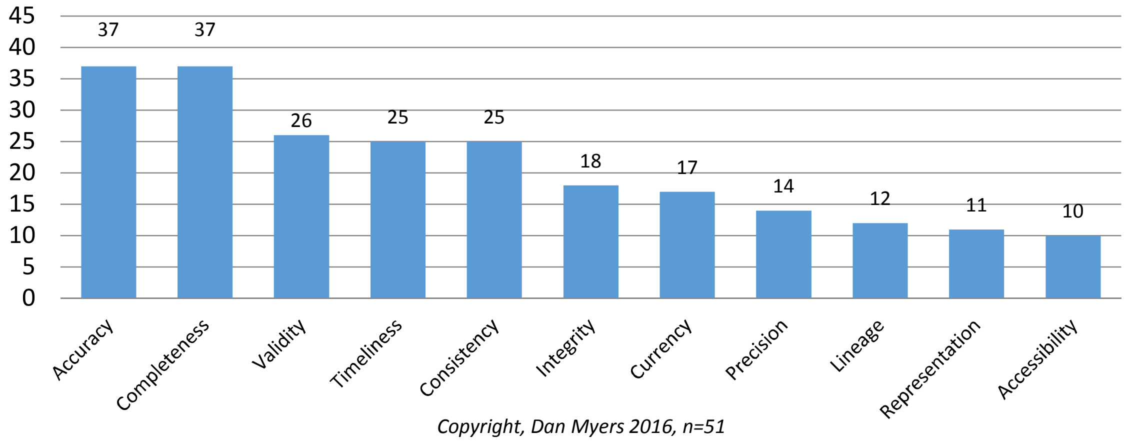
Last year we observed that most (45%) organizations that do use the dimensions of data quality on an "Ongoing basis (all projects/applications)," had a formally defined and governed set of dimensions. We found that number has even increased to 50% in 2016, which reinforces the trend we see toward intra-organizational formalization.

[1] Although technically the concepts aren't purely continuous, as one would like to use with line charts, this method of illustration lends itself well to showing how individual organizations mature over time and therefore the industry matures as well.



Popularity of Each Dimension

Figure 4a. Select any of the following dimensions currently used by your organization. If your organization's definition differs from those listed below, please comment on the difference. (Count of orgs reporting use of each dimension)



Copyright, Dan Myers 2016, n=51

The primary question of the survey was geared to get feedback regarding which Dimensions are used and how organizations define each of them. The list that we provided is a proposed set of Conformed Dimensions of Data Quality provided in the appendix. The most current version is at http://DimensionsOfDataQuality.com.

This year (2016) is the first year that we can compare year-over-year results which has made this specific question more interesting. Even though the survey response size was smaller this year, (51 in total) general observations can be made.

Summary of Ranking Changes for 2015-2016

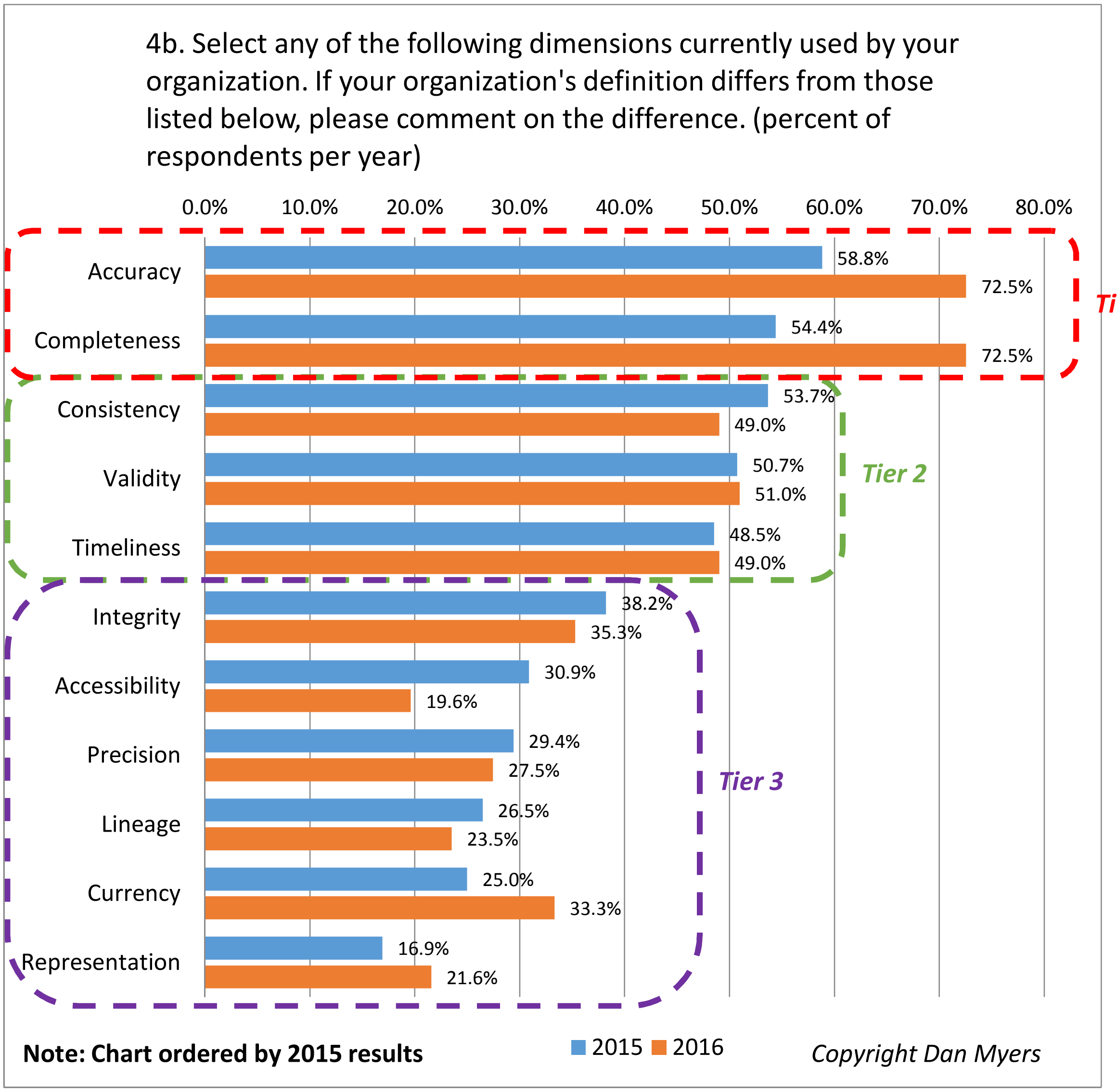
- Strong increase in usage of Accuracy and Completeness, which tied for the most highly used dimensions in 2016
- As seen in year-over-year comparison (next page) three levels of usage have become evident
  Tier 1: Accuracy & Completeness
  Tier 2: Consistency, Validity and Timeliness
  Tier 3: All others
- We observed a significant drop in use of Accessibility (from #7 to #10)
- Significant jump in the use of Currency (from #10 to #7)

Ranking Changes

2015		2016	
1	Accuracy	1	Completeness
2	Completeness	2	Accuracy
3	Consistency	3	Validity
4	Validity	4	Timeliness
5	Timeliness	5	Consistency
6	Integrity	6	Integrity
7	Accessibility	7	Currency
8	Precision	8	Precision
9	Lineage	9	Lineage
10	Currency	10	Accessibility
11	Representation	11	Representation
12	Existence	<now within Completeness>	

Unpacking the Rankings:

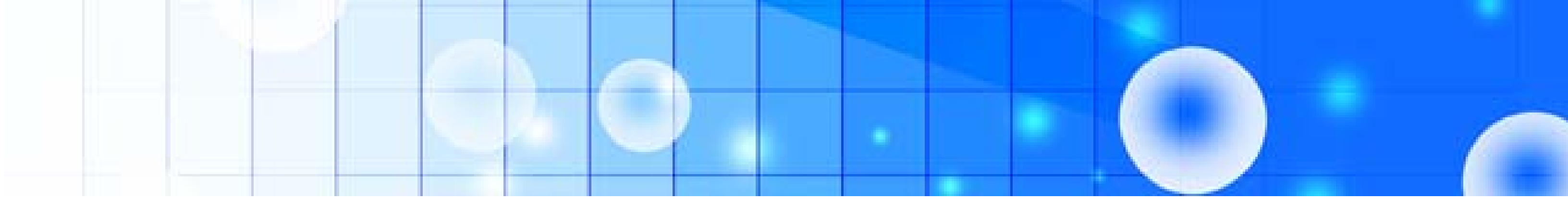
Arguably, Completeness and Validity are two of the best foundational dimensions on which to establish a data quality effort given that their explanation is strait forward and easy to test. Completeness and Accuracy tied for rank number one and Validity ranked number three in 2016. One apparent reason that these two have bubbled up near the top, two years in a row is because virtually all DQ tools accommodate out of the box functionality to measure these respective concepts. Extending on that premise, Consistency, which initially ranked 3rd in 2015, dropped to 5th in 2016- likely because it is harder to describe and is more difficult to test. We'll pay careful attention to where Consistency ranks in 2017 with a larger sample and named year-over-year survey respondent tracking. (See side-bar topic ***A Note On Survey Response Size*** on page 4)



4b. Select any of the following dimensions currently used by your organization. If your organization's definition differs from those listed below, please comment on the difference. (percent of respondents per year)

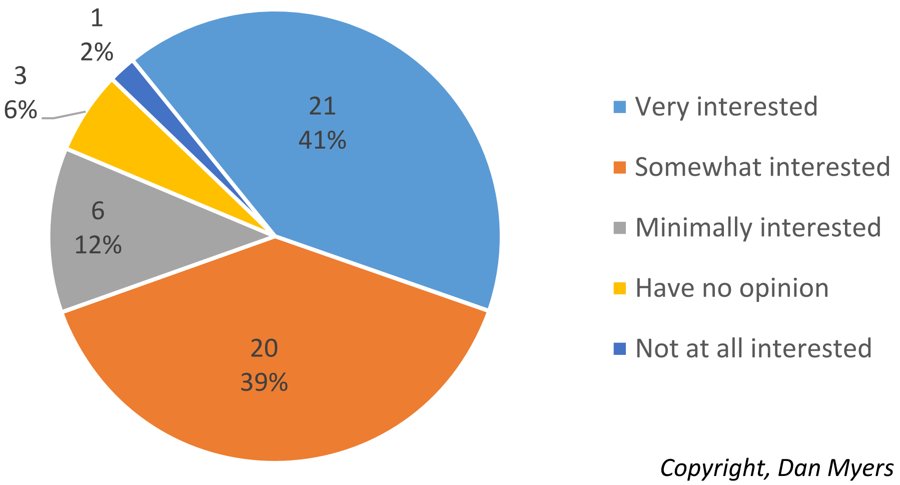
Note: Chart ordered by 2015 results

Copyright Dan Myers



Conclusion

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 2016, n=51

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. For two years now, respondents have overwhelmingly shown interest in the idea of a standard (88% in 2015 and now 80% in 2016).

We believe that although there is a self-selection bias among the people answering the survey (people experienced in the dimensions or who care more than average data management professionals about a standard), the results show that there is enough interest in a standard to actively pursue it.

For this reason a new blog on the topic of the Dimensions of Data Quality will be launched on the Conformed Dimensions of Data Quality website in 2017. The blog will be focused on identifying day-to-day examples of data quality which can be measured using the Conformed Dimensions.

Appendix

General Survey Information

Count of Full Responses:	51
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

Dimensions Listed in the Survey (Options to Choose from)

Question Text: Select any of the following dimensions currently used by your organization. If your organization's definition differs from those listed below, please comment on the differences. If you have additional dimensions please add to the "Other" field. Due to a software limitation you have to enter a comment in order to check a dimension: please enter "No Comment" if you don't want to comment on each dimension you select.

- □ Accuracy- Accuracy measures the degree to which data factually represents its associated real-world object, event, concept or alternatively matches the agreed upon system of record.
- □ Consistency- Consistency measures whether or not data is equivalent across systems or location of storage.
- □ Precision- Precision measures the number of decimal places and rounding of a data value or level of aggregation.
- □ Timeliness- Timeliness measures how quickly data is available.
- □ Accessibility- Accessibility measures how easy it is to acquire data when needed, how long it is retained, how access is controlled, and whether facts exist as data.
- □ Currency- Currency measures how quickly data reflects the real-world concept that it represents.
- □ Completeness- Completeness measures the degree of population of data values that exist in a data set. (example: columns and rows).
- □ Validity- Validity measures whether a value conforms to a preset standard (example: a domain of permitted values, domain ranges, business rule, data type, format pattern, or storage format).
- □ Integrity- Integrity measures the structural or relational quality of data sets. (example: referential integrity, uniqueness, cardinality).
- □ Representation measures ease of understanding data, consistency of presentation, appropriate media choice, and availability of documentation (metadata).
- □ 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.
- □ Other- <free form text box here>

Additional Questions Included in the Survey

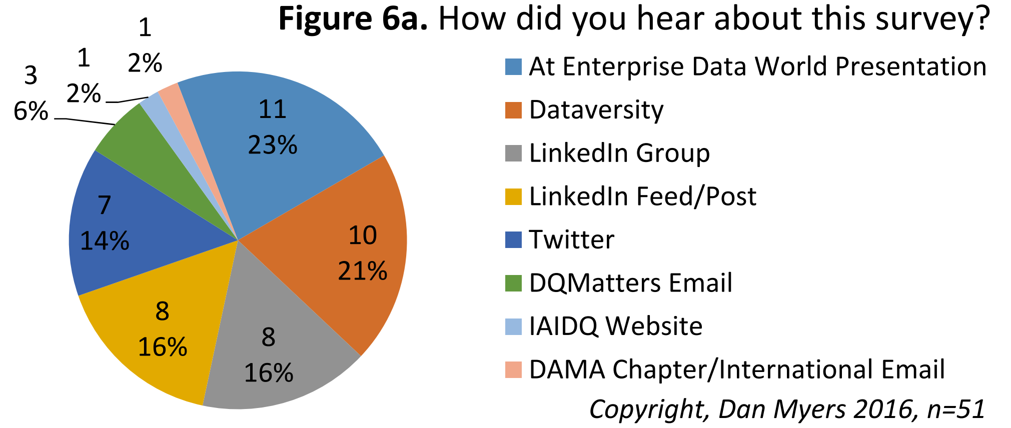
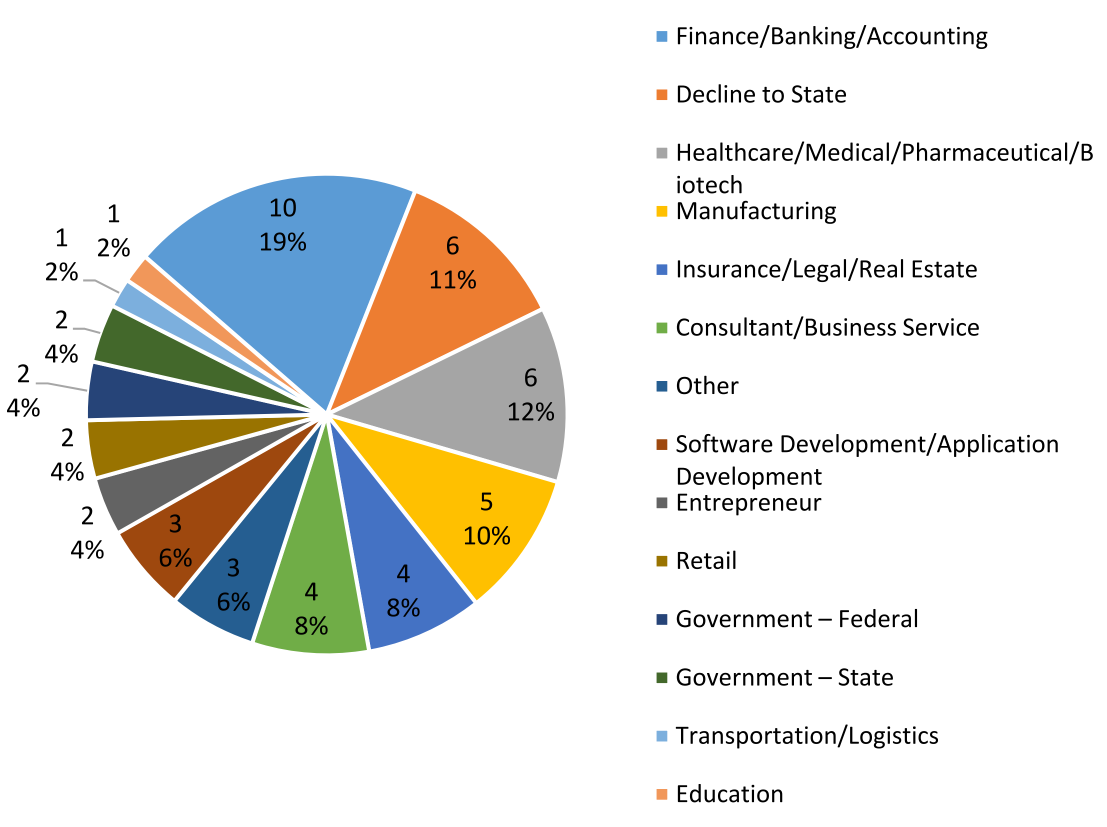


Figure 6a. How did you hear about this survey?

Copyright, Dan Myers 2016, n=51

Figure 10a. Please choose in which industry your organization is categorized.



Copyright, Dan Myers 2016, n=51

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