Measuring and Improving Data Quality
Part 3: Improving Data Quality

Written by Mike Martin DVM, MPH, Clemson University

The second part of this series addressed the fact that data quality sometimes means different things to different data users. To improve quality for data customers, data managers need to be able to measure each “quality attribute.” Because measuring quality usually involves comparing recorded data to some independently collected “gold standard,” it is impractical to measure the quality of every piece of data. A better approach is to sample the data and respond to patterns in quality. This process is similar to monitoring national herd health status by collecting diagnostic tests from a sample of animals. The resulting measures are known as metrics. This paper presents proven approaches to improving data quality based on the results of these sample quality measurements.

Improving Data Quality

Several principles of data quality improvement are universal. For example, data quality must be designed into the data production process, not added after the fact. The quality improvement cycle from the manufacturing industry applies equally well to data production and data quality improvement. Data quality improvement depends on continuous feedback to the processes producing the data. Continuous feedback is best accomplished by putting each data element to as many uses as possible, ideally as a central part of the data collectors’ day-to-day work.

Data quality must be designed into systems using proven engineering principles. Data quality is too often left to chance or given only superficial attention in the design of information systems. While good engineering principles are sometimes applied to software development, data quality is usually left up to the end user. Applying engineering principles to data quality involves understanding the factors that affect the creation and maintenance of quality data. It is helpful to look at data as the output of a data manufacturing *process*.

Data start out as attributes of the real world. They are extracted through some measurement, lab test, or examination; recorded either on paper or in a computer system; or stored in human memory prior to recording. The process of recording data may require coding, applying medical terminology, or other error-prone transformations. The data are collected, aggregated, stored, and manipulated by various systems. Finally, the data are extracted and turned into information in some form of report or statistic. Quality—or the lack thereof—results from the overall performance of these processes.

A huge volume of work exists on methods for improving the quality of manufactured products. The classic approach used in manufacturing is Total Quality Management (TQM) or one of its variations. TQM was described by C. Edwards Deming in “Out of the Crisis."[1] Total Data Quality Management (TDQM) applies the "Plan, Do, Check, Act" cycle from Deming's TQM literature to information product (IP) creation.[2] The same concepts used to improve quality in manufacturing apply to information product creation. The best and least expensive way to ensure quality (in manufactured goods or data collection) is to apply continuous process improvement to the production process, rather than attempting to inspect and rework errors out of the product in post-production.[3] For data quality improvement, the cycle becomes "Define, Measure, Analyze, and Improve" instead of the "Plan, Do, Check, Act" cycle used in product manufacturing.

First, one must define the data quality attributes most important to IP customers, and define metrics or ways to measure these attributes. Second, these metrics are used to measure the current quality of each attribute. Analysis of these measures leads to selection of areas for improvement in data quality. Once targets are selected, tools such as root-cause analysis help identify potential systematic causes.[4] This analysis guides the implementation of process improvements designed to systematically improve the IP production process. Finally, the previously defined metrics are used to monitor the progress of improvement, validating that the process changes are, in fact, producing improved data quality and helping define additional integrity rules and improvement goals.[5]

It is beyond the scope of this paper to provide details about all the tools used in TQM, but one concept should be stressed. Root-cause analysis is a process by which errors are traced, not just to their immediate (proximate) cause, but to their ultimate most fundamental root cause. Too often the response to errors is to eliminate the proximate cause. This often results in temporary improvement, but because the underlying system defect still exists in the production process, sooner or later the problem will resurface. Applying systematic procedures for identification and correction of root causes provides more effective, permanent solutions.

The feedback control system (FCS) is a key engineering feature of any system that must interact with the real world. In a FCS, changes in the real world are fed back to the system, and control features of the system are adjusted accordingly. "Attempting to build quality systems without understanding FCS is like trying to design an airplane without understanding aerodynamics."[6] The data equivalent of an FCS is continuous use of the data by those who have direct knowledge of the real world truth.

Ken Orr has made the case that data quality can only be maintained in the long run by ensuring continuous and intensive use of both the data and the metadata (data about the data). "Use-based design means focusing on exactly how the data will be used and trying to identify inventive ways to ensure that the data are used more strenuously. In many cases, this means creatively persuading the people most knowledgeable about the data to take responsibility for it."[7] Taking responsibility for the data is not simply taking the blame when they are wrong, but making the data so central to one's real job that its quality becomes important for day-to-day work.

Readers are encouraged to explore the data quality improvement literature, a tiny part of which is referenced in this paper. While not a substitute for a more in-depth study of TDQM, the final paper in this series will make some suggestions for improving data quality within Veterinary Services. It will include both general approaches as well as some specific recommendations.

For more information on measuring and improving data quality, contact Dr. Michael Martin at mmarti5@CLEMSON.EDU.

[1] W. Edwards Deming, *Out of the Crisis*, MIT Press, Cambridge, 1982.

[2] R.Y. Wang, "A product perspective on total data quality management," *Comm of the ACM*, 1998; 41(2):58-65.

[3] Philip B. Crosby, *Quality Is Free: The Art of Making Quality Certain*, McGraw-Hill, New York, 1980.

[4] D. Strong, Y. Lee, R. Wang, "10 Potholes in the Road of Information Quality", *IEEE Computer*, pp. 38-46, August 1997.

[5] Y.W. Lee, et.al. "Process-embedded data integrity," *Journal of Database Management*, 15:1, 87-103, Jan-Mar 2004.

[6] K. Orr, "Data Quality and Systems Theory," *Communications of the ACM*, Vol. 41, No. 2, pp. 66-71, Feb. 1998.

[7] Orr 1198, p. 70.