2019 Annual Report on the Dimensions of Data Quality

Year Five- Opportunity Abounds for Competitive Advantage Based on Information Quality

November, 2019
The 5th Annual Whitepaper Sponsored by

DQMatters
Data Quality eLearning

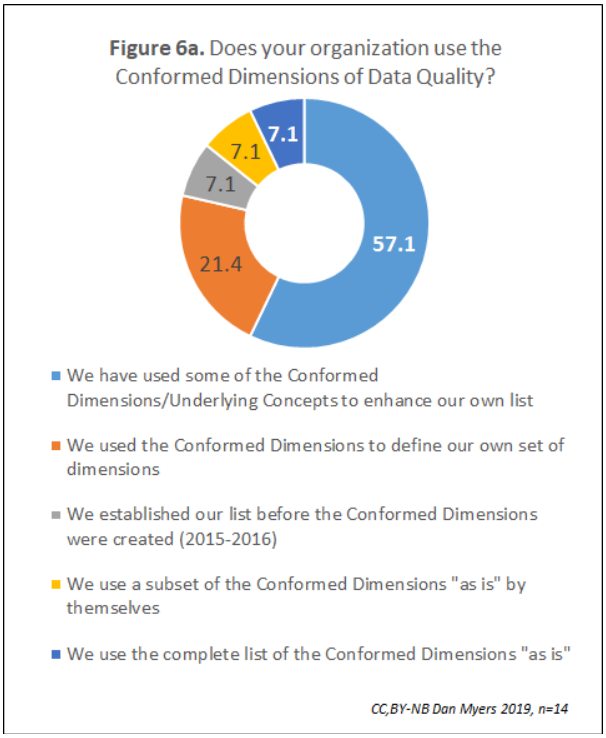
v1

Executive Summary

Survey Respondents Report that:

- 60% don't use the dimensions of data quality to categorize DQ issues
- If using the dimensions, 57% enhance their own list with the Conformed Dimensions
- 7% use the Conformed Dimensions without changes as is
- 40% report better DQ this year
- 14% decrease in the usage of Timeliness

This year's results include a mix of both promising and concerning trends in data quality practices in 2019. As highlighted in the chart on the right, many respondents (57%) are leveraging the Conformed Dimensions to improve their own DQ categorization, and as many as 7% use them out of the box.

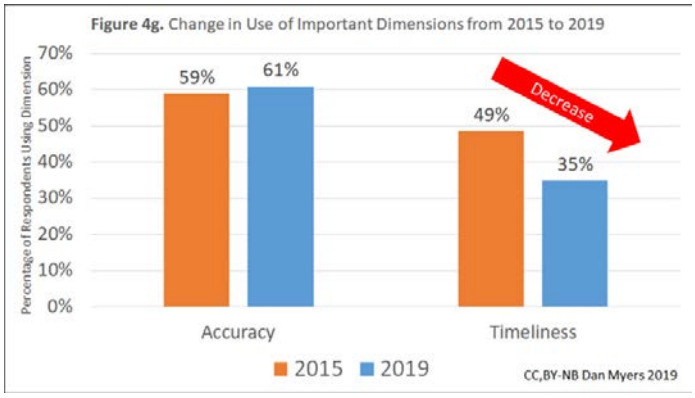


Unfortunately, about 60% of respondents don't use the Dimensions of Data Quality in any meaningful way, and we haven't seen this situation improve over the last five years we've conducted the survey.

Additionally, Our findings show that the frequency[1], and standardization[2] of the use of the Dimensions of DQ are not improving. Most data scientists are familiar with the descriptive power of the dimensions of DQ but this hasn't translated into organizational use of them during data processes.

Discuss the report findings here in the CDDQ LinkedIn Group

What's your opinion, and how did you convinced your organization to start using the dimensions of data quality? Talk about it here in the LinkedIn Conformed Dimensions group!

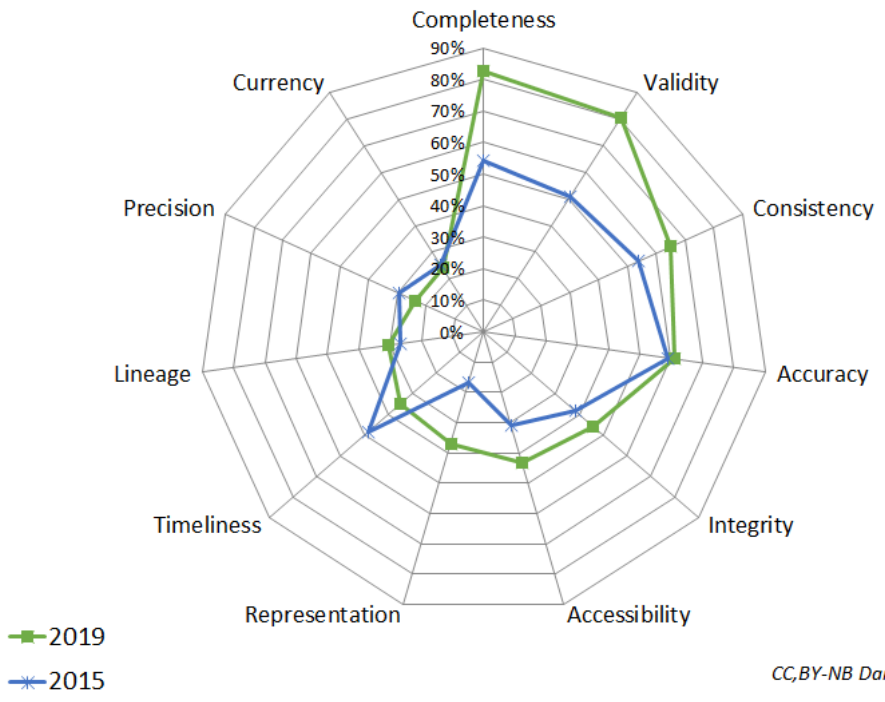


Wondering why your company just can't increase the "Accuracy" of the data? Well this year we discuss two under-utilized areas within the dimensions of data quality including the misunderstanding about **Accuracy** and the difference between **Currency** and **Timeliness**. As seen in Figure 4g, on the left, organizations have only slightly increased use of Accuracy and significantly decreased use of Timeliness and Currency. More on this in the body of the report. (Accuracy section, page 2-3)

Introduction

Every year DQMatters sponsors the Annual Dimensions of Data Quality Survey, and associated report, in order to measure the usage of the dimensions of data quality by organizations and related topics. Similar to prior years, there is good news about the increased adoption of dimensions that can be easily analyzed using data quality tools, such as Completeness, Validity, and to some degree, Integrity. Unfortunately, as highlighted in the executive summary, there hasn't been the same increase in the use of all of the other dimensions, such as, accuracy and timeliness/currency.

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. (% respondents)



Accuracy

As seen in Figures 4d (above) and 4g (prior page), Accuracy has not experienced the 20-30% increase in usage that Completeness and Validity have shown for the same period of time (2015 to 2019). We believe there are a few reasons for this. First, accuracy measures are often tied to in-person observations which are costly or those made at the time of the event. They are costly because they often require expert human collection. Second, often there is disagreement about which source (system) should be used as the system of record, or there is a lack of documentation about how the data was collected.

We believe there is a correlation between these two challenges and the nature of accuracy. Among data quality practitioners there are two primary components to the definition of accuracy:

1. **The data ties to the real-world situation-** This is used when there is a tangible, touchable object that can be observed in-person to validate the data quality.
2. **The data was recorded for a historical, or intangible concept that can't be investigated later in person-** This is the most frequent scenario as seen with intangibles such as monetary transactions, insurance/financial services, human communication, entertainment…etc.

Due to the nature of how we define accuracy, the first attribute (tied to real-world situation) often forces us to collect data in-person which is costly and takes a significant amount of time. Secondly, in cases where data is recorded (happened at point in time and therefore can't be observed in person), arguments begin to arise about what is the system of record or who's version of the truth to use. This is where DQ practice must work hand-in-hand with organizational data governance programs to define and enforce use of agreed upon sources.

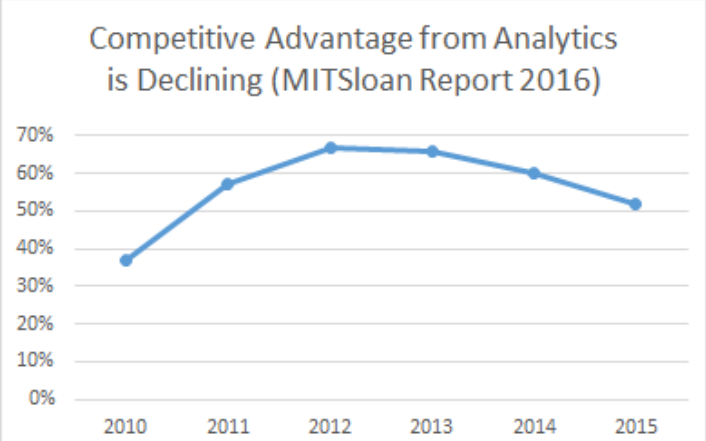
Too often the word accuracy has become a catch-all to describe data of high quality. Some people consider it to be an umbrella term including various aspects of quality. Looking forward, we expect that other tools for real-world measurement, such as Internet of Things (IOT) sensors, and object classification, using machine learning, will enable organizations to triangulate a better estimate of the real-world situation through validation from these new data sources.

How Can I Gain A Competitive Edge Using Data Quality Management?

With the rush to install BigData platforms, installed over the last five to ten years, companies have been disappointed with the analytical improvements they've seen despite early gains. Even as of 2015 the competitive advantage from analytics showed a declining trend according to the MITSloan Management Review in 2016 (see below)[1].

Given the close relationship between analytics and the quality of the underlying data we weren't surprised to find that corporate analytic strategies included three areas that intersect with data quality.[1]

1. Skills Development
2. Data Management
3. Cultural Norms for using data in decision making



Source: MITSloan Management Review, "Beyond the Hype: The Hard Work Behind Analytics Success", (Sam Ransbotham, David Kiron, Pamela Kirk Prentice), 2016.

[1] MITSloan Management Review, "Beyond the Hype: The Hard Work Behind Analytics Success", (Sam Ransbotham, David Kiron, Pamela Kirk Prentice), 2016

1. Skills Development

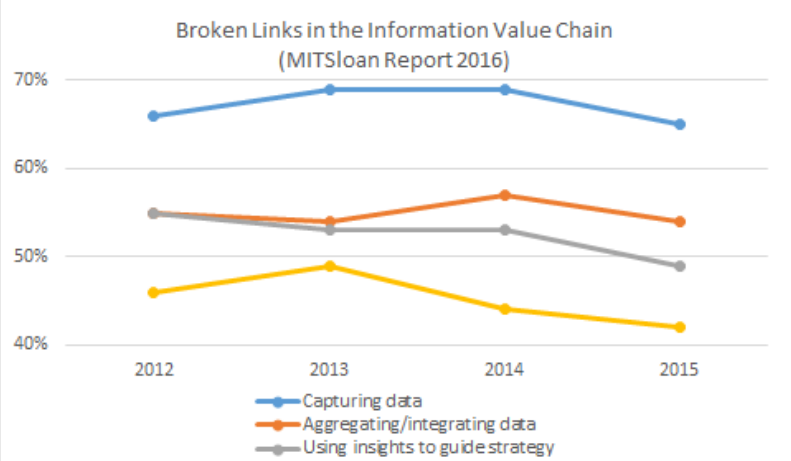
Nearly all data science educational programs include components on data acquisition and "data wrangling," but we still don't see strong organizational support for the use of the dimensions of data quality (see stats in executive summary). About 60% of our survey respondents don't use any method to categorize data quality issues. Clearly one of the areas of increase data science training, going forward, should be regarding data quality concepts and how to communicate expectations using the dimensions of data quality.

2. Data Management

The second area of intersection is Data Management, which includes programs/strategies, staff and tools used for:

- Data Quality Management (management techniques and tools to ensure fitness of use of data)
- Metadata Management (documenting where data is, where it's used, and how it's defined),
- Master Data Management (deduplicating, normalizing reusable domains of data such as customer, partner, product…etc.)
- Data Governance (Organizational system of decision rights and accountabilities)

There are additional areas[2], but just within these four (bullets above) we can see the intersection with the use of the Dimensions of Data Quality. First, as seen in the illustration below-right from the MITSloan Management Review, analytical teams are struggling to communicate data insights. Even though they have access to more data, and often can perform new analytics, they find it hard to disseminate the learning/knowledge throughout the organization.



Source: MITSloan Management Review, "Beyond the Hype: The Hard Work Behind Analytics Success", (Sam Ransbotham, David Kiron, Pamela Kirk Prentice), 2016.

Only 37% of the respondents of our survey said that they use the Representation dimension of data quality which covers key components of communication needed when sharing information[3]. This was actually an increase from when we began this survey in 2015 (when it was less than 20%, as shown in figure 4d in the Introduction).

[2] See DAMA DMBOK 2 Framework/Wheel, DAMA-DMBOK2 Framework, (DAMA International), Technics Publications, July 2017. Page 36.

[3] See Ranking of use of each dimension in Appendix 2 of this report.

Below are the Underlying Concepts within the Representation Dimension. See how many of these are consistently used when presenting information at your company.

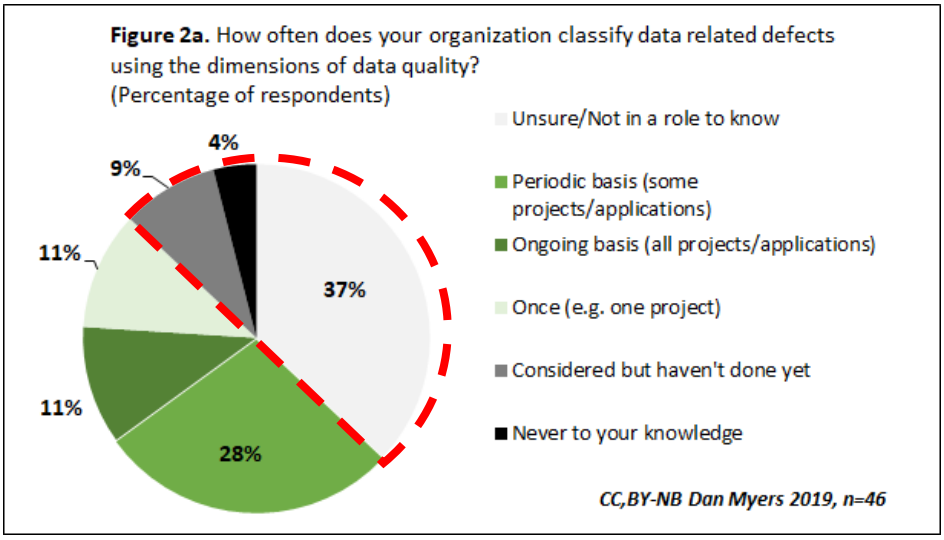
Underlying Concept Within Representation	Definition
Easy to Read and Interpret	Illustrations and charts should be self-explanatory and presented with appropriate labels, providing context.
Presentation Language	Data that is represented well is simple but elegantly formed with good grammar and presented in a standard way.
Media Appropriate	The appropriate media (e.g. Web-based, hardcopy, or audio…etc.) are provided.
Metadata Availability	Comprehensive descriptions and other information about the characteristics of the data are provided in plain language.
Includes Measurement Units	Well represented data includes the scale of measurement.

Get examples!

Clicking on the **title** of any of the Underlying Concepts (UC) above will take you to the **Conformed Dimensions of Data Quality blog** search- listing all of the blog posts associated with that respective UC.

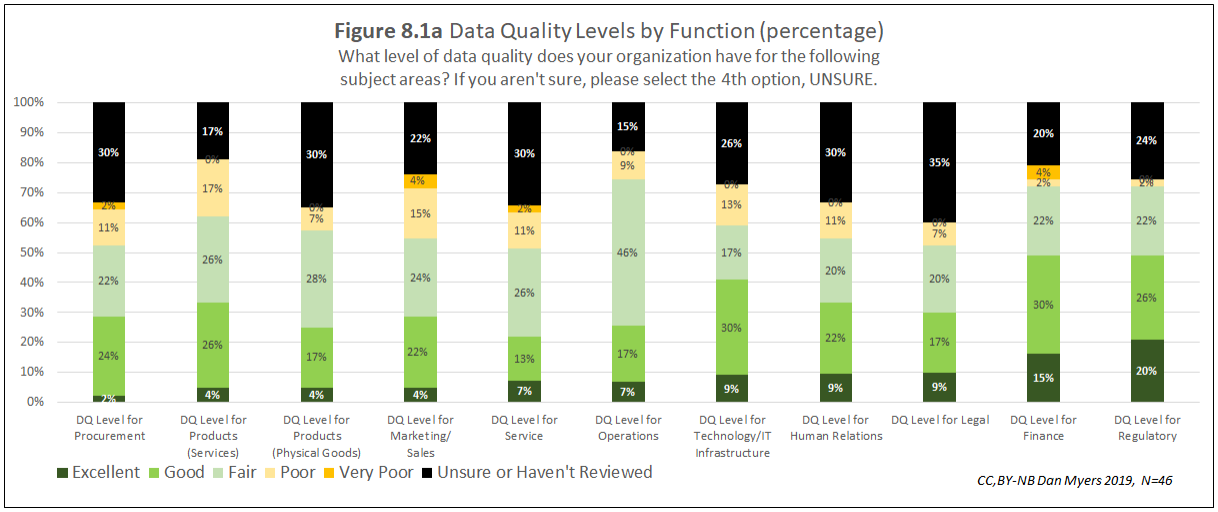
Usage of Dimensions of Data Quality

About 50% of respondents don't use the Dimensions of Data Quality in any meaningful way (see red dashed semi-circle below), and unfortunately, we haven't seen this situation improve over the last five years that we've conducted the survey. We don't want to be alarmist, but if your organization isn't using them, please start using them in some meaningful way. One way to implement this is via a phased approach- only requiring that newly collected data (or new systems) include measurement of DQ levels using the dimensions.



Data Quality Levels by Organizational Subject Area

This year was the first time that we asked respondents how they rate their data quality levels for each subject area in their organization. The highest quality data was found in: Regulatory, Finance, IT, HR and Legal subject areas.



Above, is a chart of the self-rated levels of DQ for each subject area (loosely aligned with departments). Clearly, the business prerogative is to focus resources on those areas most sensitive, either financially, or from a regulatory perspective. For example, we observed extensive regulation put into effect after the 2008 financial crisis that helped improve data quality at financial institutions. Additionally, lawsuits with financial implications that face the HR and legal departments (e.g. records management) drives some of the higher levels of DQ in those respective departments.

The interesting thing is that (other than procurement) the three departments with the worst data quality are revenue generating departments: Marketing/sales, Products (physical and services). Clearly there is an opportunity for companies to not only leverage data for better insight, but to leverage the quality of their data to beat competition to identify new customer needs and provide improve customer experience that drives revenues. Many of the CDDQ blog posts have provided examples of this.

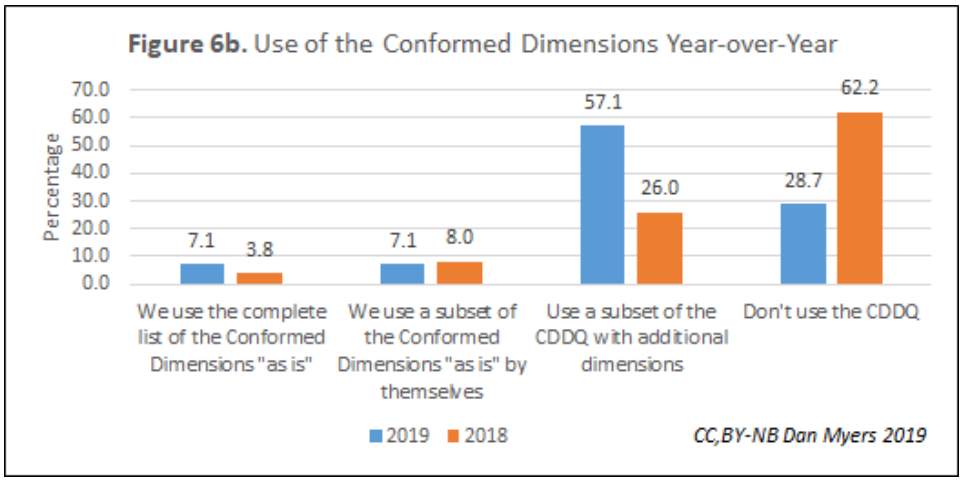
We assume that not every survey respondent will have a 360-degree view of their enterprise data quality levels, but a third of the respondents haven't even reviewed the DQ levels for five key departments (below).

1. Legal (35%)
2. HR (30%)
3. Service (30%)
4. Products -Physical Goods (30%)
5. Procurement (30%)

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

Use of the Conformed Dimensions of Data Quality

2018, was the first year that we asked respondents whether they use the CDDQ, so we were very interested to see how the numbers would compare year over year. Most organizations don't exclusively use the CDDQ, but use a subset of them to supplement their existing set (57.1%).

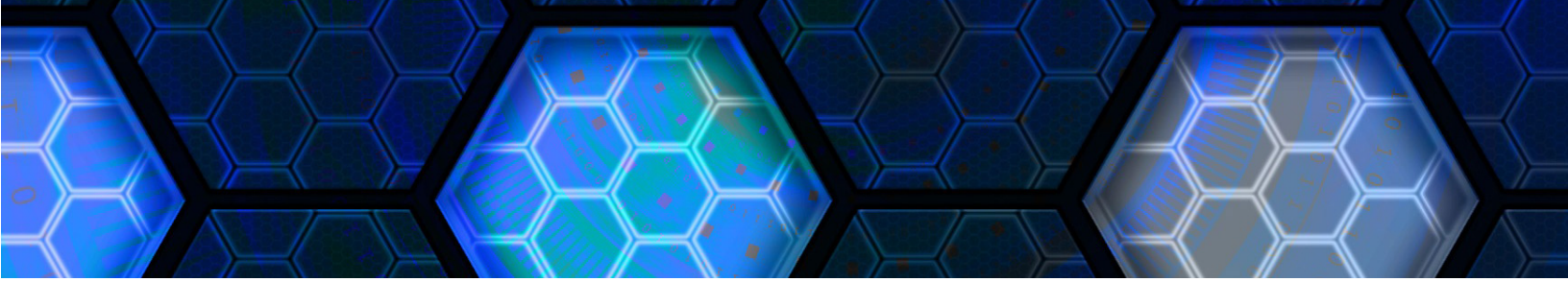


If you haven't already done the gap analysis between what your organization uses and the CDDQ – consider doing it today using the full list of definitions on the Website, and Example Metrics, published in 2018.

At the beginning of 2019, IQ International formed DQ metrics working groups with the goal of identifying and fully defining the most important DQ metrics used by DQ professionals across a broad range of industries. There are three active working groups (Healthcare, Telecommunications, and Manufacturing) as of the time this report was published. The healthcare working group has completed their first deliverable, including a list 23 recommended DQ metrics and a playbook explaining best practices for implementing DQ metrics in healthcare. All of the metrics are tagged by Conformed Dimension and Underlying Concept, so that implementation for organizations already using the CDDQ can quickly identify relevant metrics and recommended drill-paths for DQ dashboards.

IQ International

Is your organization an IQ International partner? If not consider joining as an organization (**more info here**). **Individual memberships** are also available and provide flexibility if you change roles often.



Conclusion

Each year, we identify the most interesting findings and add or change about 20-30% of the Annual Survey in preparation for the next year. Some of the core findings that we've measured over the last five years may not be included in the report each year, but we collect the data in order to provide consistent results over time. This year, we were excited to start collecting the subject area DQ health measures which enable much more detailed questions in those areas next year. Additionally, we're excited that more organizations have compared their internal definitions of the dimensions of data quality with the CDDQ. This has enabled them to enhanced their internal definitions using the CDDQ examples in the blog and example metrics.

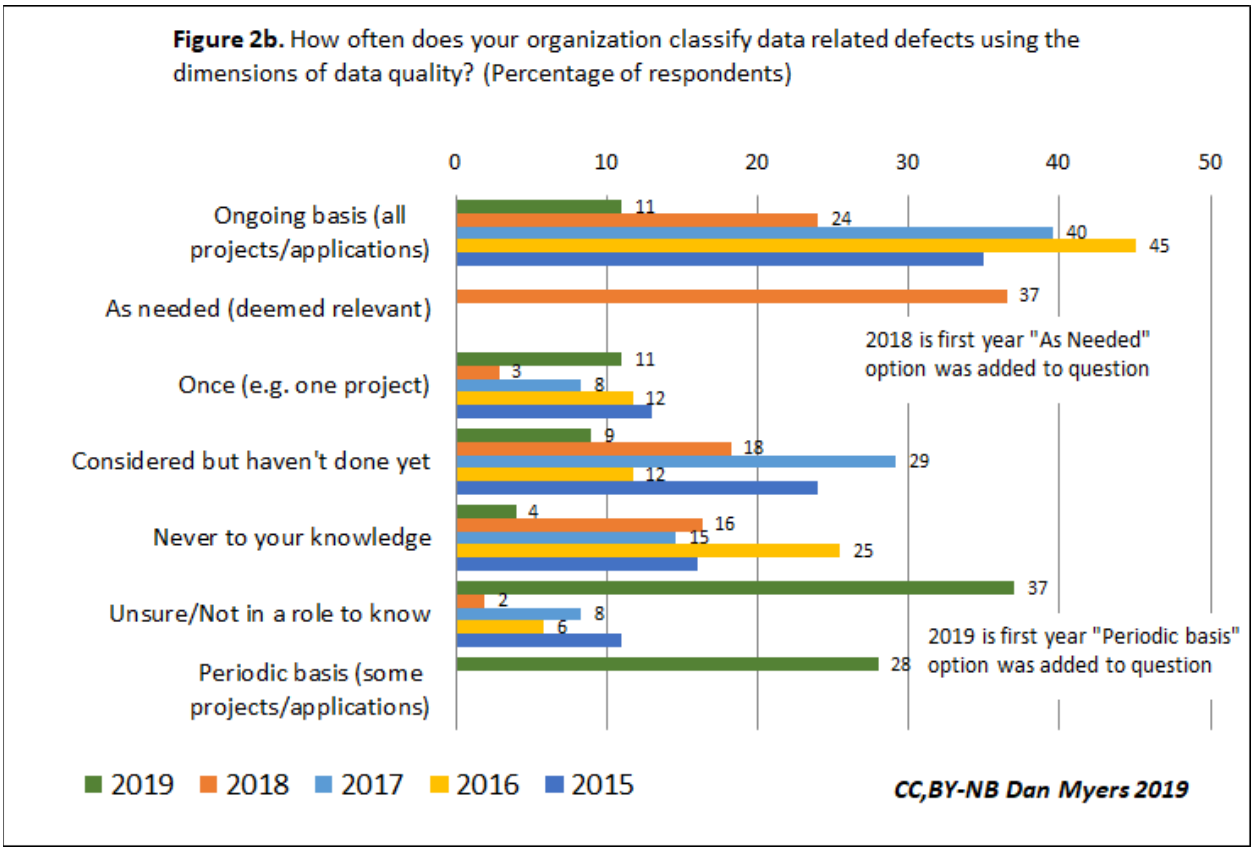
IQ International's use of the CDDQ in the DQ Metrics Working Groups is helpful to organizations that want to use an open standard (like the CDDQ) that is used by DQ professionals around the globe. If your organization isn't already participating with the working groups in some fashion, please consider doing so in order to represent your industry and gain access to highly qualified peers that are working to solve DQ challenges just like yourself.

Going into 2020, please post any questions about the survey on the LinkedIn CDDQ group and any ideas for new questions/focus areas for the 2020 survey. The CDDQ is a living standard, and as such, thrives and grows based on constructive review and ideas for improvement provided by you.

Discuss the report findings here in the CDDQ LinkedIn Group

Appendix

Appendix 1- Year Over Year Analysis of Use of Dimensions of Data Quality by Organizations



Appendix 2- Ranking of use of the dimensions of data quality (based on dimensions identified in the Conformed Dimensions of Data Quality).