

Giri Kumar Tayi, Associate Professor, School of Business, Management Science and Information Systems, University at Albany

Giri Kumar Tayi gave a presentation on data quality management tailored to users and consumers of data. The first part of the presentation focused on key ideas and broad concepts around data quality. The second part dealt with data quality in the context of the public sector, and the third part proposed one approach to deliver data quality.

Giri presented an analysis of the similarities and differences of information age and industrial age resources. Information age resources do not consist of labor, capital, raw material, and energy, but of data, information, and knowledge. Traditional and information age resources differ in their very nature. It is important to recognize these differences in order to manage them effectively. Giri described the special characteristics of data. Data is intangible- whereas you can physically touch raw material. Data is not consumable- you can use it over and over again as opposed to any raw material which has a one time use. Data is shareable— you can easily share data across agencies. Data is easy to copy at a low cost— the main cost being to produce it not to copy it. Data can be more easily and automatically transported than raw material. Data is not fungible— you cannot substitute one part for another. Data is very fragile— it can easily be erased as it is magnetic. Data is versatile— it can be used by different parties for different uses. Data does not have market valuation— it is very difficult to come up with an economic value for data. Data is not depreciable— it does not get used up even if you use it a million times. He said that data quality addresses “fitness for use” of the raw material of the information age. You have to ask yourself how fit is the data for use, and for what purpose?

Giri mentioned that you have to be very specific about data quality and think about it in terms of attributes. He presented four broad categories of attributes that are important to manage in order to have good quality of data.

Table 1: Categories and Dimensions of Data Quality

CATEGORIES	ATTRIBUTES
Intrinsic	AccuracyObjectivityBelievabilityReputation
Contextual	CompletenessTimelinessRelevancyValue Added
Representational	InterpretabilityEase of UnderstandingConcise & Consistent representation
Accessibility	AccessibilityAccess security

Source: Beyond Accuracy: What Data Quality means to Data Consumers. Wang, et al (1994).

The problem with these attributes is that some are visible— such as accuracy, timeliness, usability— and others are not. Users tend to be realistic about visible factors while being unduly optimistic about the invisible ones. There is major trade-off that takes place between these attributes. You often need to give something up and it is important to make the right trade-off. For example, all things being equal, the more timely the data the better it is. However, very seldom are all things equal. The appropriate trade-off depends on the value of the data at different points of the processing spectrum. You start with a rough piece of data and then value is added over time so the value dimensions of the data keep changing. Data quality management means different things depending on your perspective. The definition of data quality has two perspectives. From the analyst's perspective, data quality management requires a sound understanding of the nature of data, identifying the factors that determine its quality, and articulating the underlying trade-offs. From an organizational perspective, data quality management involves specification of policies, identification of techniques, and use of procedures to ensure that the organizational data resource possesses a level of quality commensurate with the various uses of data. In transforming data into information, individuals play one or more roles. The three main roles are: (1) Data producers-people or groups who generate data; (2) Data custodians- people who provide and manage computing resources for storing and processing data (IT people); and (3) Data consumers- people or groups who use data. Each group has specific processes and tasks that you need to think about.

Giri provided a set of questions that need to be answered in order to assess the quality of data in an organizational context:

1. What is the data element of interest? Is it specific numbers, findings, conclusions?
2. How was the data or information created? When? Understanding the sources of bias and knowing the methods involved in creation of the data are important;

3. Who is associated with the data creation? Different audiences perceive the credibility of the data differently;
4. From what viewpoint was the data created and why? Every data element has a perspective that results from the context of its generation;
5. What relationship does this data element have to other data elements? Discerning the relationship is an important part of assessing data quality;
6. What approval, review and filtering process has the data endured? This is very important in the public sector as it gives believability to the data.

He recommended that approaches to enhance data quality focus on improving the data itself, improving the mechanisms that collect and deliver data, and improving the ability of an individual to assess the data quality for a specific purpose. These goals are not mutually exclusive and should all be pursued.

Giri presented an intrinsic data quality problem pattern developed by Strong, Lee and Wang (1997). It starts with having multiple sources of the same data. As a consequence, mismatches occur. Because of these mismatches, believability becomes questionable. The application becomes sloppy or the organization uses it only partially. The poor intrinsic data quality becomes common knowledge. Therefore, the data may end up not being used because of little added value and poor reputation. Mismatches among sources of the same data are a common cause of intrinsic data quality concerns.

Giri addressed the issue of how to deliver high quality data or information. The problem is that in most organizations, data or information is managed as the by-product of a system or an event. However, the consumer or user of the data views it as a product, not a by-product. Organizations often focus exclusively on the hardware or software components of the system rather than the data. Moreover, these components are managed in isolation and as a result the means of producing information becomes an end in itself. Giri presented a four-step approach to delivering high quality data developed by Wang et al. (1998). The first step is to understand the consumer needs. The goal is to ensure that the data is fit for consumer use. It is a total product that exhibits all the attributes that meet or exceed the consumer's expectations. You need to look at all the dimensions: timeliness, accuracy, etc. The second step is to manage the process. The process must be well defined and must contain adequate controls, inspection and production, and delivery time management. The third step is to manage the life cycle of the data. Just as in the case of a physical product, data products should be managed over their entire life cycle, keeping in mind the nature of the data, the tasks it supports, and the changing environment in which it is used. The last step is to delegate the responsibility of data quality to a single individual. He or she could coordinate and manage the three key stakeholder groups: suppliers of raw data, producers of the deliverable data product, and consumers of the data product. Giri characterized this as an integrated, cross-functional approach.

Finally, Giri provided a set of some general data quality rules developed by Orr (1996):

- Data that is not used cannot be correct for very long.
- Data quality in an information system is a function of its use, not its collection.
- Data quality will, ultimately, be no better than its most stringent use.
- Data quality problems tend to become worse with the age of the system.
- The less likely some data attribute (element) is to change, the more traumatic it will be when it finally does change.
- Laws of data quality apply equally to data and meta data.
- Variations among the data sources' attitudes, policies, and practices contribute to uneven data quality