Appendix B
Data Quality Dimensions

Purpose

Dimensions of data quality are fundamental to understanding how to improve data. This appendix summarizes, in chronological order of publication, three foundational definitions of data quality dimensions: those of Richard Wang and Diane Strong, Thomas Redman, and Larry English. These provide context for the choices in the DQAF. In the DQAF, I have not proposed new dimensions of data quality. On the contrary, I draw a subset and have narrowed their scope to define objective measurements that can be taken from within a dataset.

Richard Wang's and Diane Strong's Data Quality Framework, 1996

In the article, "Beyond Accuracy: What Data Quality Means to Data Consumers," Wang and Strong present results of a survey conducted to understand data quality dimensions from the point of view of people using data. Their starting assumption is that "data consumers have a much broader data quality conceptualization than IS professionals realize" (p. 5). In summarizing previous work on data quality dimensions, they point out the limits of both an intuitive and the theoretical approach to data quality: Both focus on development characteristics rather than use characteristics of data (p. 7).

Wang and Strong define *data quality* as "data that are fit for use by data consumers," and they define a *data quality dimension* as "a set of data quality attributes that represent a single aspect or construct of data quality" (p. 6). To establish their dimensions, they first collected a set of 118 attributes of data identified by consumers themselves. Next they asked survey respondents to categorize the characteristics and to rate their importance in relation to data use. Wang and Strong performed factor analysis on the results and re-surveyed to understand the association of dimensions with categories of data quality. The result is a conceptual framework of 15 data quality dimensions related to four general categories of data quality: intrinsic, contextual, representational, and accessibility data quality.

- Intrinsic DQ denotes that data have quality in their own right; understood largely as the extent to which data values are in conformance with the actual or true values. Intrinsically good data is accurate, correct, and objective, and comes from a reputable source. Dimensions include: accuracy objectivity, believability, and reputation.
- Contextual DQ points to the requirement that data quality must be considered within the context of the task at hand, understood largely as the extent to which data are applicable (pertinent) to the task of the data user. The focus of contextual DQ is the data consumer's task, not the context of representation itself. For example, contextually appropriate data must be relevant to the consumer, in terms of timeliness and completeness. Dimensions include: value-added, relevancy, timeliness, completeness, and appropriate amount of data
- Representational DQ indicates that the system must present data in such a way that it is easy to understand (represented concisely and consistently) so that the consumer is able to interpret

the data; understood as the extent to which data is presented in an intelligible and clear manner. Dimensions include: interpretability, ease of understanding, representational consistency, and concise representation.

- Accessibility DQ emphasizes the importance of the role of systems; understood as the extent to which data is available to or obtainable by the data consumer. The system must also be secure. Dimensions include: accessibility and access security.

The Wang-Strong hierarchy is a very useful classification. The categories of data quality, in particular, highlight facets of data collection, storage, and use that have a direct impact on data consumers' perceptions of quality. In order for a data consumer to have a perception of intrinsic qualities of data, he or she must understand what the data is intended to represent and how the data effects that representation. If the data is not aligned with a consumer's assumptions about these things, the consumer will perceive it as inaccurate or unbelievable. Representational DQ emphasizes the role played by the specific presentation of data within a system.[1] Data can be perceived as incorrect if the consumer does not understand the conventions through which a system presents it. Contextual DQ points to the idea that the quality of data is defined to a large extent by the intended uses of data. Accessibility DQ points to another aspect of systems design. If a consumer cannot get to the data, then he or she cannot judge its quality by any other criteria.

Thomas Redman's Dimensions of Data Quality, 1996

In *Data Quality for the Information Age*, Tom Redman approaches data quality dimensions from the perspective of data modeling. Within this approach, a data item is defined in an abstract way, as a representable triple: a value, from the domain of an attribute, within an entity. This abstraction is a useful reminder of data's constructedness. Dimensions of quality can thus be associated with the component pieces of data (i.e., with the data model as well as with data values). A further dimension is data representation, which is defined as a set of rules for recording data items (p. 230). Redman identifies 27 distinct dimensions within these three general categories (data model, data values, data representation).

As noted, this approach to quality dimensions is rooted in an understanding of data structure. The first set of dimensions pertains to the conceptual view or data model. While not all data consumers are familiar with data models, models are critical to data use, since they provide a degree of context and meaning to any individual data item. Redman presents 15 characteristics of an ideal view or model of data and boils these down to six dimensions: content, level of detail, composition, consistency, and reaction to change (246–247). These characteristics are interrelated and reflect the choices that sometimes need to be made when developing a model—what to include, what to leave out, and the reasons for doing so.

The content dimensions include the relevance of data, the ability to obtain the values, and the clarity of definition. Relevance of data can be directly connected to its intended and potential uses.

[1] I have used the word "representational" in a sense different from both Strong and Wang's category of "Representational DQ" and Redman's (1999) "Data Representation" category. I have used the word "presentational" to refer to the set of characteristics that Wang and Strong categorize as "representational." "Presentational" refers to how the data itself is presented. I will use the word "representational" to refer to how data functions semioticly to represent aspects of the "real" world.

The ability to obtain data can be seen as a question of completeness of data to populate required attributes or as the appropriateness of data available for particular uses. Redman points out, for example, obstacles to obtaining data, including costs, privacy, and legal considerations, and he recognizes that many organizations resort to the use of surrogates to meet data requirements. His third consideration for content is what I have been referring to as metadata: clarity of definitions for all components of the model (i.e., entities, attributes, and their domains), as well as sources and rules for data.

The dimensions of scope include two potentially conflicting needs: for the model to be comprehensive and for it to contain only essential entities and attributes. Level of detail includes attribute granularity (the number and coverage of attributes used to represent a single concept) and precision of attribute domains (the level of detail in the measurement or classification scheme that defines the domain). Composition includes characteristics of naturalness (the idea that each attribute should have a simple counterpart in the real world and that each attribute should bear on a single fact about the entity), identify-ability (each entity should be distinguishable from every other entity), homogeneity, and minimum necessary redundancy (model normalization). Model consistency refers to both the semantic consistency of the components of the model and the structure consistency of attributes across entity types. The final dimensions of the model include robustness (its ability to accommodate changes without having to change basic structures) and flexibility (the capacity to change to accommodate new demands).

Dimensions related to data values comprise a set of four: accuracy, completeness, currency, and consistency. Redman's definition of accuracy is formulated in mathematical terms: The accuracy of a datum <*e, a, v*> refers to the nearness of the value *v* to some value *v*' in the attribute domain, which is considered as the correct one for entity *e* and attribute *a*. If the value *v* = value *v*', the datum is said to be correct (255). This equation explains the concept of accuracy, but, as Redman points out, the real challenge lies in knowing the correct value.

The next dimension of data values, completeness, refers to the degree to which values are present in a dataset. Values can be absent for different reasons, some expected, some not. The challenge with completeness comes in knowing whether or not the values are expected to be present.

Currency, the third dimension of data values, refers to time-related changes in data. Redman describes data being *up-to-date* and *current* as special cases of correctness and accuracy. Data can be correct when it is loaded to a database, but incorrect (out-of-date) if there is a change in the status of the entity being represented between the time the data is loaded and when it is used. For some data, the concept of currency (the degree to which data is up to date) is critical. For other data (of low volatility or representing "permanent" characteristics), the concept of currency is not critical.

Consistency is Redman's final dimension related to data values. He notes that the first characteristic of consistency is that two things being compared do not conflict. Consistency problems appear when datasets overlap and represent the same or similar concepts in a different manner, or when their specific content does not correspond.

Redman's classification includes eight dimensions related to data representation. Seven pertain to data formats—appropriateness and interpretability (both related to the ability of a data consumer to understand and use data), portability, format precision, format flexibility, ability to represent null values, and efficient use of storage. The eighth, representational consistency, pertains to physical instances of data being in accord with their formats. Redman concludes his description of dimensions with a recognition that consistency of entities, values, and representation can be understood in terms of constraints. Different types of consistency are subject to different kinds of constraints.

Redman's emphasis on the conceptual view or model of data differentiates his classification in a significant way from Wang and Strong's. The data model points to the constructedness of data. While model entities are said to represent real-world entities, they do so through an obvious set of choices—the selection of a set of specific attributes from a potentially infinite possible set of attributes and the definition of these attributes at a specific level of granularity and precision. As Redman rightly points out, the model provides context for understanding the data that is incorporated within it. What can be perceived as qualities intrinsic to the data must also be understood as constructs of the model.

In other respects, the two sets of dimensions share a number of the same insights and point to common concerns at the heart of data quality assessment. Both acknowledge that perceptions of accuracy or correctness and completeness are critical to data consumers—and that these aspects of data are related to data production, the way particular data represent the "real world." Both also recognize that aspects of presentation that make data easier to understand and interpret—representational consistency, appropriate format—improve the perception of the data's quality, largely by embodying conventions that make it easier for data consumers to understand what the data is intended to represent.

Redman's evolved formulation, as it appears in *Data Quality: The Field Guide* (2001), expands this classification. It identifies 55 dimensions across seven categories of quality. The new categories include aspects of what is often referred to as the data quality environment (data accessibility, privacy, commitment, and improvement), in addition to those associated with the data model, data values and data representation. This second formulation still includes space for dimensions of quality associated with aspects of the data model, including a new category for architecture. Unfortunately, it does not retain "quality of the conceptual view" as a distinct category. The dimensions associated with the model are embedded in the categories of quality of content and architecture (p. 106). I say "unfortunately" because, as I asserted above, the model is the most explicit signifier of the choices we make about what data is intended to represent, and without it being called out as a distinct category, there is a risk of losing sight of this critical knowledge when we use or assess data.

Larry English's Information Quality Characteristics and Measures, 1999

In *Improving Data Warehouse and Business Information Quality*, Larry English situates his discussion of information quality characteristics within a wider assessment of the information environment. Before assessing data quality, it is necessary to assess data definition and standards quality, information specification quality, and information architecture quality. These sets of metadata (data and data domain definitions, the data model, business rules) are critical because they provide the context needed for any uses of data.

English points out that to have high-quality information, it is necessary to have a specification for it. A specification includes a clear and precise definition of what the data represents, along with domain values and business rules for the representation. Data elements should be understood within a clear and well-defined information architecture. Data models should be designed consistently and coherently. They should not only meet requirements, but should also enable data consumers to better understand the data. Entity and attribute names should be clear, class words should be used consistently, abbreviations should be kept to a minimum, relationships should be correct, and the model should be available and understandable. Business rules should formalize relationships and be

expressed in a consistent manner (1999, pp. 83–136). If this metadata is not available, it is difficult if not impossible to assess the quality of specific data.

English adopts the word *characteristics* rather than *dimensions* to describe aspects of information quality that are important to data consumers. A *characteristic* is a feature or quality belonging to a person, place, or thing and serving to identify it. (In some ways, it is unfortunate that data quality thinkers overall did not choose *characteristic* as a starting point for describing features of data, saving *dimensions* to signify those characteristics most suitable for measurement and valuable to measure.) English identifies two sets of characteristics: inherent and pragmatic. Inherent characteristics are independent of data use. They are static characteristics of the data itself. Pragmatic characteristics are associated with data presentation and value to the accomplishment of effective work. Pragmatic characteristics are dynamic and their value (quality) can change depending on the uses of data.

The nine inherent quality characteristics are: definitional conformance, completeness of values, validity or business rule conformance, accuracy to a surrogate source, accuracy to reality, precision, nonduplication, equivalence of redundant or distributed data, and concurrency of redundant or distributed data. Definitional conformance is about the relation between data and metadata—that is, the degree of consistency between how data elements are defined and what the actual data values represent (p. 144). There are reasons why data definitions and actual data can be disparate. The two most significant reasons relate to the quality of the definition itself (definitions can be poorly written—unclear, ambiguous, incomplete, imprecise, circular) and to the volatility of the data (even if the original definition fairly represents the meaning of the data, new values can arise, the uses of data fields can "drift", etc.). English's recognition of the importance of data definition is akin to Redman's recognition of the importance of the model in providing data context.

English's other inherent qualities of data fields are straightforward. Completeness refers to the population of data fields. Completeness of any given field can have different effects on the quality of data (p. 144). Some fields are always expected to be populated; others are populated only under specific conditions. Precision refers to the granularity of the data (p. 147). Validity is defined as the conformance of data values to a domain or to a business rule. English distinguishes between validity and accuracy (p. 145). He defines accuracy as "the degree to which data accurately reflects the real-world object or event being described." Measuring accuracy to reality requires comparison to the actual object the data represents. For most situations, such measurement is prohibitively expensive (it amounts to data collection). However, a degree of accuracy can be determined through comparison to data contained in an original source (pp. 146–147).

English's final set of inherent quality characteristics refers to datasets. Duplication refers to the degree to which there are duplicate representations of the same single real-world object or event. Nonduplication is the degree to which there is a one-to-one correspondence between records and what they represent (p. 148). Duplication is a problem because most people using data assume that the same things (e.g., customers) are represented in the same manner (same name, same customer number, etc.) in a dataset. Many analyses depend on this assumption.

Duplication and data redundancy add to organizational costs in a variety of ways, from storage costs to customer dissatisfaction. The characteristic of equivalence of redundant or distributed data is a related challenge. This dimension of quality can be measured by assessing the semantic equivalence of data presentation in different representations of the same facts or concepts. Having different names for the same customer can be understood as both a duplication problem and an equivalence problem. Concurrency of redundant or distributed data presents a challenge similar to equivalent

representation, but with respect to data's up-to-date-ness, rather than its semantic consistency. There is always a lag time, or information float, between when data is created and when it is available for use. The problem of nonconcurrent data is intensified if, for example, two databases with essentially the same content will differ from each other if one is updated weekly and the other is updated monthly.

English's pragmatic characteristics, associated with data presentation and specific uses of data, include accessibility (including potential accessibility or obtainability of data and ease of actual access) and timeliness (the length of time between when data is known and when it is available for use) (p. 151). Contextual clarity is the degree to which data presentation enables a person to understand the meaning of data. Contextual clarity can include the presence of labels and other descriptive information. (I would argue that the first step to contextual clarity is English's first inherent quality characteristic, namely, the conformance of data to data definitions.) Usability is related to contextual clarity. The form in which information is presented must be suitable to its use. Derivation integrity is the correctness with which data is combined to form new information. Derivation integrity relies on both the precision of the calculation or logic that combines the data and the consistency with which the function is carried out. English's final characteristic is "rightness or fact completeness," which he defines as "having the right kind of data with the right qualities to support a given process." "Rightness" might also be defined as the degree to which data fulfills requirements for a given process.

English's 2009 work, *Information Quality Applied*, revisits and clarifies the earlier set of information quality characteristics. In his reformulation, he distinguishes between content characteristics and presentation characteristics (rather than inherent and pragmatic characteristics) and includes 12 additional dimensions, for a set of 27 all told.

There is an important difference between English's formulation and those of Wang and Strong and Redman: English's emphasis on a preliminary assessment of the information quality environment (data definition and standards, specifications, and information architecture) points to the idea that metadata and other explicit knowledge of data is critical to using it successfully.

More importantly, all three sets of dimensions recognize aspects of the reality of data that have been very influential in helping practitioners understand how to assess data quality. While demonstrating the interconnectedness of different dimensions of quality, they also point to three critical realities of data: Its representational function, its specific form (its presentation), and how these both affect its suitability to particular purposes and uses.