

## **Rethinking Data Quality**

Data quality is a persistent problem. Quality has been an issue since the dawn of the IT profession and becomes increasingly challenging as the volumes of data increase and the uses for data expand. Information technology has advanced in so many areas. Why is it that we can't overcome the data quality problem? Maybe the barriers lie in the way that we think about data quality.

Common approaches to data quality problems include processes, projects, and software products. Data quality processes involve activities such as data quality assessment, root cause analysis, and data cleansing. Quality improvement projects employ processes within a structure of business justification, planning, prioritization, resource allocation, execution, and monitoring. Data quality products primarily perform the relatively basic tasks of matching, de-duplication, address standardization, etc.

Processes, projects, products – each of these contributes to the efforts to improve data quality. But they haven't solved the problems individually or collectively. To really make substantial and sustainable differences in the quality of data we need to take a different approach; We need to think of data quality as a *profession*. We need to have data quality professionals.

### **A Data Quality Profession**

You might argue that we already have IT professionals doing data quality work. True, but therein lies the problem. IT professionals doing data quality work is not the same as data quality professionals doing that work. That may appear to be a subtle difference, but it is really quite significant. The significance becomes apparent when you consider the nature of professions.

Wikipedia describes a profession as “a vocation founded upon specialized educational training” – a good place to start understanding the significance of a data quality profession. How many data quality training classes have you attended? How many have your peers attended? Many will answer one or two, but they may have to search old memory to recall the data quality class taken back in the 1990's. The state of data quality education for most IT professionals hardly represents extensive learning and depth of knowledge.

A profession is more than simply a category of employment. Professions are distinguished by several characteristics including fundamental principles, a defined body of knowledge, best practices, and mistakes to avoid. Let's examine each of the characteristics to consider the nature of a data quality profession.

### **Data Quality Fundamentals**

*Data Quality Definition:* Quality is defined as suitability to purpose. Data serves as “business memory” which has many purposes. Among those purposes are business

transactions, business reporting, audit trails, business measurement, business analysis, prediction and forecasting, decision making, and discovery and learning. The properties of high-quality data are likely to be different for each of these purposes.

*Quality Management Principles:* Many of the principles and disciplines of quality management in other fields can be readily applied to management of data quality. Deming, Baldrige, TQM, six sigma, statistical process control may all have roles in data quality improvement.

*Data Quality Economics:* There are costs associated with defects in data quality and with activities to improve data quality. Comparative costs – the cost to correct a problem compared to the cost of living with that problem needs to be considered. Cost comparison alone is inadequate. The value of high-quality data and information must also be considered.

*Data Quality Processes:* Data quality processes range from reactive processes to repair defects when they are found to proactive processes that focus on preventing defects from occurring. Processes have real economic impact. The cost of repair is invariably higher than the cost of prevention.

*Data Quality Dimensions:* A comprehensive view of data quality encompasses several dimensions including content, structure, presentation, usage, understandability, and business alignment.

### **Data Quality Body of Knowledge**

The data quality body of knowledge is something that is evolving in a field that is still emerging as a profession. Much work remains to be done to organize an agreed upon body of knowledge. It is apparent, however, from the literature and the efforts to date, that data quality is a broad and deep field. The body of knowledge encompasses topics such as:

- Quality disciplines – TQM, Six Sigma, etc.
- DQ processes – assessment, correction, repair, prevention, etc.
- Data profiling – purpose, results, and application
- DQ methods – procedural and rule-based data quality
- DQ technology – data cleansing, matching, de-duplication, standardization, etc.
- DQ teams – collaboration, communication, roles, responsibilities, accountability, etc.
- DQ projects – assessment, data cleansing, process improvement, etc.
- DQ in IT projects – data warehousing, application development, data migration, master data management, ERP implementation, business intelligence, etc.

### **Data Quality Best Practices**

It is impractical to attempt an exhaustive list of best practices in data quality. Many experienced practitioners will agree that the following items belong on the list:

- Fix problems, not symptoms. Root cause analysis is important.
- Measurement, monitoring, and feedback are essential parts of any data quality program.
- Data quality problems usually result from systemic and global causes. Local fixes don't work.
- People, processes, and technology are all sources of data quality defects.
- Data quality management is a multi-disciplinary field that demands both business and technical knowledge and skills.
- Designated roles, responsibilities, and accountabilities are fundamental to data quality management.
- Designate authority to make decisions and take actions that is commensurate with the level of responsibility and accountability.
- Distinguish between data ownership, stewardship, and custodianship and recognize the importance of each role.
- The business must be actively engaged, involved, and participative for data quality efforts to succeed.
- Sustained data quality is achieved with long-term programs, not short-term projects.

### **Mistakes to Avoid**

Similar to best practices, mistakes to avoid are many. It is impractical to attempt a complete listing here. Among the common mistakes are:

- Saying that data quality is “everybody’s responsibility” – that is the same as saying nobody is responsible.
- Treating data quality as an IT problem.
- Lack of a business case for data quality efforts.
- Seeking the easy fix or the “silver bullet.”
- Reactive data quality management using a repair-only approach.
- Lack of data quality standards – especially for master data and shared reference data.
- Absence of measurement or lack of targets that give the measures context.
- Lack of expertise, both business and technical.
- Believing that data knowledge or data modeling skill is an acceptable proxy for data quality skills.
- Insufficient measures – counting defects but failing to quantify cost or impact.

### **In Conclusion**

This short article cannot possibly describe all of the complexities and challenges of managing data quality. A quick look at the scope – fundamentals, body of knowledge, best practices, and common mistakes – makes it apparent that data quality is a challenging and multi-faceted field. And it needs to be recognized as a distinct field that overlaps with but is not contained in any of IT, data warehousing, business systems, database development, or information systems. It is time that data quality becomes a profession and skilled practitioners make the leap to becoming data quality professionals.