# 2018 Annual Report on the Dimensions of Data Quality

**Year-four**

Go Green-
Reuse, Reuse, Reuse

September, 2018
The 4th Annual Whitepaper
Sponsored by

DQMatters
Data Quality eLearning

# Contents

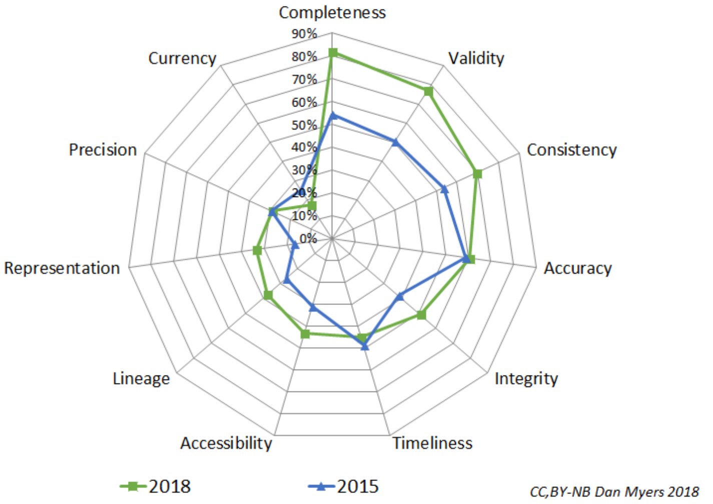
# Executive Summary

This year marks the fourth year that we have conducted the Annual Dimensions of Data Quality Survey, and new insights gleaned from your responses are driving strategic changes to the Conformed Dimensions of Data Quality (CDDQ) standard over the next year. The purpose of this survey is to measure the usage of the dimensions of data quality by organizations and related data quality topics. Below is a summary of the 2018 findings, but don't stop there, take the time to read through the details and sign up for the affiliated blog about the Conformed Dimensions of Data Quality in order to reap the benefits of detailed articles describing real-world data quality issues and measurement techniques using the CDDQ.

## Summary of Findings

- **Over 70% of the Dimensions of DQ are used more than in 2015**. As industry maturity continues, organizations are using a larger variety of the dimensions of data quality (right).
- **100% of organizations reporting excellent data quality use the dimensions of data quality.** Most of these have an enterprise-wide definition of each dimension in place.
- **37.8% of organizations taking the survey use the Conformed Dimensions in some form**
- **35% of organizations report that industry regulators require specific DQ metrics** and another 33% require self-selected measures of DQ to measure submissions (appendix 5).

**Figure 4d. Change in Use of Individual Dimensions Over Time**
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 respond



CC,BY-NB Dan Myers 2018

The last four years have shown high interest in using a standard set of dimensions of data quality (annual reports 2015-2017), so this year we asked, how many respondents were actually using the "Conformed Dimensions of Data Quality," and if they still weren't using them, we asked them why they were not. We found that 37.8% of organizations are using the CDDQ in some form and 3.8% are using them as is, out of the box so to speak. During follow-up interviews, we found that respondents want example metrics in order to speed up delivery and provide busy data stewards that don't want to reinvent the wheel. More about this in the *Exciting Updates* section of the report.

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

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

**Blog Signup:**
http://dqm.mx/addq18-cddqblogsignup

## Introduction

In 2015, a set of Conformed Dimensions of Data Quality (CDDQ) were drafted and published based on a literary review of six author/organization's versions of the dimensions of data quality[1]. Then in 2016, the first Annual Dimensions of Data Quality Survey was conducted asking organizations which of those definitions were used and whether anything was missing from the proposed CDDQ standard. Since that time, this survey has been conducted in April-May each year and the report released later in the summer.

**Request Copies of Prior Whitepaper Years Here!**

## Exciting Updates

1. **Metrics Available-** Last year this report observed that many industries are really missing out by not using the dimensions of data quality, so this year we asked what organizations need in order to better implement the dimensions of data quality. Respondents said that they would like example metrics (72%), but more specifically- industry specific example metrics (74%) in order to assist implementation within their organizations (See Figure 12a in Appendix 2). So as step one of two- we've release one example (typically generic in nature) metric for each of the CDDQ's Underlying Concepts. We hope that in the following years organizations will begin to volunteer industry specific examples- thereby completing the solution.

**Example DQ metrics can be accessed here:**

2. **Foreign Languages Available-** Additionally, because a majority of Information Quality literature has been published in English, there is added value to having at least the basic definitions of the Conformed Dimensions in other foreign languages. We've been working with practitioners and researchers to translate the CDDQ into other foreign languages in order to support their needs and expand the value of the CDDQ. So far, we have translations in Portuguese and German that have been published on the website and others are under development. If you're interested in helping in this area- please reach out to Dan Myers directly (Dan@DQMatters.com).

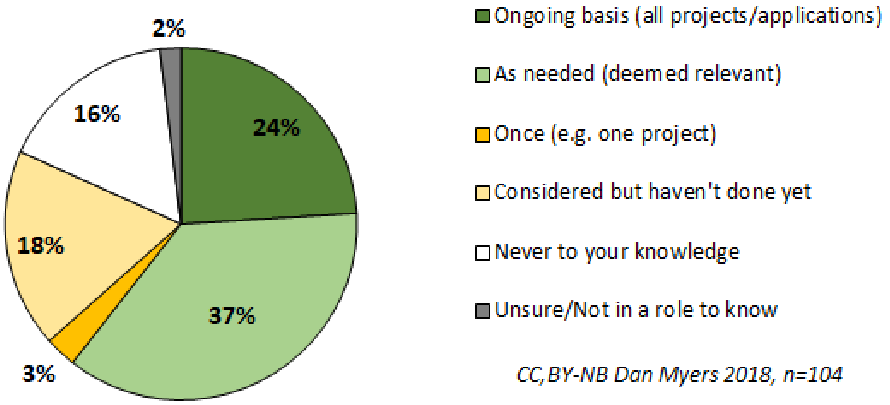
**Links to Foreign Language Translations are here:**

[1] Link to Dan Myers' 2013 articles titled, "Dimensions of Data Quality Under the Microscope" in Information-Management.com.

## Use of the Dimensions of Data Quality in General

In order to identify whether the use of the dimensions of data quality is subjective- meaning people may apply them inconsistently- we decided to add an additional option to the following question this year. “How often does your organization classify data related defects using the dimensions of data quality?” This year we added a 6th option titled, “As needed (deemed relevant)” and this caused almost half of the people that previously (2017) said they were used the dimensions in an “Ongoing basis” to choose this new category instead. Apparently, the application of the dimensions is still loose and may not be well governed. Arguably, there will be projects that are purely functional in nature- and do not warrant the use of the dimensions, but in today’s data focused economies those are few and far between. We recommend that organizations set clear guidelines for when the dimensions should be used and effectively educate end-users how to apply them so that they don’t opt-out of a potentially valuable exercise.

**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 2018, n=104*

The year-over-year version of this figure is available in Appendix 3.

***Do you have further insight or a different experience in your workplace? Tell us about it on the CDDQ LinkedIn Group here:***

Comparing the change in usage between 2015 and 2018, we can see increased adoption, and those answering the survey are knowledgeable on the topic[1]. Those that have only used them once has dropped from 13% to only 3%[2]. Optimistically, this means they do it more than just once now. Additionally, those that haven’t even tried it have dropped[3]- hopefully because they more have already tried their use- or even use them regularly.

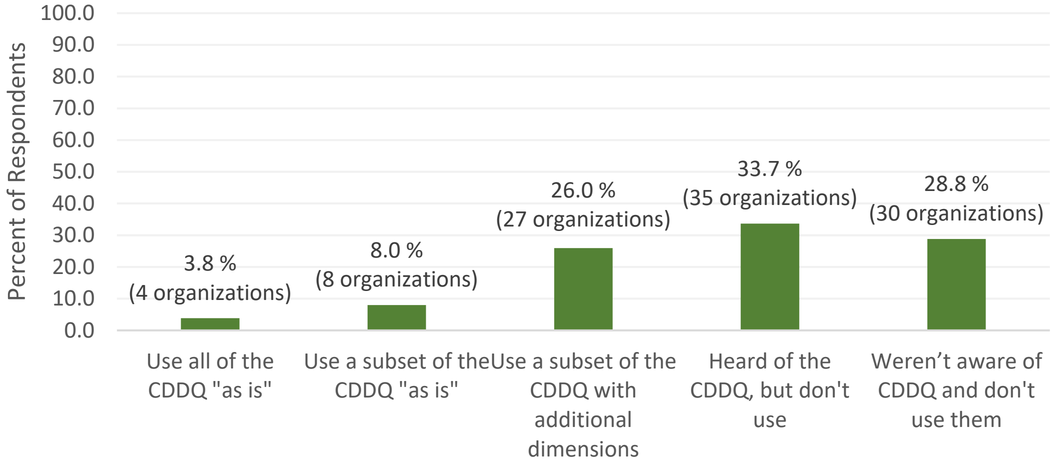
**Organizational use of the Dimensions of Data Quality to Classify Data Related Defects**

| | 2018 | 2015 | Change |
|---|---|---|---|
| Ongoing basis (all projects/applications) | 24% | 35% | -11% |
| As needed (deemed relevant) | 37% | n/a | +37% |
| Once (e.g. one project)[2] | 3% | 13% | -10% |
| Considered but haven't done yet[3] | 18% | 24% | -6% |
| Never to your knowledge | 16% | 16% | No Change |
| Unsure/Not in a role to know[1] | 2% | 11% | -9% |

## Use of the Conformed Dimensions of Data Quality

Until 2018, it hasn't really made sense to ask people whether they are using the Conformed Dimensions of Data Quality, because the standard was only proposed in proposed in 2016. In 2018, we saw that more than 11% of the respondents use either all or some of the Conformed Dimensions "as is" and an additional proportion (26%) use a subset of them in conjunction with other organization-specific dimensions (see Figure 6a, below).

**Figure 6a.** Does your organization use the Conformed Dimensions of Data Quality?



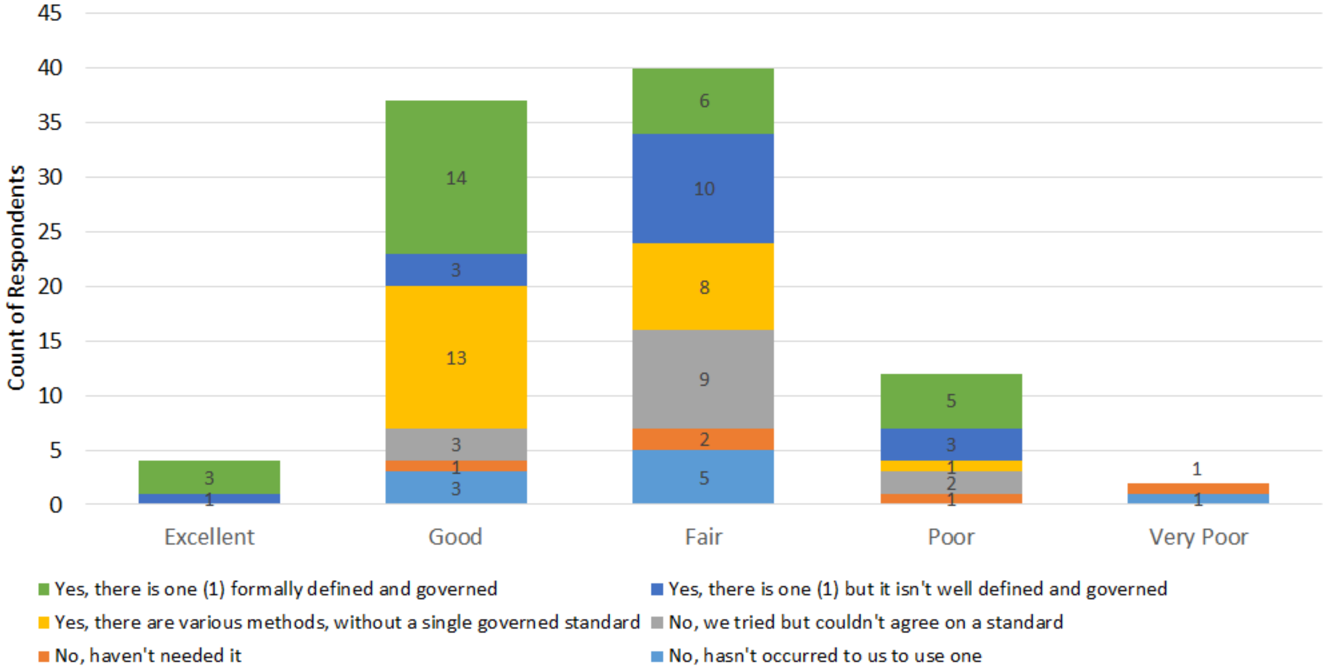
*CC,BY-NB Dan Myers 2018, n=104*

Given that the CDDQ is so new, this relatively high adoption rate is encouraging. Some of the respondents subscribe to the Conformed Dimensions Blog or have taken the survey in prior years, which would seem to explain the adoption. Based on the results, it's clear that a high percentage of organizations are using the CDDQ to compliment (add to) their own list of dimensions (26%). Readers of this report are encouraged to conduct a gap analysis between their existing list of dimensions with those of the CDDQ in order to identify areas of improvement. The definitions are provided at http://dimensionsOfDataQuality.com and the associated blog provides case studies regarding the use of each dimension. For those organizations that use the ISO/IEC 25012:2008 Data Quality Model's Dimensions of Data Quality, a gap analysis has already been done between the CDDQ and ISO dimensions and can be downloaded here (free to IQ International members).

## True Understanding of Data Quality Starts with Definition and Categorization

Survey results show that the best organizations (self-reporting “Excellent” data quality) have just one set of dimensions of data quality and most are well governed.

**Figure 3c.** Overlap between self-supported organizational data quality and the level of governance/use of a standard set of dimensions of data quality.



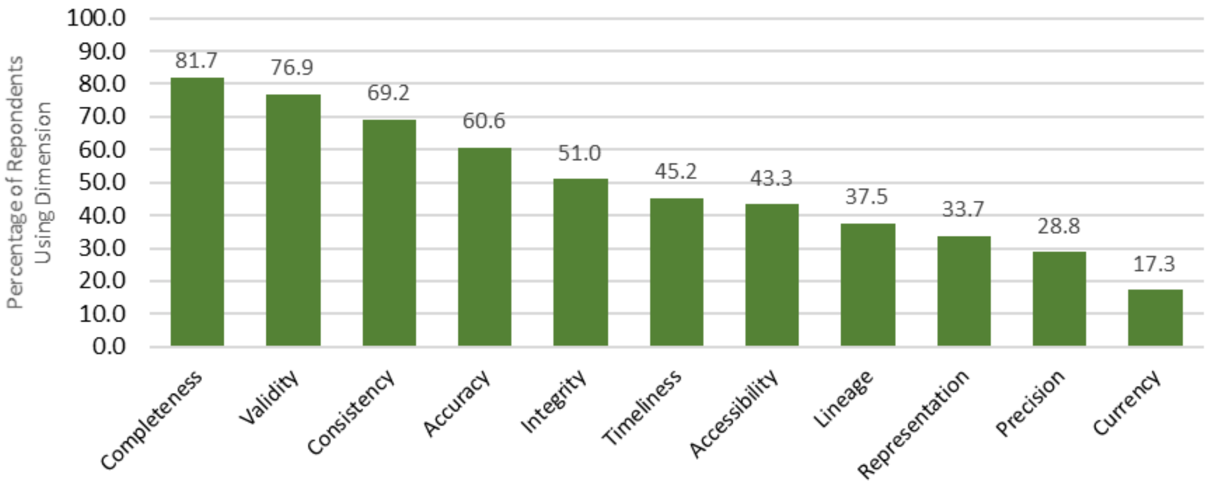
*CC,BY-NB Dan Myers 2018, n=104*

Organizations that can’t agree on a single standard set of definitions should use predefined dimensions and underlying concepts in order to move from their current levels to a higher level of quality. We do see that organizations with “Good” data quality more frequently have a single defined and governed standard. A large portion of those with “Fair” data quality also may have only one, but typically it isn’t well defined and governed.

Perhaps future surveys will need to take into account the length of time that the standard has been in place, how well they’ve been communicated and more, because we see that five respondents report “Poor” data quality in spite of having a single defined and standard set of definitions- which seems anomalous.

## Popularity of Each Dimension

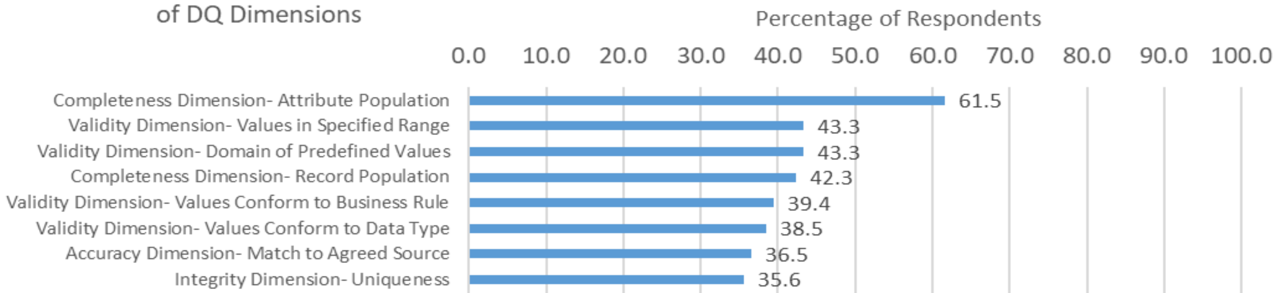
**Figure 4a (percentage).** Select any of the following dimensions currently used by your organization.



*CC,BY-NB Dan Myers 2018, n=104*

Each year we survey how widely each of the dimensions of data quality are used and this year we took that to the next level, by not just asking what dimensions, but specifically,  which Underlying Concepts are used. The 2018 findings reveal that when measured at the Underlying Concept level, Attribute Population is by far the most used area of measurement (61.5% of respondent organizations use it). See Appendix 1 for the complete bar chart.

**Figure 4f.** Ranking of Industry Use of the Underlying Concepts of DQ Dimensions



Generally, this can be attributed to the simplicity of explanation and large number of software tools on the market that offer this type of Null analysis per data column.

The changes in rankings from last year (right) show that there is a significant change in emphasis on accuracy- which we attribute to the two following reasons:

1. Accuracy is defined in two ways (see definitions below. Measurement is either focused on a real-world audit of the data, which can be costly, or a comparison to an agreed upon source of record. Although the latter is usually cheaper- it isn't always easy to agree what is a reliable source of truth and ensure that over time. The latter is effectively a Consistency measure. In 2018, the Consistency dimension (yellow) was the third most popular dimension used by respondents of the survey.

**Figure 4e.** Ranking Changes

| 2017 | | 2018 | |
|---|---|---|---|
| 1 | Accuracy | 1 | Completeness |
| 2 | Completeness | 2 | Validity |
| 3 | Consistency | 3 | Consistency |
| 4 | Validity | 4 | Accuracy |
| 5 | Timeliness | 5 | Integrity |
| 6 | Integrity | 6 | Timeliness |
| 7 | Accessibility | 7 | Accessibility |
| 8 | Currency | 8 | Lineage |
| 9 | Precision | 9 | Representation |
| 10 | Lineage | 10 | Precision |
| 11 | Representation | 11 | Currency |

*CC,BY-NB, Dan Myers 2018*

**Definitions:** Conformed Dimensions of Data Quality Accuracy Dimension- (release 3.5)

- **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.

2. The second reason that we attribute to the de-prioritization of Accuracy is that people often can't agree upon a definition for accuracy, but can define and relatively easily measure Completeness, therefore focus is put on Completeness. As argued in the past, Completeness is a foundational aspect that ensures that all the intended data is present. Typically, an organization's focus on key dimensions will change over time, but we'd hope that the usage is cumulative, rather than measuring one at the expense of foregoing the other.

## Do we need a different set of dimensions for each different industry?

In years past, the survey has asked whether respondents think that a separate set of the dimensions of data quality are required for different industries (e.g. P&C Insurance, Healthcare…etc.). In 2017, most respondents said that it was either Very Important (46%) or Somewhat Important (27%) that there should be an industry specific set of dimensions (see appendix #4). But until now, we haven't heard a good argument for a different set of dimensions, but rather different implementations of metrics based on the same dimensions and underlying concepts. Now that we've provided example metrics for each of the underlying concepts.

**For example:** the following two industries both can use the Completeness *Dimension* implementing the *Underlying Concept* of Attribute Completeness with two different metric names.

| Industry | Metric Name | Definition | Formula |
|---|---|---|---|
| Financial Services | Fill Rate | Rate of population of records for a given column | For a given column, the count of not null rows divided by the total number of rows in the set. |
| **Example** | A bank offers loan products, such as personal loans, credit cards, and mortgages. Customers are subject to a credit screening process to determine their repayment probability to support the extension credit. Once approved, the bank disburses the loan amount to the customer. Within the customer system, a table contains a "Funds Disbursed Amount" column which is dependent on the "Loan Approved" column flag. Therefore the "Funds Disbursed Amount" column will be NULL until the loan is approved by the Bank. The "Fill Rate" is the population of "Funds Disbursed Amount" divided by the number of loans approved (records where the "Load Approved" flag is set. | | |

Note that industry, metric name and definition can be worded differently, as long as the metric formula is the same. At some point, we guess that practitioners working across industries will desire a standard- even at the metric level.

| Industry | Metric Name | Definition | Formula |
|---|---|---|---|
| Retail | Column Population | For a given column, the count of not null rows divided by the total number of rows in the set. | For a given column, the count of not null rows divided by the total number of rows in the set. |
| **Example** | A retail company sells T-shirts, Diapers and Pants, but only directly ships pants to end-customers via its Website and T-shirts and Diapers are distributed through grocery store chains. The transactional table that lists sales of items has three rows in it (albeit small) with one T-shirt sale and one Diaper package sale and one pant sale. The DIRECT_SHIP column of this table stores a "Y" when the item is shipped directly to the end-customer. In this scenario we see that the DIRECT_SHIP column is 1/3 (33.3%) Not Null or has a Column Population of 33.3%. | | |

## Post-survey Interviews with Respondents

During follow-up interviews of some respondent's, key pain points were highlighted. Clearly each organization is at a different phase in their journey. The following topics were often discussed across by at least 3 of the 7 respondents:

- Data Quality Dashboards
- Automation (e.g. profiling or RPA)
- Organizational Silos and Communications Issues
- Collecting, and defining lineage
- Defining and normalizing KPIs across organization

**Banking & Financial Services**

Both interviewees in the banking area highlighted current efforts to automate data quality measurement using dashboards- one of which also frustrated with the complexity of DQ tool infrastructure setup. This organization finds it easier to implement stewardship due to the strong governance support based on compliance business cases, but they acknowledge inconsistencies in business KPIs and misunderstandings about how to communicate data quality issues (e.g. using the language of the dimensions of data quality). One respondent is looking forward to tagging existing DQ rules with associated dimensions in order to enable drill down based on similar dimensions. The other is interested in matching existing internal corporate survey results about subjective employee satisfaction with objective DQ rule results.

**Retail and Data Aggregation for Retailers**

Due to the physical nature of the retail space, where a tangible product is manufactured, displayed, and purchased, the concept of Accuracy often deals with how closely the labels on the product match the specifications provided by the retailer. Due to the sheer size of operations, samples are tested and error rates assumed for the population of products. Compared to financial services or other non-tangible product-based companies the retailers have an easier hurdle given these real-world auditable aspects of quality. Similar to those organizations discussed in other industries, they struggle with breaking down organizational silos. Even though one unit successfully implements tools and DQ metrics, they have been unable to gain executive support for cross functional implementation that would enable enterprise-wide cost savings and improved customer experience.

### Utilities

The single utility respondent (highlighted in last year's interviewee section as well) has continued to make progress in the Information Quality space, expanding quality concepts to measurement of lineage and timeliness. They built on the concepts of lineage measurement outlined in the Conformed Dimensions, specifically the use of *End-to-End Graphical Documentation*, and contextualized them for use in complex data acquisition workflows.

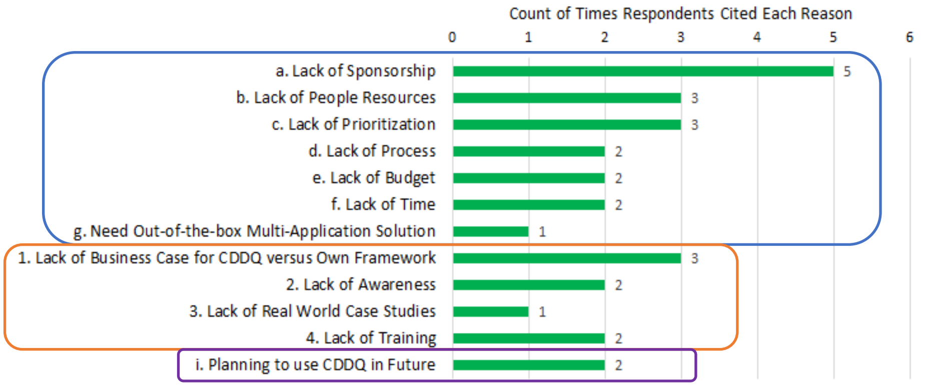
This organization understands the high value of measuring lineage, but is faced with the challenge of implementing this approach broadly across data acquisition processes. This organization has achieved a level of consistency of metadata (or "cross-walk") to identify the differences of data naming between systems, but the actual consistency of data moved under those names still has room for improvement.

Like the retail cases discussed above, some data quality management issues arise from the complexity of dealing with real-world phenomena that can only be measured via sampling strategies. Sampling methodologies are routinely accompanied by estimates of accuracy arrived at by widely accepted statistical methods. In the environmental area the standards and measurement processes developed by organizations like the US Environmental Protection Agency (EPA) leads to some level of conflict between the conformed dimensions of data quality vocabulary and the established terminologies and concepts of the scientific community for example the use of the *Precision, Accuracy, Representativeness, Completeness and Consistency – PARCC* as a set of quality dimensions and / or measures.

## Conclusion

Originally, the scope of this report was broad- and specifically designed to understand the use of the dimensions of data quality generically. With more attention and positive momentum toward the use of the Conformed Dimensions we've chosen to add more content focused on their use and lessons learned. This year we branched the survey- asking a new question of a specific set of respondents- those who said they **were** interested in using a standard version of the dimensions of data quality, but who **don't** currently use the Conformed Dimensions. We asked them what resources they needed in order to use the CDDQ? And we found some valuable information within their responses, as you can see below.

**Figure 6b.** You stated that you are interested in an industry standard version of the dimensions of data quality but then stated you don't use the Conformed Dimensions (CDDQ). What resources are needed to begin using the CDDQ within your organization?



*n=28, CC,BY-NB Dan Myers 2018*

- **Group I:** General Reasons Not Specific to the Conformed Dimensions
- **Group II:** Genuine hurdles that face the Conformed Dimensions Standard
- **Group III:** Those planning to use the Conformed Dimensions later

This list of answers was developed by grouping similar statements collected through a free-form text box field of the survey. Now, it is apparent that some of the answers have nothing to do with the Conformed Dimensions compared to generic dimensions of data quality developed in-house (group I outlined above in blue). Clearly the hardest part of any information quality initiative is gaining sponsorship that paves the way for the need for human, urgency, time, process, monetary, and time related investments.

It should be noted that although only one respondent mentioned that they want a software tool that includes the Conformed Dimension measures integrated within it (see 'g' above), we are hopeful that major DQ tool industry vendors like Informatica, SAP, Oracle, IBM, Talend…etc. will begin including a Conformed Dimensions library of predefined dimensions and metrics for their customers.

Group II includes a list of genuine hurdles that present themselves, such as, ‘if our company’s list of dimensions works, why change now’? This is valid question that has yet to be fully answered. Admittedly, the Conformed Dimensions offer the most value to those who are starting from scratch, but they also can be used when there is a lack of agreement about the definitions for dimensions- and only an outside 3rd party can offer an amicable choice that is perhaps mutually respected due to its adoption rate and level of documentation. Until now the lack of awareness of the CDDQ has been a challenge, but given the uptake this last year and the roll out of example metrics and definitions in other languages (Portuguese and German), prospective end-users are encouraged to reconsider the use of the standard.

DQMatters
Data Quality eLearning

We foresee that additional case-studies of the use of the CDDQ will be available next year in time for the report, and preferably presented by the implementing organizations themselves- rather than facilitators and stewards of the CDDQ. Although DQMatters currently offers training on the CDDQ, there is nothing stopping other vendors from offering training because the definitions and even example metrics are all Creative Commons licensed.

In conclusion, there are many positive things going for the Conformed Dimensions and organizations using them. If you aren’t already a part of the community building them out and strengthening the framework, please get involved. Check out the sample metrics here, and contribute your own here.

# Appendix

1. The 2018 findings reveal that when measured at the Underlying Concept level, the reason Completeness is so universally used is because of the focus on Attribute Population (see Figure 4f below).

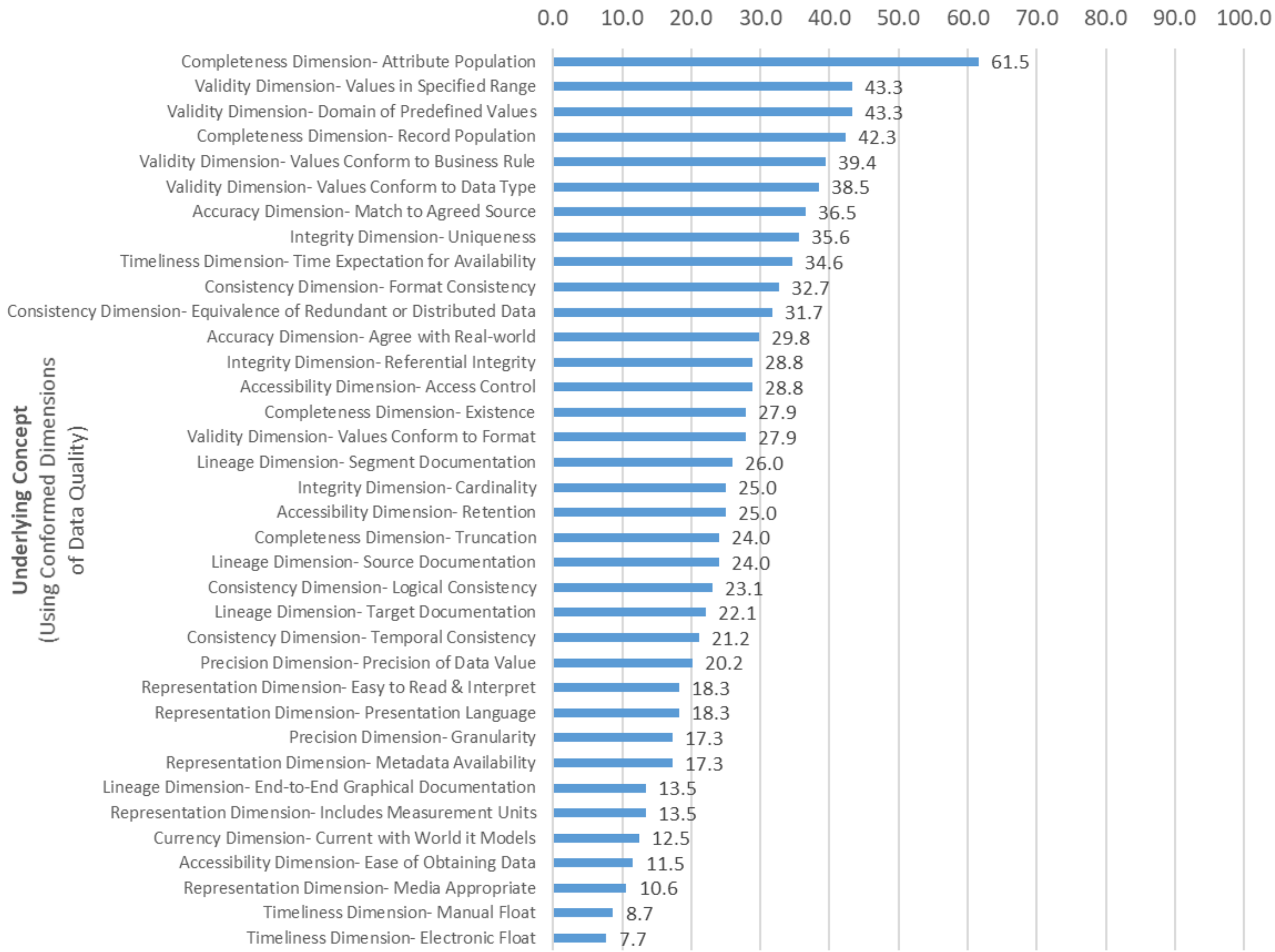


**Figure 4f.** Ranking of Industry Use of the Underlying Concepts of DQ Dimensions

*CC,BY-NB Dan Myers 2018, n=104*

2. Figure 12a

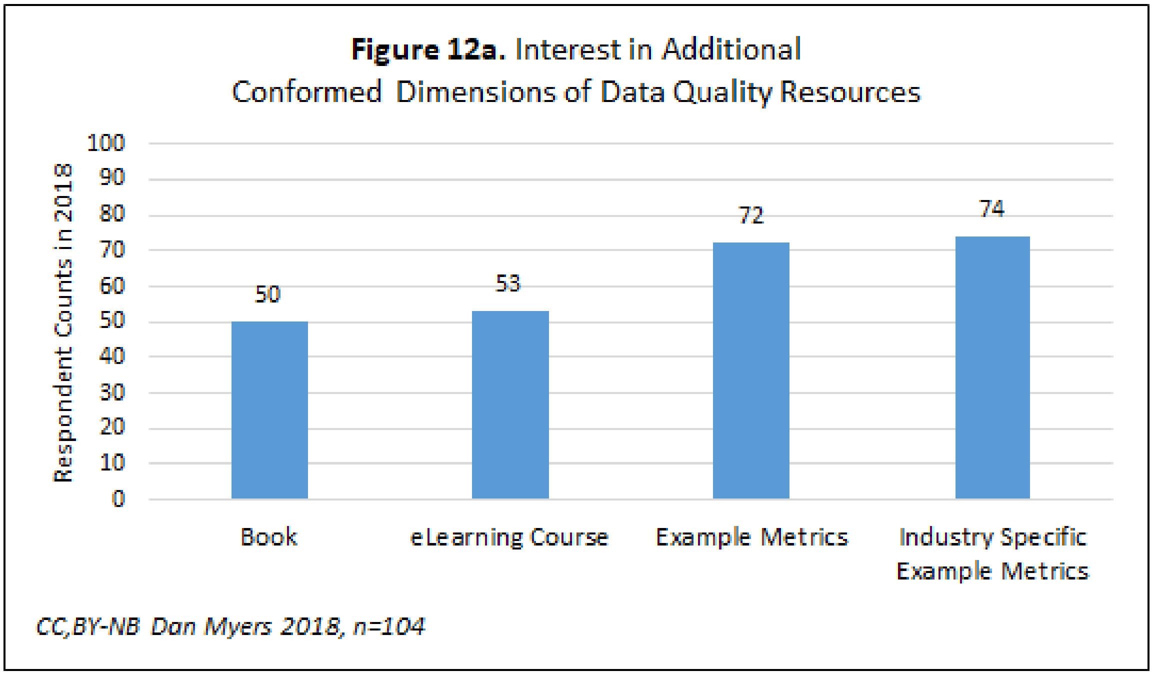


**Figure 12a.** Interest in Additional Conformed Dimensions of Data Quality Resources

*CC,BY-NB Dan Myers 2018, n=104*

3. Figure 2b

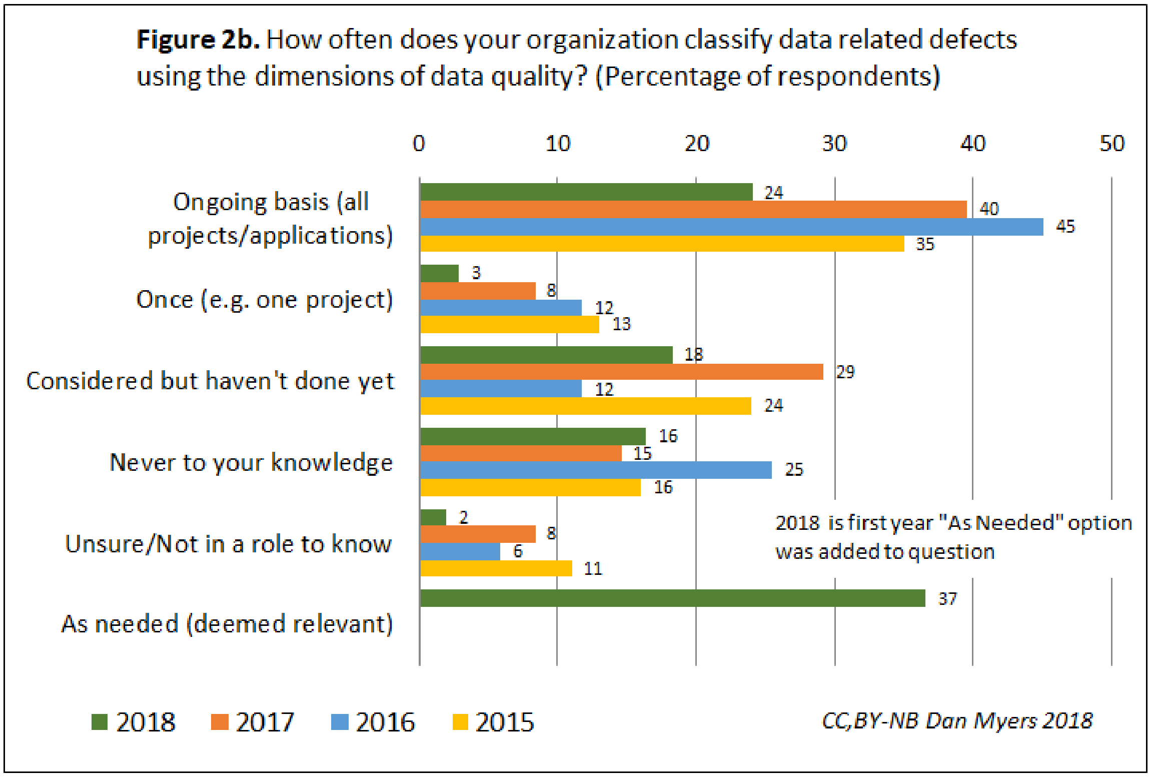


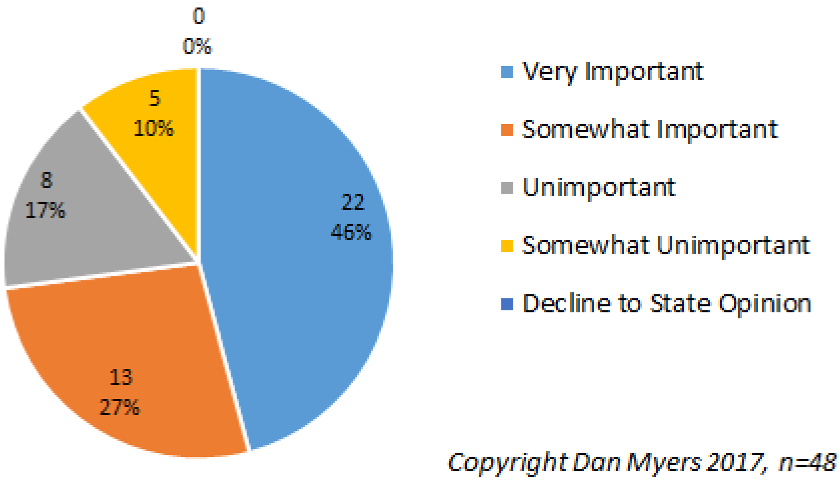
**Figure 2b.** How often does your organization classify data related defects using the dimensions of data quality? (Percentage of respondents)

2018 is first year "As Needed" option was added to question

*CC,BY-NB Dan Myers 2018*

## 4. 2017 Results, Figure 5

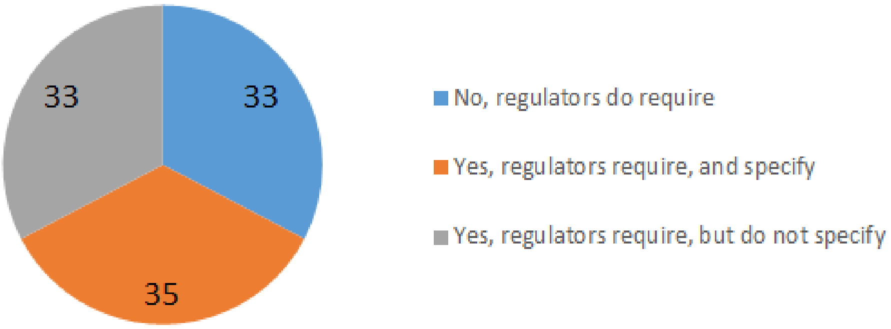
**Figure 5.** In your opinion, how important is it that the standard set of dimensions be designed specifically for your industry (e.g. healthcare, retail, banking...etc)?



*Copyright Dan Myers 2017, n=48*

## 5. Figure 15a

**Figure 15a.** Regulations Require DQ Measures



**Survey Question:** Do regulators/authorities in your industry have quantifiable measures embedded into rules and laws that govern the way your company/ organization conducts business and works, requiring you to submit reports/data? If yes, please identify w

*CC,BY-NB Dan Myers 2018, n=101*