ISSN 2688-6170

kcFED Technical Briefings

Discussing Data: Evaluating Data Quality

Thealexa Becker

November 2019

TB 19-03
https://doi.org/10.18651/TB/TB1903

Federal Reserve Bank of Kansas City

Discussing Data: Evaluating Data Quality

By Thealexa Becker[1]

Abstract

Data-driven organizations are becoming increasingly aware of data quality in order to mature their data activities. Data quality is of particular concern to research functions, but many existing frameworks are not well suited for use by researchers. This paper discusses existing data quality frameworks and focuses on one that meets the needs of research functions. A modified version of that framework is described along with details for use with a wide range of data.

[1] Data Scientist, Thealexa.Becker@clev.frb.org, Federal Reserve Bank of Cleveland

The Federal Reserve Bank of Kansas City, like many organizations, has become more data-driven. Maturing as a data-driven institution often means creating a data management strategy that addresses all aspects of an organization's data activities. One critical part of any data management strategy is data quality. There are many frameworks for evaluating data quality in an organization, but many do not incorporate some particular considerations of research functions.

Research functions within organizations like the Federal Reserve Bank of Kansas City employ researchers with training to understand and address shortcomings in their data when producing research. This paper presents an adapted framework for evaluating data quality that accounts for specific needs of research activities. This framework begins by defining the components of data quality and then suggesting how they might be assessed. It then provides a standardized rubric for measuring each component of data quality and creating a composite data quality rating. Finally, the paper discusses how users can apply these ratings when beginning a project.

Choosing a data quality framework

The Data Management Body of Knowledge, otherwise referred to as the DAMA-DMBOK, is one of the common data management reference guides for individuals and organizations employing a data management strategy. The DAMA-DMBOK provides several potential frameworks for observing data quality, each one with a slightly different focus but similar core components. Overall, the frameworks present variations on how to view data quality underpinned by a common set of data quality dimensions.

When considering which framework is best suited for use in a research function, we first need to evaluate how information about data quality is used and perceived. Research functions often manage a large number of data sets, many acquired from third party sources. These data are not all necessarily in the same sub-field but may be used together for analysis. Additionally, research staff are often well-equipped to evaluate data quality in a non-standardized way. In many cases, their evaluation of a given data set's quality will make its way into their analysis. Data that are not well-regarded by the research community in terms of quality are often cautioned against while alternative data sources are suggested in their place. There is an unstructured and built-in system of passing on information about data quality among the research community. Well-known and high-quality data are used often, poor quality data are used irregularly. Choosing from the frameworks presented in the DAMA-DMBOK is a matter of aligning the needs of the line of business with the focus of the framework.

The DMBOK highlights four frameworks for data quality: the Strong-Wang framework, the Redman framework, the English framework, and the DAMA UK framework. Each of these alternative frameworks judges data quality through a different lens. The Strong-Wang framework views data quality from the perspective of a consumer, with a focus on perception. There are multiple dimensions identified in this framework that relate to how easy a consumer would find it to work with the data. The Redman framework, on the other hand, is more concerned with data quality as it relates to the structure of the data, with a strong focus on detailed compositional aspects of the data. The English framework, while most similar to the DAMA UK framework, puts forward data quality as a set of characteristics that are either inherent or pragmatic. The pragmatic characteristics exist to allow the evaluation of data based on its intended use. One, in particular, a white paper for the DAMA UK workgroup, lists six core dimensions of data quality along with several other characteristics that are a good fit with how research functions think about data quality. The eight components of data quality addressed in this paper are adapted from this framework. It is worth understanding why this framework is more appropriate than several other contenders presented in the DMBOK, since, as the DMBOK itself notes, there is no "correct" choice.

There are a few reasons why the DAMA UK framework is best suited for use in a research function, given these above considerations. First, this framework is concise, consisting of only a few components to evaluate. This makes it less time-consuming for staff in a research function to evaluate the data they manage. Second, the components in this framework are general enough to apply to most data being used. Given that much of the data being used in research will have varying formats, having relatively general components that can apply broadly makes this framework flexible. Finally, this model is a standardized way of conveying the same concepts researchers talk about in an unstructured way. The components in this framework reflect many statistical and practical considerations often discussed in analysis about the pros and cons of data use without veering too far into technological or business use concepts. In brief, this framework is a concise, flexible standardization of common research concepts that does not delve too deeply into tangential concerns more appropriate for another business function.

Components of Data Quality

Rating the quality of data with a single word is not useful. A user who describes a dataset as having "poor" quality might be referring to a specific aspect of the data they found lacking. A different user may not be concerned with that deficiency, making this simplified rating ineffective. Conveying a more complete evaluation of data quality requires standard guidance.

Data quality can be broken down into a discrete set of measurements[2]. Four are objectively measured data quality dimensions: completeness, consistency, uniqueness, and validity. Objective

[2] Two components of data quality from the source framework are not addressed: value and confidence. While these concepts might be more appropriate to consider in a different line of business, they are not as useful for a research function. The data having value and users having

measurements can be taken using a clearly defined metric, usually rooted in a formula or coding procedure. Four are subjectively measured data quality dimensions: accuracy, flexibility, timeliness, and usability. Subjective measurements require judgment from subject matter experts because although most users might agree in general, more thorough knowledge of the data is needed to assess.

Below are the four objective and four subjective components of data quality defined:

Completeness

Definition: The proportion of data that is stored against the potential for 100 percent of the data to be stored. In other words, is there missing data?

How to measure: Most software used to handle data have the capabilities to check to see if there are missing data. Missing observations are more difficult to capture, but in some cases, they can be accounted for if you are able to compare to a full list of respondents.

Consistency

Definition: Whether or not, given a definition, data would represent the same object the same way.

How to measure: Most data sets should come equipped with a data dictionary. Such a document would detail the definitions of all variables in the data. Consistency can be measured by how often the definition of a given concept changes throughout time.

confidence in the data are the minimum threshold for acquisition and use. If there is not confidence or value, the data simply is not used.

Value is not a useful component of data quality to use in a research function for two reasons. First, the notion of "value" is subjective in research. Second, data are often obtained from third party producers for purposeful use in research activities. Therefore, if data are no longer "valuable" from a cost perspective, contracts to use the data are simply not renewed.

For similar reasons, confidence is not as useful a measure for use in a data quality framework. Since most data are acquired externally for research functions, confidence lies more with the data producer rather than the data product itself. As research functions often do not control the production of acquired data, if the confidence in external data is too low, it just will not be used or considered at all.

Uniqueness

Definition: The degree to which the data does not contain multiple entries for the same object, based on how the object is defined. In other words, no duplications.

How to measure: Most software used to work with data contain a command or sequence of commands designed to root out duplicate entries. What is needed is a clear definition of what constitutes a duplicate entry.

Validity

Definition: Whether or not the data conform to a defined domain of values. For instance, if answers to age are filled with whole numbers greater than 0 and less than 120.

How to measure: Data dictionaries should include value ranges for data. Data can be assessed against those value ranges to determine what percentage of entries fall outside that range. Many statistical programs are equipped with features to accomplish and track this.

Accuracy

Definition: The degree to which the data correctly describes the object of the data.

How to measure: Accuracy is arguably the hardest dimension of data quality to measure. It can be difficult to know what the "right" answer or entry in the data is. Accuracy is best measured by spot-checking against known data.

Flexibility

Definition: The degree that the data are comparable with other data, compatible with other data, or can be repurposed for a use different than its stated objective.

How to measure: What data a given user might want to combine could vary. How flexible a given data set is depends on how big the pool of data it can be effectively combined with is.

Timeliness

Definition: How likely the data are to change or be updated, and for what reason. This could also refer to if the data are "up to date," or, have a lag of data release. For instance, if it is 2020 but the most recent release is from July 2015, then the data are not timely.

How to measure: Most data have a release or update schedule. Comparing the frequency of data updates with the frequency of use and how recent the latest release is will determine the timeliness.

Usability

Definition: The degree that the data are relevant, accessible, maintainable, and simple at the desired level of precision.

How to measure: This is a subjective measure, as the experience level of a given user might impact the assessment of usability. Think about three ways in which data might be difficult to use:

1. Technology – Are there technology limitations that make the data difficult to use?

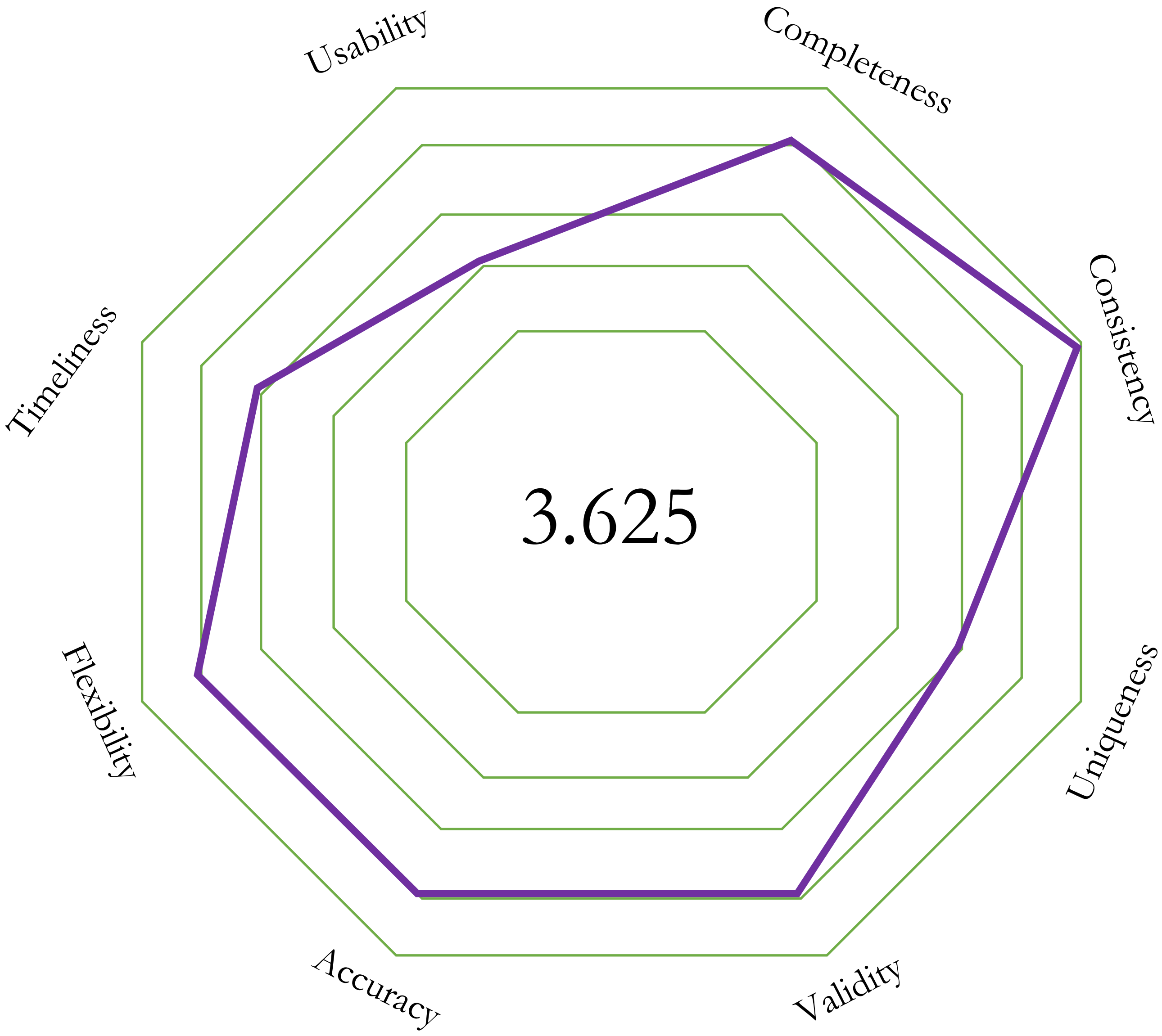
2. Format – Does the data come in a format that makes it difficult to use?
3. Content – is there information that is difficult to find or nonexistent that make the data difficult to use?

This dimension would be measured in more of a general sense. If there are a number of users concerned about the same type of usability across experience levels, perhaps the data are not in a usable format.

Data Quality Rubric

This section provides a rubric for rating each data quality component on a scale of 1 to 5 with 1 denoting poor and 5 denoting excellent. For a single contributed rating, these eight ratings are averaged to create the composite data quality rating. For multiple contributors, the ratings for each component are averaged to create a composite score for a given component. Those eight composite component ratings are then averaged to create a composite data quality rating.

The ratings can be visually represented in the octagon chart example found below. Each point on the octagon represents a component of data quality. The purple, superimposed octagon represents the rating of a dataset for each component. The average of these ratings produces the score, represented in the middle.



Ratings should be assigned using the following criteria (and subject matter expertise, as appropriate):

Completeness

1- Most data records are incomplete, including critical information.
2- Many data records are incomplete, but critical information is present in most cases.
3- Some data records are incomplete, but all critical information is present.
4- A few data records are incomplete, but all critical information is present.
5- All records are complete.

Consistency

1- Data have no consistent definitions.
2- Data have some consistent definitions.
3- Data have a majority of consistent definitions, but the inconsistent definitions are difficult to reconcile.
4- Data have a majority of consistent definitions and inconsistent definitions can be reconciled.
5- Data either have completely consistent definitions or all inconsistencies have been reconciled.

Uniqueness

1- The data are replete with duplicate entries that are impossible to cull from the data.
2- The data contain many duplicate entries that can be removed.
3- The data contain some duplicate entries that can be easily removed.
4- The data contain a few duplicate entries that are easily removed.
5- The data contain no duplicate entries, or they have all been removed as a routine practice.

Validity

1- A large share of data does not fall in acceptable value ranges and the data cannot be corrected.
2- A large share of data does not fall in acceptable value ranges, but some of the data can be corrected.
3- Some data do not fall into acceptable value ranges, but some of that data can be corrected.
4- Some data do not fall into acceptable value ranges, but all of that data can be corrected.
5- All data fall into acceptable value ranges.

Accuracy

1- Data are not accurate.
2- Data have a range of significant accuracy issues.
3- Data have some accuracy issues.
4- Data are mostly accurate.
5- Data are as accurate as possible.

Flexibility

1- Data are not able to be combined with any other data.
2- Data can be combined in specific instances with a few other data sets.
3- Data can be combined in a high-level manner with a few other data sets.
4- Data can be combined with a few other data sets.
5- Data can be combined with many other data with ease.

Timeliness

1- Data are not released in a timely manner and have no release schedule.
2- Data are not released in a timely manner and are not released on a regular schedule, although a schedule exists.
3- Data are not released in a timely manner, but they are released on a regular schedule.
4- Data are released with a lag, but the data are released on a regular schedule.
5- Data are released in a timely manner on a regular schedule.

Usability

1- There are significant technology, format, or content limitations that prevent basic use.
2- There are technology, format, or content limitations that hinder general use.
3- There are technology, format, or content limitations that make general use more difficult but can be overcome by an expert user.
4- There are technology, format, or content limitations that make advanced use more difficult but do not prevent basic or general use.
5- Technology, format, or content limitations are either fully addressed or are easy to overcome by a novice user.

Using Data Quality Ratings in Practice

Not all data quality issues can be resolved, but that does not prohibit the effective use of the data. Measuring data quality is as much about awareness and identification of problems in the data as it is about finding solutions to those problems. Providing a complete picture of data quality will help future users make decisions based on their needs.

A researcher or analyst beginning work on a dataset will be at a distinct advantage if they have information about that data's quality. Pre-processing data for use, often one of the most time-consuming parts of analysis, is necessary because the extent of data quality concerns may be unknown and unaddressed. A user who has a composite data quality rating for a dataset can focus pre-processing work on identified concerns.

Beyond using data quality ratings to inform data pre-processing, these ratings provide a natural means of comparison between similar data sets. Many times, researchers or analysts looking to answer a particular question have several data source choices. If this area of analysis is new to the user, it can be difficult to compare data options. However, if these data sets have data quality ratings, a user would be able to understand, at a high level, some of the advantages or challenges in working with each data set.

For organizations that manage multiple data sets, data quality ratings are a useful resource for reference purposes. In addition to any documentation about the data, a well-maintained data quality rating can be provided to prospective users before beginning analysis. At the conclusion of any analysis, those users can then provide their own assessment of the data and update the ratings. In this way, an organization can create a crowdsourced measure of the experience of working with a given data set. With this information, an organization can make decisions about continuing to use that data, making substantial improvements to the data, or recommending data as a go-to resource.

Conclusion

Incorporating data quality ratings for known data sets is useful for both individuals and organizations. Standardized data quality ratings allow individual users a means of both comparing data and understanding inherent issues in the data that need to be addressed in the analysis. These same ratings can be maintained by organizations in order to recommend and inform staff on data use for analysis.

References

DAMA UK Working Group. 2013. *The Six Primary Dimensions for Data Quality Assessment: Defining Data Quality Dimensions.* United Kingdom: DAMA UK. Available at https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf

DAMA International. 2017. *DAMA-DMBOK: Data Management Body of Knowledge.* 2nd ed. Basking Ridge, NJ: Technics Publications.