# DATA QUALITY: THEORY IN PRACTICE

**Dr. Richard Y. Wang**
**Director, MIT Information Quality Program**

We do not always notice data quality (DQ) problems until it is too late. However, organizations are increasingly aware of the overwhelming advantages of high-quality information. They are also painfully aware of the significant costs of low-quality information – costs that translate into hard dollars, reduced productivity, waste, and myriads of other consequences – even affecting quality-of-life.

## DQ Landscape

There is a growing demand for DQ initiatives as organizations' awareness of the importance of their DQ increases. Efforts in the Federal government include DoD DQ Guidelines, Data Quality Act, and the reference models issued by the Office of Management and Budget (OMB). In the private sector, the DQ landscape includes a variety of vendors, such as Acxiom, Ascential, Evoke, Firstlogic, and Group 1. These vendors utilize different technologies, such as ETL (Extraction, Transform, Load), name matching, cleansing and profiling, business rule extraction, and customer recognition systems to provide DQ services to organizations.

Leading organizations no longer treat data quality as an unknown. Rather, people today accept that they do have DQ problems but may not know how to deal with such problems. Furthermore, many new DQ issues have emerged as organizations began their journey to data quality: real-time cleansing and matching, metadata management, DQ dashboards, and new government regulations. Real-time cleansing and matching have become an important issue in the real-time nature of e-business practices and processes today. The issue of metadata management has gathered attention because standards, definitions, and application metadata sharing are key to solving many DQ problems.

In helping organizations tackle their DQ problems, we have learned many lessons. A decade ago DQ job titles were unheard of, whereas today more and more organizations have formal positions and job descriptions for DQ analysts, managers, directors, and vice presidents. Meanwhile, traditional DQ efforts have gone beyond the simple issue of measuring accuracy. It is well established that in improving data quality, organizations must treat DQ as a multi-dimensional concept beyond accuracy[1,2]. Methods and tools for performing DQ assessments have been developed and widely accepted in practice. In addition, understanding the systems, processes, and management practices of an organization has become as important as understanding its data. Furthermore, many organizations have found that resolving their DQ problem is not a single-phase process. Rather, it is a journey, where one solution may lead to new problems and employees at all levels must come together and to solve the DQ problems.

[1] Richard Wang & Diane Strong, *Beyond Accuracy: What Data Quality Means to Data Consumers.* Journal of Management Information Systems, 1996. 12(4): p. 5-34.
[2] Yang Lee, Leo Pipino, James Funk & Richard Y., *Journey to Data Quality.* Forthcoming, MIT Press.



**An MIT Perspective**

Early research efforts in DQ at MIT led to the development of the Total Data Quality Management (TDQM) cycle: *Define, Measure, Analyze*, and *Improve*. The *definition* component identifies DQ dimensions. The *measurement* component produces DQ metrics. The *analysis* component identifies root causes for DQ problems and calculates the impacts of poor-quality data. Finally, the *improvement* component provides techniques for improving DQ. A core set of data quality requirements has also

been identified to perform the TDQM Cycle for continuous improvement and has been published in leading journals and books. The past research efforts of TDQM at MIT have helped lay the foundation for some of the more current activities, including those related to managing information as a product, customer-centric DQ management, corporate household research, and data quality training programs.

It is important to recognize that DQ is multi-dimensional and includes many factors beyond accuracy, including objectivity, completeness, timeliness, ease of understanding, and believability. In the measurement component data integrity analysis becomes the fundamental activity by applying Codd's five integrity constraints, namely domain, entity, referential, column, and user-defined integrity, to the data at hand. One of the key activities of the analysis component is the use of information product maps (IPMAPs) to analyze the data flow in an organization. IPMAPs, a powerful modeling technique, address DQ problems and bring awareness of these problems to organizations by identifying their root causes. Once the root causes of DQ problems have been identified, the improvement component begins with the DQ Filtering System (DQFS). The DQFS is part of the undergoing research efforts at MIT to develop an innovative, practical, and integrated approach that facilitates the development, production, maintenance, and management of high-quality information products throughout their life cycle. The vision behind the DQFS is to extend the existing body of research work and to develop a scalable system for large-scale operations and small organizations.

**Implementation Issues**

Without a solid foundation of high-quality data, "dirty data" can chip away at an organization's ability to function effectively. A data quality initiative that is well defined within the context of an organization may still encounter difficulties with implementation. Some of the difficulties faced by organizations when executing their data quality initiatives include dealing with data standards, handling secondary information, reconciling science DQ and general data quality, integrating disparate disciplines, and scrapping and re-working.

**Conclusion**

Today's EPA conference provides us with the unique opportunity to discuss the issues facing quality systems for environmental programs. Data quality is a core issue in improving environmental data that is fundamental to successful environmental programs. Over the last

two decades, DQ research efforts, government leadership, and industry practice together have yielded significant results for DQ implementation and institutionalization. Today data quality is practiced and researched at a rapid pace. The lessons learned in applying solutions to solve DQ problems in other settings could be adopted to manage the quality of environmental data, which in turn would enhance the EPA Quality Community's ability to contribute to environmental protection efforts.

# Keynote Speaker



## Dr. Richard Y. Wang
## Director, MIT Information Quality Program

Dr. Richard Y. Wang is Director of *MIT Information Quality* (MITIQ) Program and Co-Director for the *Total Data Quality Management Program* at the Massachusetts Institute of Technology. He had served as a professor at MIT for a decade prior to heading the MITIQ program. He has also served as a professor at the University of Arizona, Boston University and as a visiting Professor at the University of California, Berkeley. Dr. Wang received his Ph.D. in Information Technology from MIT.

Dr. Wang has put the term *Information Quality* on the intellectual map with myriad publications. In 1996, he organized the premier *International Conference on Information Quality*, which he has served as the general conference chair, and is currently serving as Chairman of the Board. Dr. Wang's books on information quality include *Quality Information and Knowledge* (Prentice Hall, 1999), *Data Quality* (Kluwer Academic, 2001), and *Journey to Data Quality* (MIT Press, forthcoming).

Dr. Wang's current research and industry practice focus on extending information quality to enterprise issues such as data architecture, data governance, and data sharing. He is involved in the *Leaders in Enterprise Architecture Deployment* (LEAD) project, which is sponsored by the U.S. government. Additionally, he heads a *customer centric information quality management* (CCIQM) working group. At MITIQ program, Dr. Wang offers certificate programs and executive courses on information quality management, as well as a planned course on systems integration in 2005. He can be reached at rwang@mit.edu, http://mitiq.mit.edu