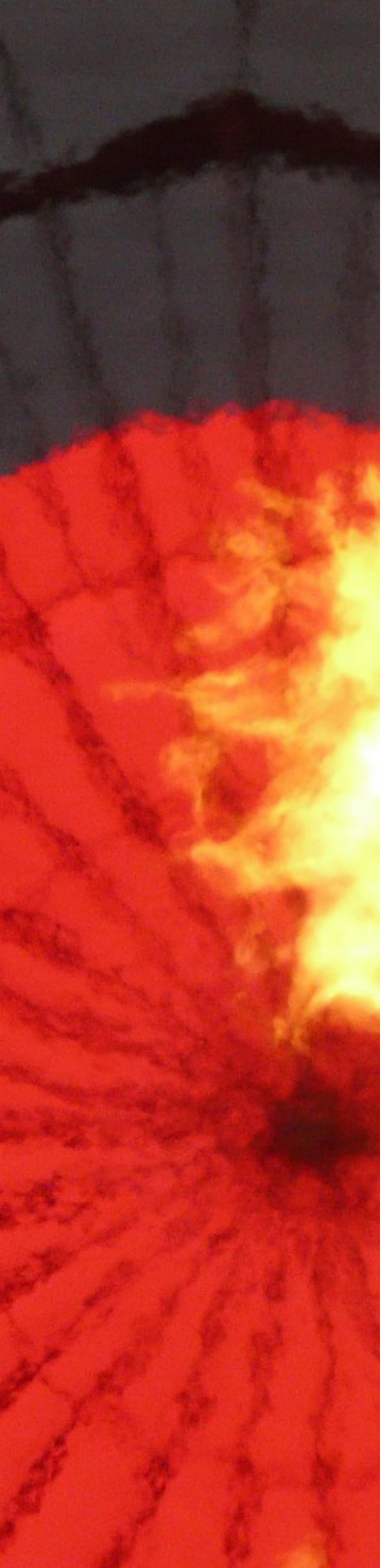


2017 Annual Report on the Dimensions of Data Quality

Year-three:
Significant Increase in Measurement of "Accuracy"
But Many Industries Still not Using the Dimensions of Data Quality

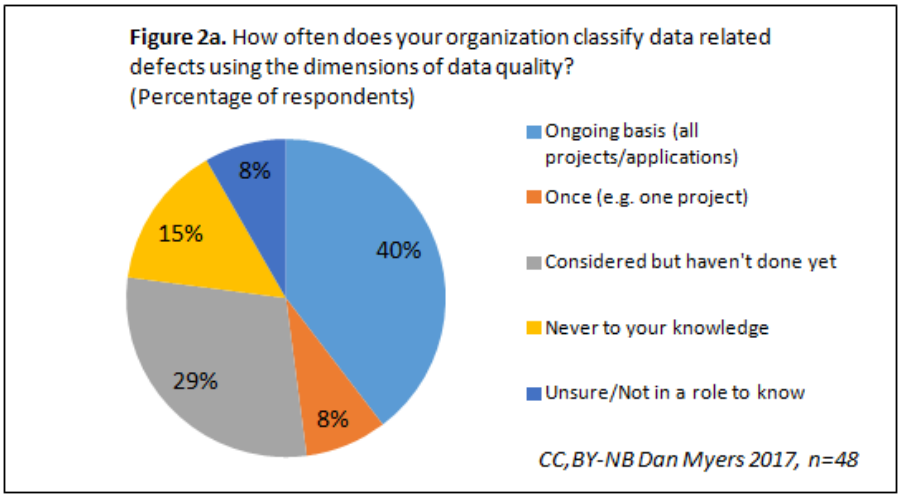
August 15, 2017
The 3rd Annual Whitepaper
Sponsored by

DQMatters
Data Quality eLearning

Executive Summary

This year marks the third year we've held the Annual Dimensions of Data Quality Survey, and we continue to uncover industry trends in data quality and affirm finding from prior years. The purpose of this survey 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. Below is a summary of the 2017 findings, but don't stop there, take the time to read through the details on the following pages and sign up for the affiliated blog about the Conformed Dimensions of Data Quality (see below). There were 51 complete responses to the survey.

Summary of Findings



- 40% of respondent's organizations classify data related defects using the dimensions of DQ on an ongoing basis. (See figure 2b at right)
- Many industries falling behind without use of the dimensions of data quality (e.g. Utilities, Chemicals, Mining, Petroleum, Textiles, Federal Government…etc.)
- Accessibility dimension jumped from 10th to 7th this year, which we loosely associate with the larger focus on Data Lakes and having more data available in one place for data consumers.
- Top 4 most popularly used dimensions are: **Accuracy, Completeness, Consistency,** and **Validity**

Discussion

This year we noticed a significant jump in the number of organizations measuring Consistency from around 49%, last year, to almost 73% in 2017. Even Accuracy, which often demands real-world observation by humans, had an increase of nearly 10%, now measured by 81% of organizations (up from only 59% in 2015).

In order to help explain the proposed Conformed Dimensions of Data Quality we started a new blog in January, 2017. If you haven't already signed up, please do so using the URL or QR code (bottom right).

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

Blog URL:
http://dimensionsofdataquality.com/blog

Blog Signup:
http://dqm.mx/cddqblog2017

Introduction

Request Copies of Prior Whitepaper Years Here!

This is the third year that we've conducted the Annual Survey about organizational use of the "Dimensions of Data Quality" (e.g. Accuracy, Completeness, Validity…etc). The original purpose of this survey was *to measure how frequently different dimensions of data quality are used and the level of interest in a cross-industry agreed upon standard set of dimensions of data quality.* It is still focused on the use of the Dimensions of DQ, although less on the desire for a standard and going forward more on how the Conformed Dimensions of Data Quality (CDDQ) are used. Additionally, this year, we've included information from end-users of the CDDQ and started a new LinkedIn group dedicated to the discussion of the standard in order to answer questions and facilitate sharing about use of the dimensions of DQ in general.

In the following paragraph, we review the value of using a standard set of dimensions, but honestly it can't be said better than one of the survey respondents who stated that, "the use of the Conformed Dimensions is has prevented fist fights over what is the 'correct' definition of each dimension during implementation". This is at the heart of why we are passionate about the standard as a communication tool for measuring and explaining quality.

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

A Little History Helps

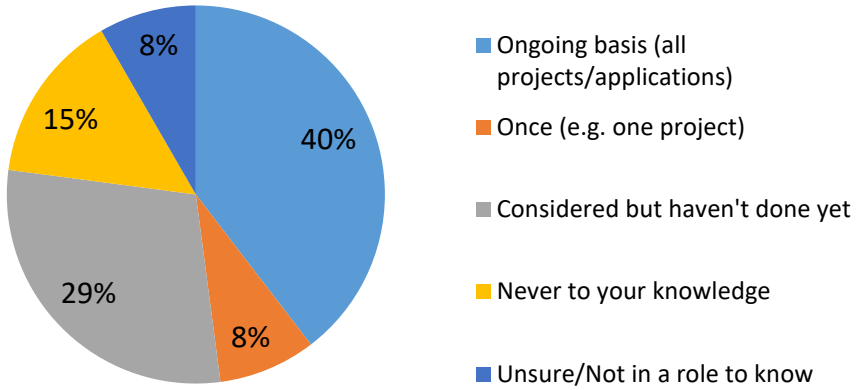
In a series of articles, addressing the lack of agreement on the Dimensions of Data Quality in Information-Management.com in 2013, Dan Myers proposed a conceptual list of dimensions that agrees with most authors' definitions. Based on that work and discussion with data management industry leaders, Dan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

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

Usage of the Dimensions

In 2017, the survey shows that 40 percent of respondents' organizations use some form of the Dimensions of Data Quality in an "Ongoing Basis", with the sum of those using them either once or in an ongoing basis, nearly the same as last year- around 50 percent.

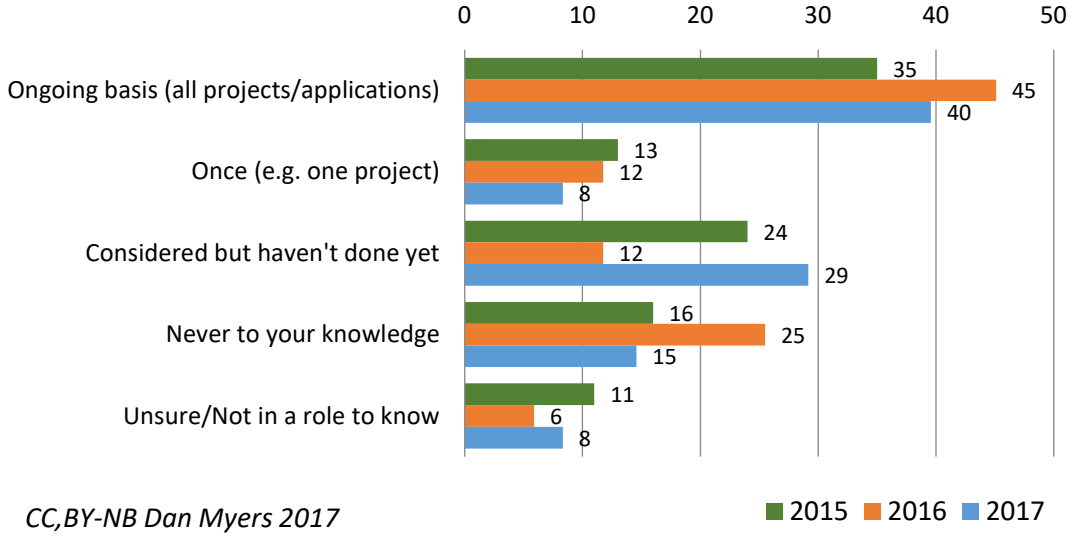
Figure 2a. How often does your organization classify data related defects using the dimensions of data quality? (Percentage of respondents)



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

In 2016, there was a significant drop in the number of respondents that reported they have considered the use of the dimensions, but hadn't done it yet (down to 11.8%), so we interpreted that to mean that there may have been an influx of organizations beginning the information quality journey, but because the 2017 numbers (29%) are closer to 2015 (24%) we believe that was likely a survey cohort specific nuance instead.

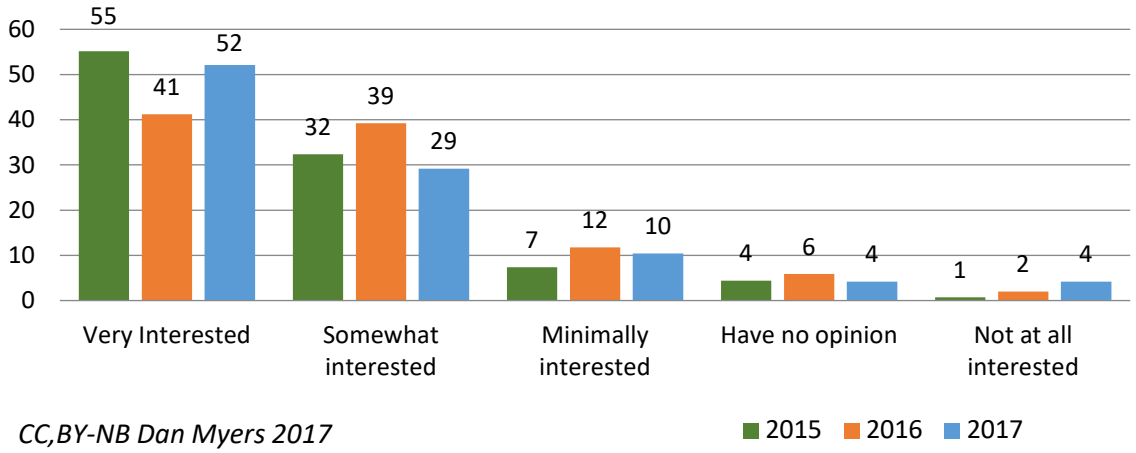
Figure 2b. How often does your organization classify data related defects using the dimensions of data quality? (Percentage of respondents)



CC,BY-NB Dan Myers 2017

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? (percent of respondents)



CC,BY-NB Dan Myers 2017

In 2016, we identified a significant drop in percentage of respondents "Very Interested" in the standard, but with the new 2017 data we see that was more likely a single-year anomaly. We expect that as more people hear about the standard and additional academic validation of the CDDQ (see call-out on the following page), becomes available we suspect these numbers will even increase.

Additionally, in January of 2017 a new blog was established on the CDDQ website and so far seven posts have been completed in 2017 (see QR code on page 1 for link).

The blog is geared toward broader audiences (not just DQ practitioners), but with valuable articulation of the CDDQ underlying concepts and how they apply in a real-world context. Feedback about the blog has been positive and the blog will become an important communications channel regarding changes to the standard over time. The most popular blog post was the April blog titled, "Data Quality Lessons Learned at Starbucks" (http://dqm.mx/cddqreport-b4).

A Note On Survey Response Size

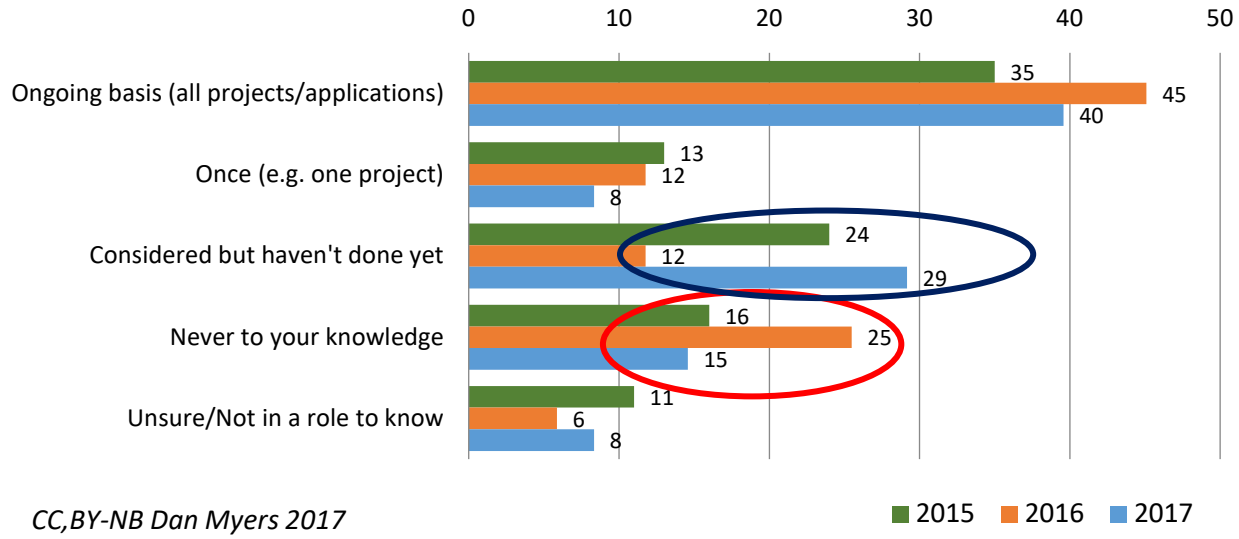
In 2016, we were looking for a way to increasing the over-all number of survey respondents, so we started collecting names and contact information for respondents that are interested in providing their input annually. From 2016 to 2017 this grew from 30% to 46% showing a greater interest and likelihood of adoption.

Would you 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 of respondents)



CC,BY-NB Dan Myers 2017

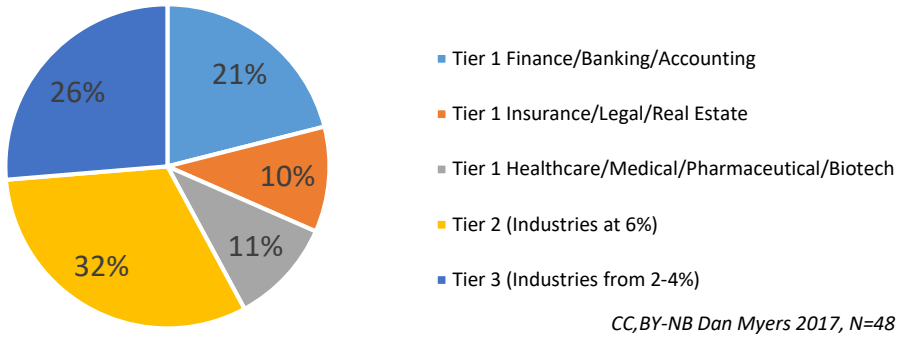
The basic purpose of this annual survey is to understand the current level of usage of the dimensions of data quality, and the question in Figure 2b (above) is at the heart of understanding that concept. The new 2017 data shows that last year's jump in responses about not having used the dimensions (red circle above) was likely a 2016 specific anomaly- much different than 2015 and 2017. Similarly the drop in those considering it but not having done it (blue circle) looks like a year-specific anomaly.

If our survey sample is representative of all organizations, somewhere around 40% of organizations are using the dimensions of data quality, and hopefully with more clearly articulated definitions available in the CDDQ, and now examples (via the blog), other organizations will use them as well. One concern we have, however, is that certain industries seem prone to not adopting the dimensions as well as others. In table 1 (below) we've grouped the industries that most frequently use the dimensions of data quality into tiers:

Grouping	List of Industries in Group
Tier 1	Finance/Banking/Accounting (19.1%) Healthcare/Medical/Pharmaceutical/Biotech (10.6%) Insurance/Legal/Real Estate (8.5%)
Tier 2 (Industries at 6%)	Government – State/Retail/Manufacturing/Software Development/Application Development/Consultant/Business Service/Other
Tier 3 (Industries from 2-4%)	Utilities/Chemicals/Mining/Petroleum/Textiles/Government – Federal/Media/Entertainment/Transportation/Logistics
Tier 4 (No representation 0%)	Entrepreneur/ISP/Web Host/IT Services Outsourcer/Education/Government/Military/Public Administration

Of course the larger concern is whether information quality understanding and program development is occurring within these industries (outside the scope of this report). More specifically, if our readers have any influence over the adoption of the dimensions in tiers 2, 3, and 4, please contact us to find out how we can assist. The cost of implementing the CDDQ is negligible compared to the value these organizations will gain, so step up and get involved by improving your organizations data quality.

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



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

Here is the breakdown of respondents who **do** use the dimensions of data quality in some capacity by tier.

- **As seen on the right, respondents in Tier 1 are:** generally industries in the for-profit service sectors with high levels of regulations.
- Industries that currently **Do Not** use the dimensions of data quality include:
  1. **Not-for-Profit:** Governments (Military, Federal, Local, Public Administration, Education[public])
  2. **Produce Physical Products:** (often commoditized)
- **Unexplained industries include:**
  1. Retail, which has both commodity priced segments but also very high-end niche segments.
  2. Media/Entertainment industries have sizeable resources.
  3. Software/Application Development/ISP/IT Service Outsourcer

	Count	%
Tier 1		
Finance/Banking/Accounting	8	21%
Insurance/Legal/Real Estate	4	11%
Healthcare/Medical/ Pharmaceutical/Biotech	4	11%
Tier 2 (Industries at 6%)		
Government – State	3	8%
Retail	2	5%
Other	2	5%
Software Development/Application Development	2	5%
Consultant/Business Service	2	5%
Manufacturing	1	3%
Tier 3 (Industries from 2-4%)		
Chemicals/Mining/Petroleum/Textiles	2	5%
Non-profit other than listed above	2	5%
Utilities	2	5%
Transportation/Logistics	1	3%
Government – Federal	1	3%
Government – Local	1	3%
Media/Entertainment	1	3%
Tier 4 (Unrepresented)		
Entrepreneur/ISP/Web Host/IT Services Outsourcer/Education/Government/ Military/Public Administration	0	0%

Some of the unexplained areas (e.g. retail, media, and software/tech) are likely due to the small sample size, because our survey includes so few respondents from these industries. In order to better understand the drought of focus on quality in these tier 2, tier 3 and tier 4 industries we interviewed a few respondents in these industries. Our findings are available in the next section.

Interviews with Survey Respondents

Example 1. Mining Industry
Organization: $5 Billion (earnings) Mining Company

Historical Context: The mining industry hasn't fundamentally changed over last 100 years, resources are extracted from the ground and shipped to another place (sometimes pre-shipping refining is involved). So rather than product differentiation (like consumer retail products) mining companies are focused on reducing costs and risk mitigation in order to drive increased shareholder returns.

Recent Advances: Data has become more critical in order to manage risk but regulatory threats are still relatively minor compared to financial services organizations where individuals are at risk of going to jail if the data is misleading, fraudulent, or of poor quality.

Overall Tenor: With new advances in Big Data and operational Business Intelligence, C-level leaders are expecting more of their data teams, but back office manual scrubbing and validation, that was previously done over several weeks, can't keep up with these timeliness and accuracy demands. For that reason data quality has crept into scope, but not received the attention it needs. Primary constraints include the lack of data quality mechanisms within legacy plant control systems, and increase in the number of variables of data now collected, and complex/undocumented integration conducted in data lakes. Generally speaking data quality isn't getting any worse, but the challenge is getting harder for these reasons.

Regarding the CDDQ: The interviewee said that, the "money saved [by using the CDDQ] comes from the lack of fist fights at the beginning phases of a data analysis/quality effort" by not arguing so much about what each dimensions means.

Example 2. Utilities
Organization: Large Public Utility in the Pacific Northwest

Historical Context: The primary business of Utility providers is service provision via complex infrastructure. Asset management practices, now adopted by leading Utilities around the world have data driven decision making as their objective, but for an organization in which data stewardship and governance are in a low stage of maturity, the amount of work required to generate analytical reports about the state of our infrastructure is far too intensive at the current time. The time and institutional knowledge required to generate a comprehensive picture of asset conditions result in incomplete analysis and decisions that are less than optimal.

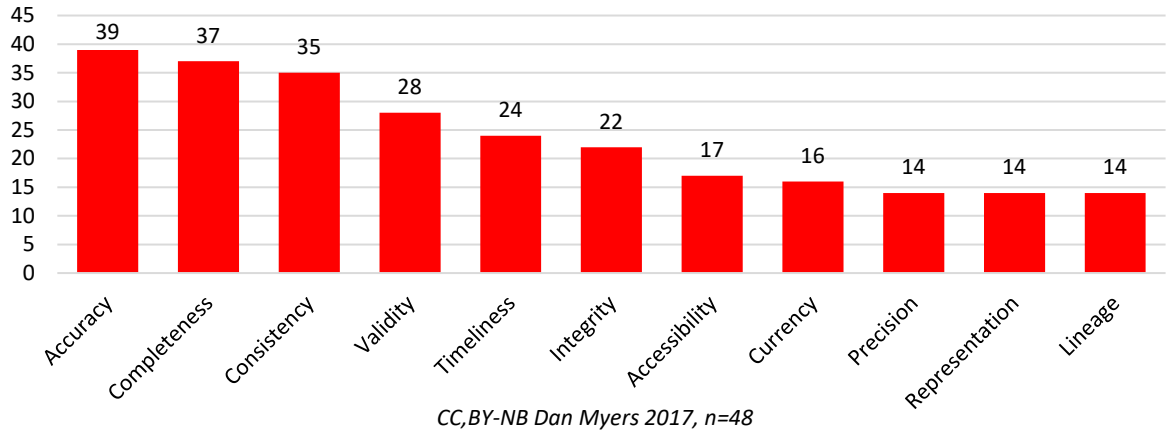
Recent Advances: New leadership at the municipal level, and within the Utility, have placed greater emphasis on demonstration of measurable results for our customers both internal and external. Utility industry regulation is focused on service levels and demonstration that they are being delivered at the lowest lifecycle costs while at the same time meeting national water quality standards and providing stringent protection of the environment. Such demonstration requires multiple data sources and complex analyses. For this reason there is a rising level of effort on formalizing data stewardship and data governance. The formation of industry associations such as the Water and Waste Water CIO Forum (http://watercioforum.com) is a national reflection of the increased emphasis on integration of technologies in Utility service provision and the shift in culture towards higher levels of information management maturity.

Overall Outlook: New initiatives and staffing are being put in place to facilitate the shift from foundational levels of enterprise information management towards formally measured metrics in a number of Data Quality Dimensions.

Regarding the CDDQ: The availability of a consistent and broadly accepted vocabulary of data quality dimensions has supported the communication to many stakeholders that there is a trend towards a maturing practice of measuring data quality. The extension of the traditional concepts of quality to a broader and unified view that aligns with the reality of our technology infrastructure helps us to provide a tool for utility decision makers to specify service levels for data management that match the complexity, frequency and impact of the decisions we make.

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 organizations using each dimension)



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

Another one of the key questions of the survey was geared to get feedback regarding was, which DQ dimensions are used and how organizations define each of them. The list that we provided was the Conformed Dimensions of Data Quality (CDDQ) which is available in the appendix. The current version of the CDDQ is at http://DimensionsOfDataQuality.com. The results of this question are shown in the bar chart above.

2017 was the second year that we've had year-over-year results which has made this specific question more interesting. Unfortunately, this report doesn't include research about which of the following fluctuations in ranking is part of organizational behavior (e.g. leadership's political choices) rather than a change in actual data quality needs. Having said that, the jump in ranking for Consistency (5th to 3rd) and Accessibility (10th to 7th) are particularly noteworthy. We don't know for sure, but it seems that the strong push by BI/Analytics vendors regarding ways to present your data (e.g. dashboards, heat maps, 3D maps…etc) may have shifted the focus on Accessibility in the rankings. If you (our readers) have other hypothesis, we encourage you to please, share them.

Summary of Ranking Changes for 2016-2017

- **Accuracy was reported as the most used dimension** in 2015, and now in 2017 we have returned to that ranking, confirming our earlier observation that these two dimensions are at the heart of recent DQ efforts.
- **Accessibility jumped from 10th to 7th this year** which we loosely associate with the larger focus on Data Lakes and having more data available in one place for data consumers.
- **Consistency climbs back to 3rd position** from 5th place.

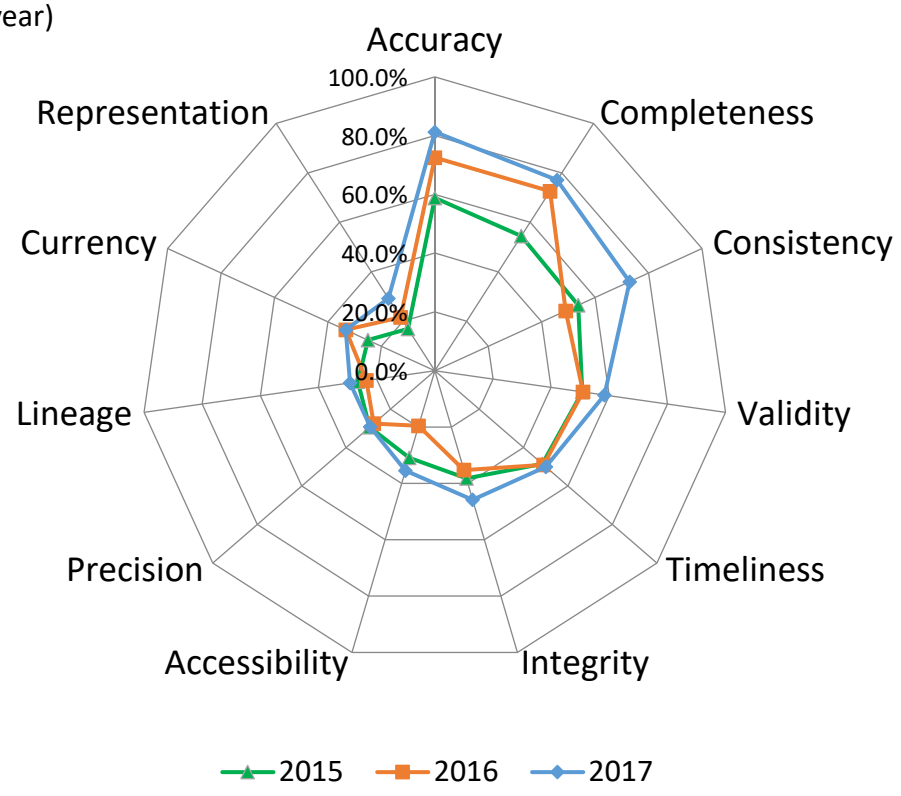
Ranking Changes

2016		2017	
1	Completeness	1	Accuracy
2	Accuracy	2	Completeness
3	Validity	3	Consistency
4	Timeliness	4	Validity
5	Consistency	5	Timeliness
6	Integrity	6	Integrity
7	Currency	7	Accessibility
8	Precision	8	Currency
9	Lineage	9	Precision
10	Accessibility	10	Lineage
11	Representation	11	Representation

Unpacking the Rankings

During our discussion about the usage of the dimensions, we discussed the need for a standard for the dimensions of data quality for every industry. Although we don't conceptually understand this need, we still have relatively high number of respondents that report that this is needed. About 46% say it is Very Important, and 27% say it is Somewhat Important. During our discussions with one of the respondents regarding this topic, it was pointed out that the need isn't likely for a whole new set of dimensions, but rather a different weighting for importance of each dimension, based on industry. Assuming this is true, we could identify the generally agreed upon weightings per industry (e.g. mining: Accuracy must be ≥ 90% and Completeness ≥ 80%...etc). Using this information we could plot each industry on a radar-chart (somewhat similar to the following chart where we've charted respondent usage of the dimensions by year). Then these unique "fingerprints," as our interviewee so adeptly named them, could be represented graphically for reuse by people unfamiliar with that industry's needs.

Figure 4d. 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)



CC,BY-NB Dan Myers 2017

We understand that this idea is a bit abstract and needs additional development and thorough testing, so we're going to begin a series of interviews, followed by focus groups and surveys to identify applicability and a proposed methodology. If you consider yourself a proponent of having an industry specific standard, e.g. unique to healthcare, please contact us to get involved we'd love to get your input. info@dimensionsofdataquality.com

Relationship Between use of Dimensions of Data Quality and Organizational Data Quality Levels

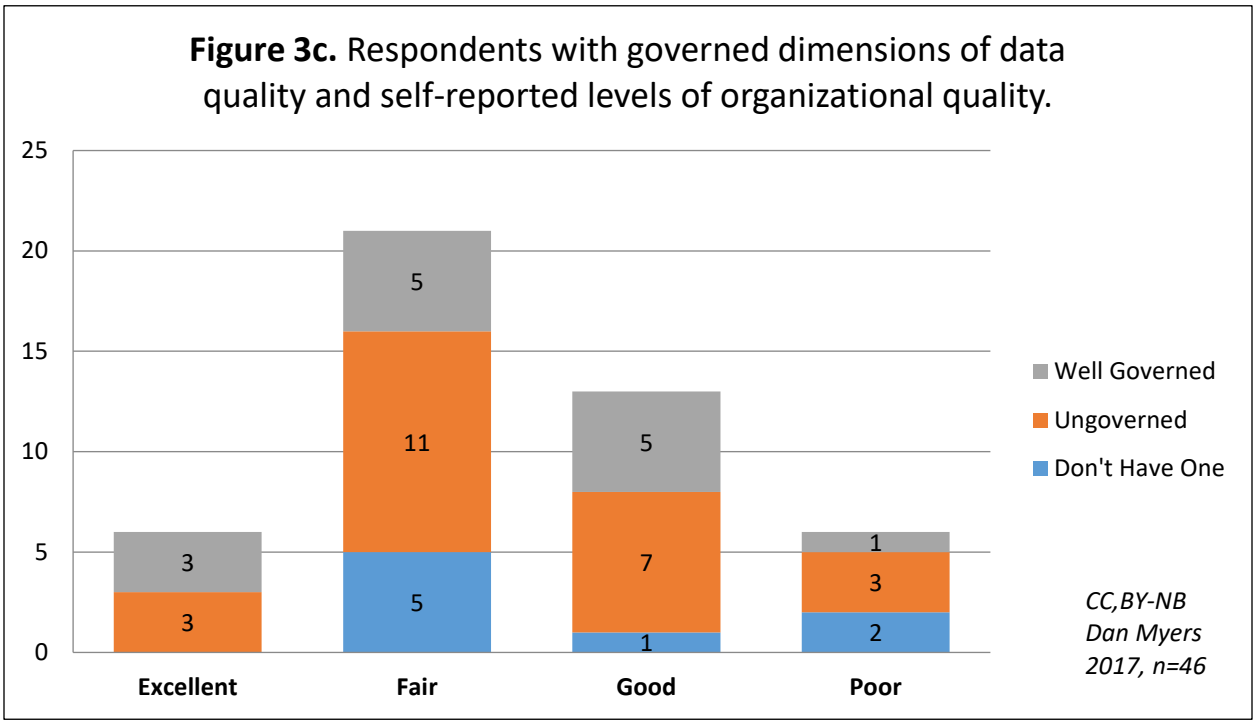
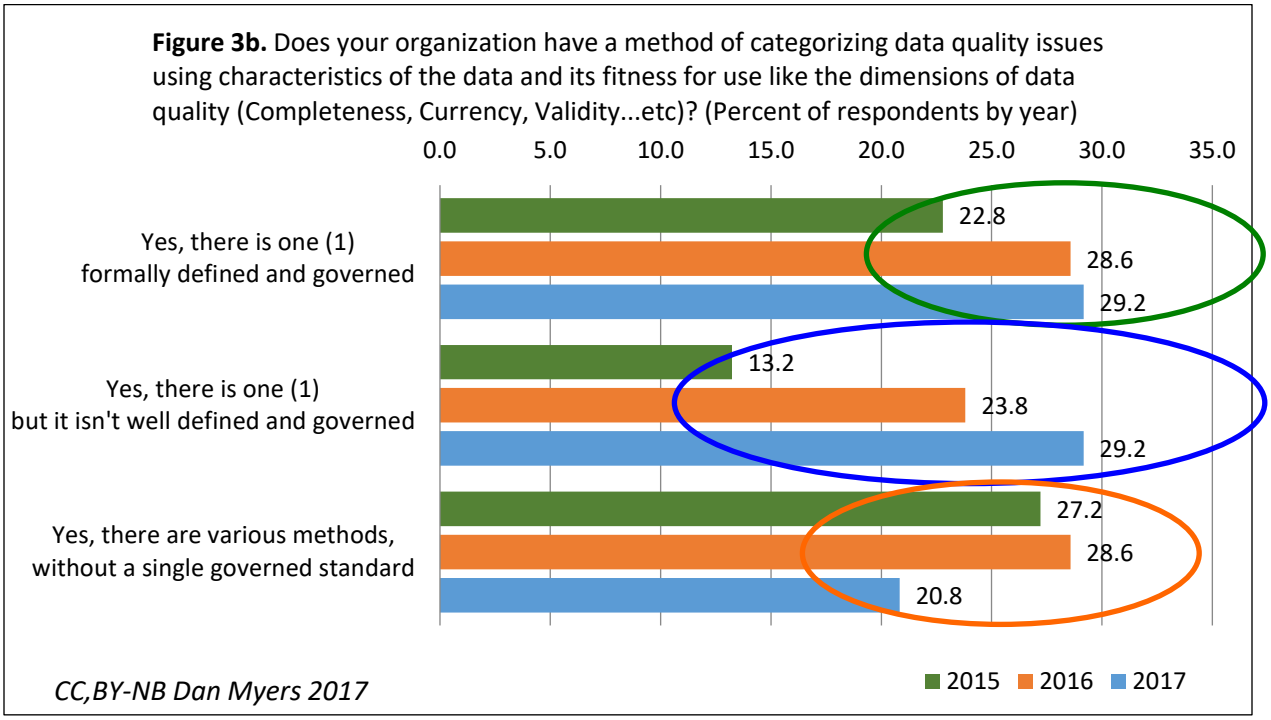


Figure 3c. Respondents with governed dimensions of data quality and self-reported levels of organizational quality.

In order to better understand whether organizations with well defined and governed dimensions of data quality reported higher levels of data quality, we tabulated the results (show above). Although we expected that organizations with a well governed set of dimensions of data quality in use, would report higher levels of data quality, we didn't necessarily find that to be the case. We did notice that only organizations who do use the dimensions also self-reported levels of "Excellent". Interestingly enough though, some organizations that don't use the dimensions, report "Fair" levels and one reports a "Good" level.

We do have to say that the sample size is small and we'd love to see this same analysis done with thousands of respondents. Next year when you see the survey advertised, please consider taking it and forwarding it with a note to data management professionals you know- asking them to take the survey.

Conclusion



In conclusion, we are generally optimistic that organizations are improving the strength of their data quality measurement through the increased use of the dimensions of data quality. As shown above, there has been an increase in the use of the dimensions of data quality each year that we have conducted the survey.

1. **Most desirable** (Green ellipse above): The number of organizations which a single formally defined and governed list of dimensions has grown since 2015 (22.8% to 29.2%)
2. **Better than nothing** (Blue ellipse above): The number of organization with a standard- though not well defined and governed also has significantly increased (13.2% to 29.2%)
3. **Standardization and normalization continues** (Orange ellipse above): Since 2015, we've seen the number of organizations that have fragmented approaches become smaller (27.2% to 20.8%)

If you are currently using the Conformed Dimensions, in any fashion, consider telling us about it or presenting your organization's success at a data management conference or professional organization near you. Additionally, recommended blog topics are welcome and we'll be sharing updates throughout the next year.

Is your organization looking for Information Quality speakers for corporate events? Why not bring the author of this paper, Dan Myers (MBA/IQCP), onsite for outcomes-focused IQ training, leveraging the Conformed Dimensions of Data Quality and Information Quality Certified Professional (IQCP^{sm}) training material. Contact us: info@DQMatters.com

Appendix

Survey Methodology Information

Count of Full Responses: 48
Dates Survey was Open: April 1st, 2017 to May 2nd, 2017

This survey was administered online and advertised via LinkedIn (and LinkedIn groups, Twitter, CDDQ Website, referral and prior-year sign-up. It was conducted and promoted with a near-zero dollar.

There is likely a response bias- given that only respondents (organizations) aware of the dimensions of data quality concept may feel comfortable completing the survey, and they may naturally over-represent organizations that have already implemented a version of the dimensions.

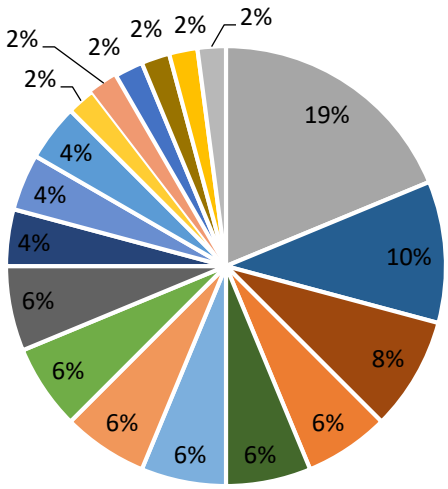
Dimensions Listed in the Survey (Options to Choose from)

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>

Usage of the Dimensions of Data Quality by Industry

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

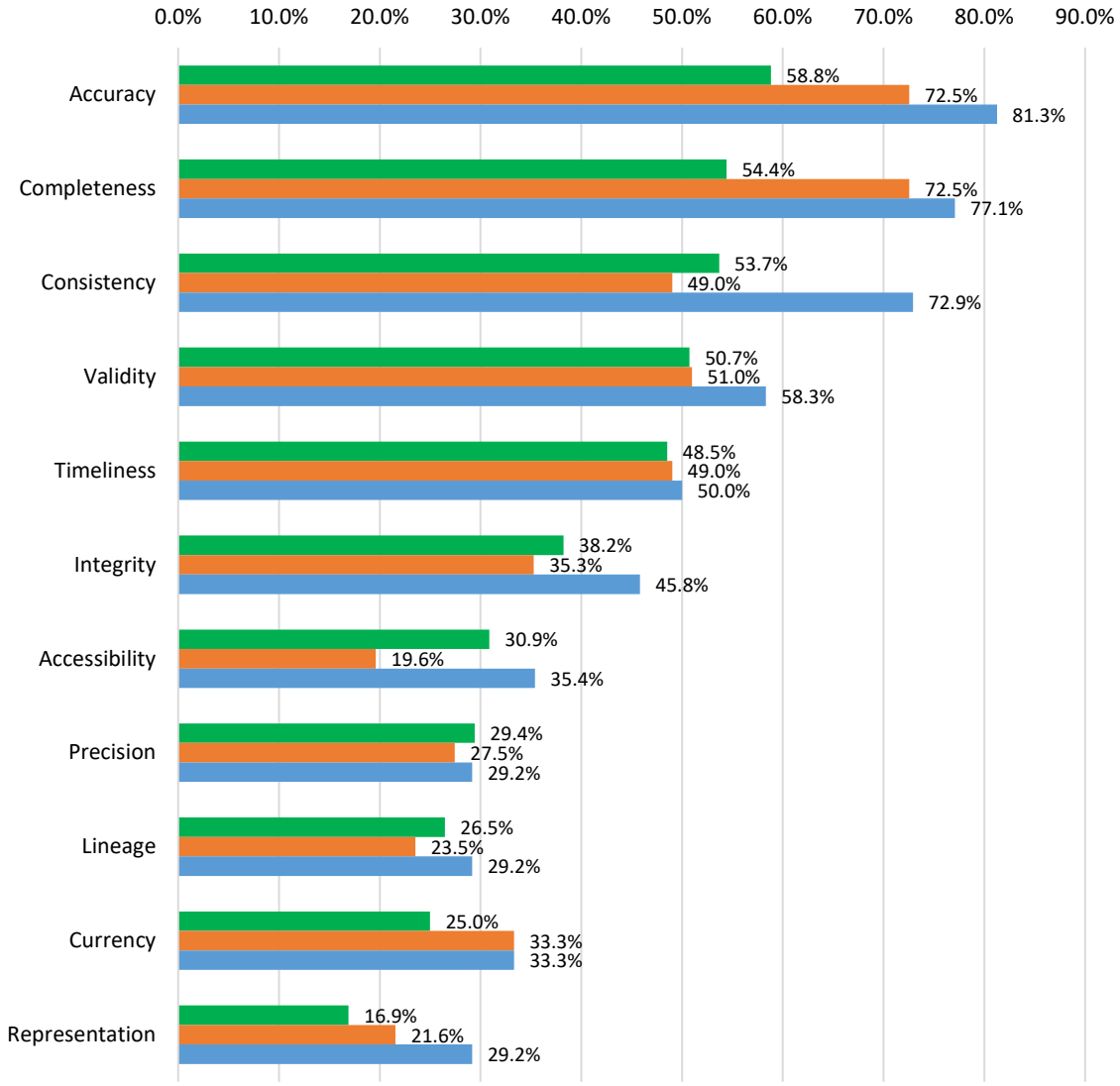


- Finance/Banking/Accounting
- Healthcare/Medical/Pharmaceutical/Biotech
- Insurance/Legal/Real Estate
- Consultant/Business Service
- Other
- Retail
- Software Development/Application Development
- Government – State
- Manufacturing
- Non-profit other than listed above
- Utilities
- Chemicals/Mining/Petroleum/Textiles
- Transportation/Logistics
- Entrepreneur
- Government – Local
- Media/Entertainment
- Government – Federal
- Telecommunications/Carriers

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

Change in Responses Over the Last Three Years

Figure 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 (first year) results 2015 2016 2017 *CC,BY-NB Dan Myers 2017*