Richard Y. Wang

A Product Perspective on Total Data Quality Ma

To increase productivity, organizations must manage information as they manage products.

The field of information quality (IQ) has experienced significant advances during its relatively brief history. Today, researchers and practitioners alike have moved beyond establishing information quality as an important field to resolving IQ problems—problems ranging from IQ definition, measurement, analysis, and improvement to tools, methods, and processes. However, theoretically-grounded methodologies for Total Data Quality Management (TDQM) are still lacking. Based on cumulated research efforts, this article presents such a methodology for addressing these problems. The purpose of this TDQM methodology is to deliver high-quality *information products* (IP) to information consumers. It aims to facilitate the implementation of an organization's overall data quality policy formally expressed by top management [10].

nagement

The terms data and information are often used synonymously; in practice, managers differentiate information from data intuitively, and describe information as data that has been processed in some manner. Unless specified otherwise, this article will use "information" interchangeably with "data."

The results of the research appearing in this article contribute to the IQ field by developing concepts and principles for defining, measuring, analyzing, and improving IP. We developed a survey-based diagnostic instrument for IQ assessment, from which a software tool has been developed to collect data and plot IQ dimensional scores for the individual, organizational role, and overall averages once data has been collected [4]. We've also developed a pragmatic methodology based on current research, and will illustrate how this methodology can be applied in practice.

Table 1.
Products vs. information manufacturing

	Product Manufacturing	Information Manufacturing
Input	Raw Materials	Raw Data
Process	Assembly Line	Information System
Output	Physical Products	Information Products

An analogy exists between quality issues in product manufacturing and those in information manufacturing, as shown in Table 1. Product manufacturing can be viewed as a processing system that acts on raw materials to produce physical products. Analogously, information manufacturing can be viewed as a processing system acting on raw data to produce information products. The field of product manufacturing has an extensive body of literature on Total Quality Management (TQM) with principles, guidelines, and techniques for product quality. Based on TQM, knowledge has been created for IQ practice [6, 8]. An organization would follow certain guidelines to scope an IQ project, identify critical issues, and develop procedures and metrics for continuous analysis and improvement. Although pragmatic, these approaches have limitations.

The limitations arise from the nature of raw materials used in information manufacturing, namely data. Data can be utilized by multiple consumers and not depleted, whereas a raw material can only be used for a single physical product. Another dissimilarity arises from timeliness. One could say that a raw material arrived just in time, but one would not ascribe an intrinsic property of timeliness to the raw material. Other dimensions such as the believability of data simply do not have a counterpart in product manufacturing. In short, many research issues need to be addressed in order to develop a methodology for TDQM. Much research has been conducted toward this

goal, and the proposed TDQM methodology builds upon and utilizes these efforts.

Research Foundations

Methodologies developed for any field must be discipline-based and rigorous so that they can be repeatedly tested and employed by others. These methodologies should also introduce applicable concepts that capture pertinent ideas in different operational environments. The proposed TDQM methodology is based on accumulated research and extended practical experiences. To present the methodology, we first introduce the concepts of TDQM cycle and information products.

The TDQM Cycle. Defining, measuring, analyzing, and improving information quality continuously is essential to ensure high-quality IP. In the TQM literature, the widely-practiced Deming cycle for quality enhancement consists of: Plan, Do, Check, and Act. By adapting the Deming cycle [7], we develop the TDQM cycle. The definition component of the TDQM cycle identifies important IQ dimensions [11] and the corresponding IQ requirements. The measurement component produces IQ metrics. The analysis component identifies root causes for IQ problems and calculates the impacts of poor quality information. Finally, the improvement component provides techniques for improving IQ. They are applied along IQ dimensions according to requirements specified by the consumer.

Table 2.
Information quality categories and dimensions

IQ Category	IQ Dimensions
Intrinsic IQ	Accuracy, Objectivity, Believability, Reputation
Accessibility IQ	Access, Security
Contextual IQ	Relevancy, Value-Added, Timeliness, Completeness, Amount of data
Representational IQ	Interpretability, Ease of understanding, Concise representation, Consistent representation

The Information Product. We refer to an information manufacturing system as a system that produces information products. The concept of IP is introduced to emphasize the fact that the information output from an information manufacturing system has value that is transferable to the consumer. We identify four roles:

- *Information suppliers* are those who create or collect data for the IP.
- *Information manufacturers* are those who design, develop, or maintain the data and systems infrastructure for the IP.
- *Information consumers* are those who use the IP in their work.
- *IP managers* are those who are responsible for managing the entire IP production process throughout the IP life cycle.

We illustrate these four roles with a financial company's client account database. A broker who creates accounts and executes transactions has to collect from clients the necessary information for opening accounts and executing these transactions. The broker, therefore, is a supplier. An information-systems professional who designs, develops, produces, or maintains the system is a manufacturer. A financial controller or a client representative who uses this system is a consumer. Finally, a manager who is responsible for the collection, manufacturing, and delivery of customer account data is an IP manager.

Information Quality. Just as a material product has quality dimensions associated with it, an IP has IQ dimensions. IQ has been viewed as fitness for use by information consumers, with four IQ categories and fifteen dimensions identified [11]. As shown in Table 2, the intrinsic IQ captures the fact that information has quality in its own right. Accuracy is merely one of the four dimensions underlying this category. Contextual IQ highlights the requirement that information quality must be considered within the context of the task at hand. Representational and accessibility IQ emphasize the importance of the role of information systems.

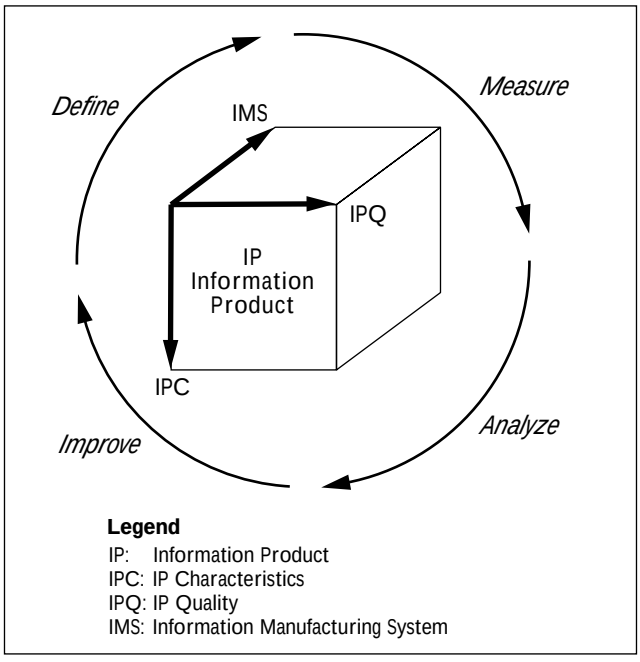

Figure 1.
A schematic of the TDQM methodology

Overview of the TDQM Methodology

In applying the TDQM methodology, an organization must: (1) clearly articulate the IP in business terms; (2) establish an IP team consisting of a senior executive as the TDQM champion, an IP engineer who is familiar with the TDQM methodology, and members who are information suppliers, manufacturers, consumers, and IP managers; (3) teach IQ assessment and IQ management skills to all the IP constituencies; and (4) institutionalize continuous IP improvement.

A schematic of the TDQM methodology is shown in Figure 1. The tasks embedded in this methodology are performed in an iterative manner. For example, an IP developed in the past may not fit today's business needs for private client services in an investment bank. This should have been identified in the defining phase if private client representatives are involved. If not, it would be the IP team's responsibility to ensure this need is met at a later phase; otherwise this IP will not be fit for use by the private client representatives.

In applying this TDQM methodology, one must first define the characteristics for the IP, assess the IP's information quality requirements, and identify the information manufacturing system for the IP [3]. These tasks can be challenging for organizations that are not familiar with this methodology. Our experience shows, however, that after these tasks have been performed once and the underlying concepts and mechanisms are understood, it becomes relatively easy to repeat the work for another IP. Once these tasks are accomplished, other work—measurement, analysis, and improvement—ensues.

The TDQM Methodology: Define IP

Define IP Characteristics. The characteristics of an IP are defined at two levels. At the higher level, the IP is conceptualized in terms of its functionalities for information consumers. As in defining what constitutes an automobile, it is useful to first focus on the basic functionalities and leave out advanced capabilities (for example, optional features for an automobile such as air conditioning, radio equipment, and cruise control).

Continuing with the client account database example, the functionalities are customer information needed by information consumers to perform the tasks at hand. The characteristics for the client account database include items such as account number and stock transactions. The functionalities and consumers of the system are identified in an iterative way. The consumers include brokers, client representatives, financial controllers, accountants, and corporate lawyers (for regulatory compliance). Their perceptions of what constitute important IQ dimensions need to be captured and reconciled.

At a lower level, one would identify the IP's basic units and components and their relationships. Defining what constitutes a basic unit for an IP is critical as it dictates the way the IP is produced, utilized and managed. In the client account database, a basic unit would be an ungrouped client account.

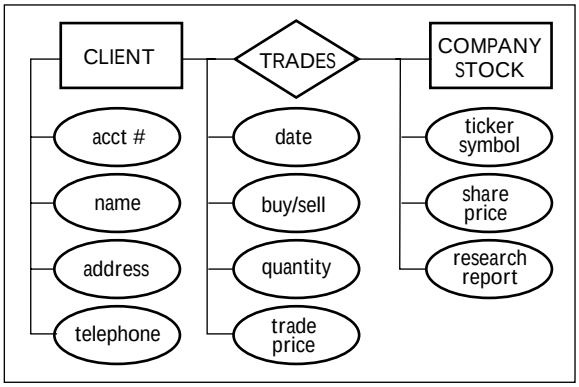

Figure 2.
A client account schema

In practice, often it is necessary to group basic units together (just as eggs are packaged and sold by the dozen). A manager of mutual funds would trade stocks on behalf of many clients, necessitating group accounts; top management would want to know how much business the firm has with a client that has subsidiaries in Europe, the Far East, and Australia. Thus, a careful management of the relationship between basic accounts and aggregated accounts, and the corresponding processes that perform the mappings are critical because of their business impacts.

Components of the database and their relationships can be represented as an entity-relationship model. In the client account database, a client is identified by an account number. Company stocks are identified by the companies' stock ticker symbols. When a client makes a trade (buy/sell), date, quantity of shares and trade price is stored as a record of the transaction. An ER diagram is shown in Figure 2.

Define IQ Requirements. With the characteristics of the IP specified, the next step is to identify IQ requirements from the perspectives of IP suppliers, manufacturers, consumers, and managers. We have developed an instrument for IQ assessment and corresponding software to facilitate the IQ assessment task.

After data has been collected from information suppliers, manufacturers, consumers, and IP managers, it is entered into the survey database for the IQ assessment software tool to perform the query necessary for mapping the item values in the surveys to the underlying IQ dimensions [4]. Figure 3 illustrates the capability of the software tool through data collected from a manufacturer and a consumer.

The result from the first dimension indicates that the

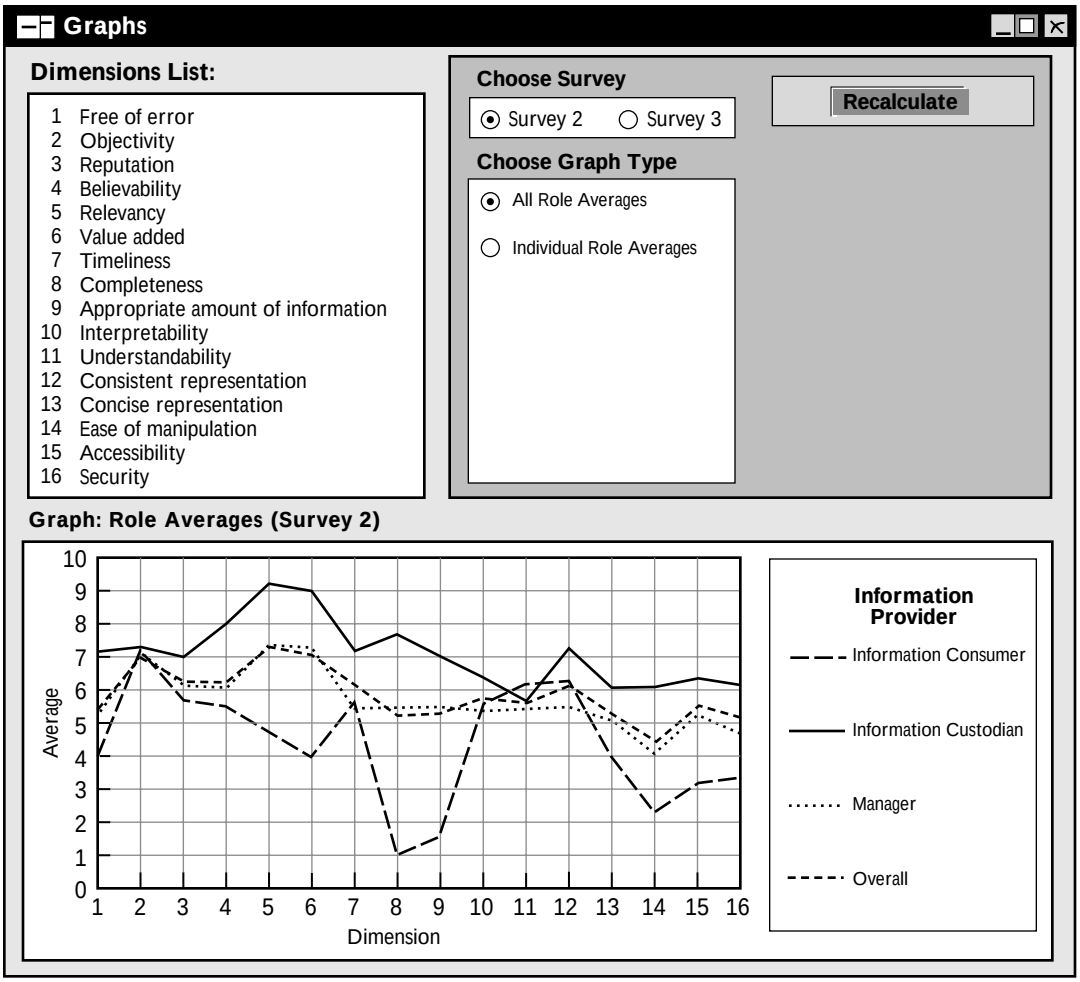

Figure 3. Dimensional assessment of IQ importance across roles

manufacturer (referred to in Figure 3 as Information Custodian) believes the IP to be largely free of error (score 7 on a scale from 0 to 10 with 10 being completely free of error), whereas the consumer does not think so (with a score of 4). Both the manufacturer and the consumer indicate the IP contains objective and relatively important data (score 7). The biggest contrast shows up for Dimension 8, completeness. Although the manufacturer assesses the IP as having reasonably complete data (score 7.6), the consumer thinks otherwise (score 1)!

From the IP characteristics and the IQ assessment results, the corresponding logical and physical design of the IP can be developed with the necessary quality attributes incorporated [9]. Timeliness and credibility are two important IQ dimensions (Dimensions 3 and 7) for an IP supporting trading operations. In Figure 4a), timeliness on share price indicates the trader is concerned with how old the data is. A special symbol, "√ inspection" is used to signify inspection requirements such as data verification.

The IQ requirements are further refined into more objective, measurable characteristics [9]. These characteristics are depicted as a dotted-rectangle as shown in Figure 4b. For example, timeliness is redefined by age (of the data), and credibility of the research report is redefined by analyst name. The quality indicator collection method, associated with the telephone attribute, is included to illustrate that multiple data collection mechanisms can be used for a given type of data; values for collection method may include "over the phone" or "from an existing account."

The quality indicator media for research report is to indicate the multiple formats of database-stored documents such as bitmapped, ASCII, or postscript. The quality indicators derived from "√ inspection" indicate the inspection mechanism desired to maintain data reliability. The specific inspection or control procedures may be identified as part of the application documentation. These procedures might include independent, double entry of important data, front-end rules to enforce domain or update constraints, or manual processes for performing certification on the data.

Define Information Manufacturing System. Equally important to the task of identifying IQ dimensions is the identification of the information manufacturing system for the IP. Figure 5 illustrates an information manufacturing system which has five data

units (DU_1–DU_5) supplied by three vendors (VB_1–VB_3). Three data units (DU_6, DU_8, DU_{10}) are formed by having been passed through one of the three data quality blocks (QB_1–QB_3). For example, DU_6 represents the impact of QB_1 on DU_2 [3].

There are six processing blocks (PB_1–PB_6) and accordingly six data units (DU_7, DU_9, DU_{11}, DU_{12}, DU_{13}, DU_{14}) that are the output of these processing blocks. One storage block (SB_1) is used both as a pass-through block (DU_6 enters SB_1 and is passed on to PB_3) and as the source for database processing (DU_1 and DU_8 are jointly processed by PB_4). The system has three consumers (CB_1–CB_3). Each consumer receives some subset of the IP. The placement of a quality block following a vendor block (similar to acceptance sampling) indicates that the data supplied by vendors is deficient in regard to IQ.

In the client account database, identifying such an information manufacturing system would provide the IP team with the basis for assessing the values of IQ dimensions for the IP through the Information Manufacturing Analysis Matrix and studying options in analyzing and improving the information manufacturing system [3].

Summary. The IP definition phase produces two key results: (1) a quality entity-relationship model that defines the IP and its IQ requirements, and (2) an information manufacturing system that describes how the IP is produced, and the interactions among information suppliers (vendors), manufacturers, consumers, and IP managers.

With these results from the IP definition phase, an organization has two alternatives. First, the organization can develop a new information manufacturing system for the IP based on these results. The advantage of this approach is that many IQ requirements can be designed into the new information manufacturing system, resulting in quality-information-by-design analogous to that of quality-by-design in product manufacturing. Many of the IQ problems associated with a legacy system can also be corrected with the new system. The disadvantage is that a new system would require more initial investment and significant organizational change.

Alternatively, the organization can use these results as guidelines for developing mechanisms to remedy the

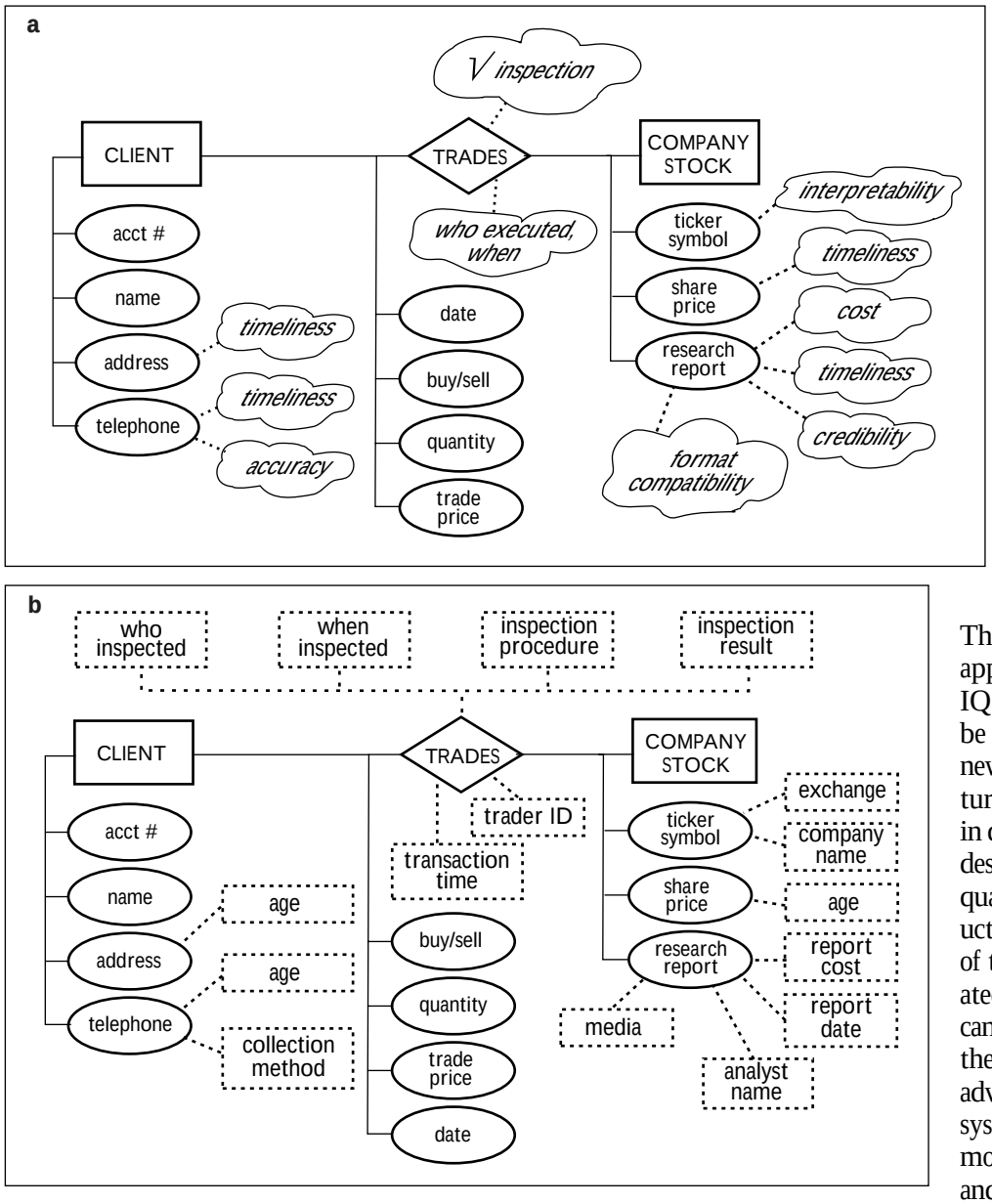

Figure 4.
a) IQ added to the ER diagram
b) A quality entity-relationship diagram

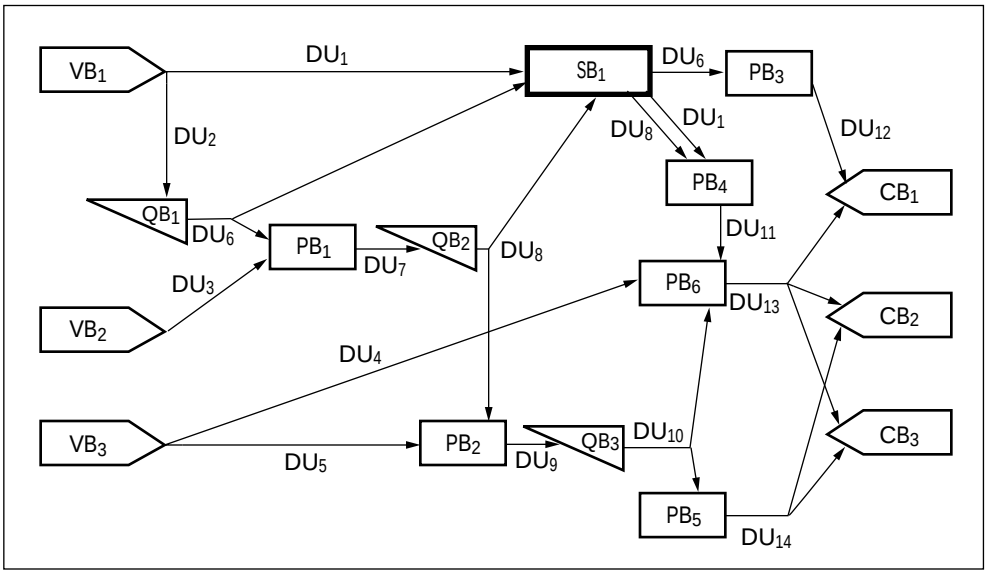

Figure 5.
An illustrative information manufacturing system

deficiencies of the existing system. Ultimately, however, a new information manufacturing system will need to be developed as the business environment changes which, in turn, will change the consumers' information needs.

The TDQM Methodology: Measure IP

The key to measurement is the development of IQ metrics. These IQ metrics can be the basic IQ measures such as data accuracy, timeliness, completeness, and consistency [1, 12]. In the client account database, IQ metrics may be designed to track, for example:

- The percentage of incorrect client address zip codes found in a randomly selected client account (free of error)
- An indicator of when client account data were last updated (timeliness or currency for database marketing and regulatory purposes)
- The percentage of non-existent accounts or the number of accounts with missing value in the industry-code field (completeness)
- The number of records that violate referential integrity (consistency)

At a more complex level, there are business rules that need to be observed. For example, the total risk exposure of a client should not exceed a certain limit. This exposure should be monitored for clients who have many accounts. Conversely, a client who has a very conservative position in one account should be allowed to execute riskier transactions in another account. For these business rules to work, however, the IP team needs to develop a proper account linking method and the associated ontology to make the linkage.

There are also information-manufacturing-oriented IQ metrics. For example, the IP team may want to track:

- Which department made most of the updates in the system last week
- How many unauthorized accesses have been attempted (security)
- Who collected the raw data for a client account (credibility)

Other IQ metrics may measure the distribution of the DQ-related collective knowledge across IP roles. Whatever the nature of the IQ metrics are, they are implemented as part of a new information manufacturing system or as add-on utility routines in an existing system. With the IQ metrics, IQ measures can be obtained along various IQ dimensions for analysis.

The TDQM Methodology: Analyze IP

From the measurement results, the IP team investigates the root cause for current IQ problems. The methods and tools for performing this task can be simple or complex. In the client account database, one can introduce dummy accounts into the information manufacturing system to identify sources that cause poor IQ. Other methods include statistical process control (SPC), pattern recognition, and Pareto chart analysis for poor IQ dimensions over time.

We illustrate other types of analysis through the case of the Medical Command of the Department of Defense that has developed IQ metrics for information in their Military Treatment Facilities (MTF). In that case [5], one must analyze the assumptions and rationale underlying the IQ metrics such as:

- What the targeted payoffs are
- How the IQ metrics link to the factors that are critical to the target payoffs
- How representative or comprehensive these IQ metrics are
- Whether these IQ metrics are the right set of metrics

The target payoffs could be twofold: (1) the delivery of ever-improving value to patients and other stakeholders, contributing to improved healthcare quality; and (2) improvement of overall organizational effectiveness, use of resources, and capabilities. It would be important to explicitly articulate the scope of these metrics in terms of the categories of payoffs and their linkages to the critical factors.

To provide the best health care for the lowest cost, different types of data are needed. MTF commanders need cost and performance data, managed care support contractors need to measure the quality and cost of their services, and patients need data they can use to know what kind of services they would receive from different health plans. The types of data needed can fall into several categories: patient, provider, type of care, use rate, outcome, and financial. Based on the targeted payoffs, the critical factors, and the corresponding types of data needed, one can evaluate how representative or comprehensive these IQ metrics are and whether these metrics are the right set of metrics.

The TDQM Methodology: Improve IP

IP improvement phase ensues once the analysis phase is complete. The IP team needs to identify key areas for improvement such as: (1) aligning information flow and work flow with the corresponding information manufacturing system, and (2) realigning the key characteristics of the IP with business needs. As mentioned earlier, the Information Manufacturing Analysis Matrix [3] is designed for these purposes. Ballou and Tayi [2] also develop a methodology for allocating resources for IQ improvement. Specifically, an integer programming model is developed to determine which databases should be chosen to maximize IQ improvement given budget constraints.

Conclusion

We have developed the concepts, principles, and procedures for defining, measuring, analyzing, and improving information products. We have also developed an IQ survey software instrument for information quality assessment. Based on these and cumulated research efforts, we have presented a Total Data Quality Management methodology, and illustrated how this methodology can be applied in practice.

The power of the TDQM methodology stems from the cumulative multidisciplinary research and practice in a wide range of organizations. Fundamental to this methodology is the premise that organizations must treat information as a product that moves through an information manufacturing system, much like a physical product, yet realize the distinctive nature that the IP exhibits.

Consumers are more likely to find problems with the information they use, particularly contextual IQ. The IP problems, however, should not be left for consumers to recognize and resolve. The IP team must proactively improve the quality of the IP continuously. To this end, information manufacturers as well as information suppliers need to expand their knowledge about how and why the consumers use information. Conversely, information consumers need to understand how information is produced and maintained so that the communication among the different roles can be effective. The TDQM methodology has been shown to be effective for improving IP, particularly when top management has a strong commitment, as expressed in the organization's IQ policy. Organizations of the 21st century must harness the full potential of their data in order to gain competitive advantage and attain strategic goals. The TDQM methodology has been developed as a step to meeting this challenge. C

References

1. Ballou, D.P. and Pazer, H.L. Modeling data and process quality in multi-input, multi-output information systems. *Management Science 31*, 2 (1985), 150–162.
2. Ballou, D.P. and Tayi, G.K. Methodology for allocating resources for data quality enhancement. *Commun. ACM 32*, 3 (Mar. 1989), 320–329.
3. Ballou, D.P., Wang, R.Y., Pazer, H., and Tayi, G.K. Modeling information manufacturing systems to determine information product quality. *Management Science* (1997).
4. Cambridge Research Group. *Information Quality Survey: Administrator's Guide.* Cambridge Research Group, Cambridge, MA, 1997.
5. Corey, D., Cobler, L., Haynes, K., and Walker, R. Data quality assurance activities in the military health services system. In *Proceedings of the 1996 Conference on Information Quality*. (Cambridge, Mass., 1996), pp. 127–153.
6. Cykana, P., Paul, A., and Stern, M. DoD Guidelines on data quality management. In *Proceedings of the 1996 Conference on Information Quality*. (Cambridge, Mass., 1996), pp. 154–171.
7. Deming, E.W. *Out of the Crisis.* Center for Advanced Engineering Study, MIT, Cambridge, MA, 1986.
8. Firth, C.P. and Wang, R.Y. *Data Quality Systems: Evaluation and Implementation.* Cambridge Market Intelligence Ltd., London, 1996.
9. Wang, R.Y., Kon, H.B., and Madnick, S.E. Data quality requirements analysis and modeling. In *Proceedings of the 9th International Conference on Data Engineering.* (Vienna, Austria, 1993), pp. 670–677.
10. Wang, R.Y., Storey, V.C., and Firth, C.P. A framework for analysis of data quality research. *IEEE Trans. Knowledge and Data Engineering 7*, 4 (1995), 623–640.
11. Wang, R.Y. and Strong, D.M. Beyond accuracy: What data quality means to data consumers. *J. Manage. Info. Syst. 12*, 4 (1996), 5–34.
12. Wang, R.Y. and Lee, Y.W. *Integrity Analyzer: A Software Tool for Total Data Quality Management.* Cambridge Research Group, Cambridge, MA, 1998.

Richard Y. Wang (rwang@mit.edu), a leading authority in data quality research and practice, is Co-Director of the Total Data Quality Management Program at Massachusetts Institute of Technology.

Research has been supported in part by MIT's Total Data Quality Management (TDQM) Research Program, Cambridge Research Group, and Naval Command, Control, and Ocean Surveillance Center's (NCCOSC) contract # NM66001-91-D-0103.