SECTION 34
SECOND-GENERATION DATA QUALITY SYSTEMS

Thomas C. Redman

INTRODUCTION

This section discusses the ubiquity and poor quality of data. More specifically, it explains why the manager must be concerned about poor data quality and discusses the approaches to improvement. Most people recognize that if there are errors in data, they should be corrected. And indeed, good software packages can help find certain errors in even the largest databases. But new data are created at enormous rates, so the job of error detection and correction never ends. Leading enterprises have achieved superior results with so-called second-generation data quality systems, which are those focused on preventing future errors. This section is a guide to the leader who wishes to upgrade to a second-generation data quality system and reap the benefits of improved customer satisfaction, lowered costs, and more confident decision making.

This section begins with an outline of the steps taken by one executive to do so. It then summarizes critical differences between data and other resources and the complexities they engender for data quality. Next, it sketches the business case for second-generation data quality systems. Lastly, it describes good practice in defining and implementing them.

THIS REPORT IS WRONG!

To illustrate one way that implementation of a second-generation data quality system could proceed, consider a fairly large service enterprise, organized in several business units and staff functions. The enterprise is profitable, but its industry is extremely competitive and becoming more so. In this example, the head of one business unit is frustrated that summaries of financial performance are late, incomplete, and inaccurate. Important decisions are delayed, then made in haste. Poorer decisions that are harder to implement are the direct result.

This executive decides to start a small program to explore the benefits of a second-generation data quality system. He or she selects an important data-driven business problem. (Experience suggests that areas and problems involving money, such as billing, revenue, customer accounts, and marketing, are especially impacted by poor data and can produce case studies with bottom-line results rather quickly.) A *customer need analysis* is conducted to determine who is concerned about this problem and to understand their data needs. Next, primary *sources of* (*internal*) *data are identified.* A *planning project* identifies two or three important gaps, and *improvement projects* follow. Upon completion, *controls* are implemented to hold the gains.

The people responsible for these projects share their successes with this executive's staff. The executive urges further projects and each staff member agrees to conduct a test project or two. In conducting these tests, several staff members recognize that the data cross their functions and they begin to explore *process management* as a means of working together. Another acquires important data from a commercial source, and decides to explore *supplier management.*

Not all projects succeed. But enough do and the reports that originally frustrated the executive are now much better. The staff decides to devote a portion of each bi-weekly staff meeting to data quality, effectively forming itself as a *Data Council.* And it, at the vocal urgings of the executive, decides to be more vigorous in its efforts. It agrees to aggressive improvement targets after a high-level planning effort and decides it must formalize its approach to meet them. Council members also recognize the need for *vision* and *policy,* and devote time over several months crafting them.

In parallel, our executive decides it is time to inform the CEO of his/her unit's data quality program. He or she naturally wants *recognition,* both personally and for those who led improvement efforts. More importantly, projects increasingly need the cooperation of other units, which are reluctant to provide it (importantly, few data quality systems impact beyond the span of control of their most senior sponsor). And the cycle repeats itself across the enterprise.

In time, the enterprise and executive discover the techniques described herein. They will also learn some important features of data.

PROPERTIES OF DATA

In many respects, the techniques used by the executive and enterprise mentioned previously are as described throughout this handbook. But data differ from other resources in some critical ways, and these differences can have important, yet subtle, impact on the data quality program. These differences (Levitin and Redman 1998) include the following:

Consumability: Unlike other resources, data are consumed with use.

Copyability: Data records can be copied for a fraction of the cost of the original. You simply can't do this with other resources.

Computer Storage: Data, unlike other resources, can be stored cheaply and easily in almost unimaginable quantities on computers.

Depreciability: Data, can, in principle, be immortal. While their value does not diminish with use, the utility of most data deteriorates rapidly in time.

Fragility: Paper data records are occasionally accidentally destroyed, but they are more apt to be lost. Data stored in computers are much more fragile.

Intangibility: Perhaps the most obvious difference between data and other resources is data's intangibility. While data recordings can be seen and touched, data themselves are intangible.

Nonfungibility: Fungibility means that one unit of a resource can easily be exchanged for another, assuming another unit is available. But data "units" are inherently unique. You simply cannot substitute one person's date of birth for another's.

Renewal: Whenever pertinent features of the real-world change, data values change and/or new ones are created. New data result from everyday business—each customer transaction, each shipment, indeed, practically every activity leads to new data. This happens at astounding rates. This property of data, called "renewal," does not really apply to other resources. In most cases, there is an inherent lag time until all databases are updated.

Shareability: To a larger degree than any other resource, data may be shared.

Source: In contrast to other resources, data are generated by a tremendous number of sources. In many cases, the original sources of many data sets are undocumented or even unknown. The Internet is exacerbating this problem.

Transportability: Data are also unlike any other resource in the degree, ease, and speed with which they can be transported over long distances.

Valuation: Neither markets nor standard accounting practices exist for most data.

Versatility: Data collected and used for one purpose are often used in other applications. Data-driven marketing is one such example. But some alternate uses of data are illegitimate. Data about a person's age, for example, cannot be used in a hiring decision.

Implications for Data Quality. The following are some of the more obvious ways that properties of data influence the data quality program. First, that which is "out of sight is often out of mind." As data are stored neatly away in computers, there seems to be a tendency to pay them less attention, particularly given the other compelling and highly visible problems facing the enterprise. So many enterprises are not even aware of their data quality issues. This problem is exacerbated by the lack of accepted methods of valuation. It is much easier to spur management action when there are clear monetary costs or benefits.

Second, more than anything else, the high rates of data creation (renewal) ensure that error detection and correction won't work well.

Third, since data are not tangible, they have no physical properties. This complicates measurement. Managers and technicians know how to measure physical properties such as length, viscosity, time, and impedance. But all important data quality dimensions are abstract and so are difficult to measure. For example, you can't tell by direct examination whether most data are correct.

A number of properties help create difficult political situations [see Davenport et al. (1993), Strassman (1994)]. It is interesting that while data are shareable, this is the exception rather than the rule in most enterprises. Instead, since data are relatively inexpensive to store, copy, and transport, organizations within an enterprise often acquire, store, and manage their own. This immediately raises issues of ownership—issues that are among the most brutal in many enterprises. It also makes any kind of centralized planning for data and standards difficult to establish and enforce.

A number of properties (copyability, transportability, nonfungibility, fragility, and cost of storage) conspire to increase redundancy and contribute to a "save everything" mentality, including data that are no longer useful. Yet all redundancy is not bad. For example, redundancy allows users to work with data in the environments of their choosing (for example, it helps make the office-at-home feasible).

BUSINESS CASE FOR SECOND-GENERATION DATA QUALITY SYSTEMS

Second-generation data quality systems make good business sense. The case is summarized as follows:

- Data are used by every activity conducted by the enterprise.
- Most data are of poor quality.
- Current efforts to find and correct errors don't work well.

- Left alone, the problem will become more critical as data become even more important. The Internet exacerbates this problem.
- Second-generation data quality systems cost less and produce better results.

The following subsections explain these points in more detail.

Data Are Ubiquitous. In the previous example, the "call to action" stemmed from poor decisions. It is axiomatic that decisions will be no better than the data on which they are based. Nor should one expect any other activity that takes poor data as inputs to yield superior results. And indeed, data are ubiquitous—virtually every activity in which the enterprise engages requires data. Consider the following:

- Data are both critical inputs to and outputs of almost all "work" performed by an enterprise. Data are used to serve customers, manufacture products, manage inventory, and so forth.
- Data support managerial and professional work and, as in the previous example, are critical inputs to almost all decision making at all levels of the enterprise. Data are the means by which the enterprise knows about its other resources—financial, human, and so forth.
- Data may be combined in almost unlimited ways in the search for new opportunities, market niches, process improvements, and new products and services.
- Because definitions of common terms like "customer" and "service" are captured in data, they (data) contribute to the enterprise's culture. They "fill the white space" in the organization's chart.
- Enterprises strive to convert tacit "knowledge" into data. For example, a company's salespeople may have warm personal relationships with the company's most important customers. But for the company as a whole to serve these customers, important aspects of the relationships must be specified in data. (Some authors have noted that "data" are the raw material for "information," which in turn are the raw materials for "knowledge." The reverse direction is even more important. Specifically, knowledge created or developed by an individual or group must eventually become structured data so others can apply that knowledge.)

It is interesting to note that these activities can be taking place simultaneously, using the same data. Figure 34.1 depicts a typical scenario. The end-to-end process of information supply, new data creation, processing, and use is called an "information chain." "Information products" are the outputs along the way—the data recorded in databases, reports, analyses, and so forth.

Most Data Are of Poor Quality. At the time of this writing, the best known data quality problem is the so-called "year 2000 (Y2K) problem." Examples of other common data quality problems include:

- *Low Accuracy:* Accuracy is probably the most carefully studied data quality issue. Numerous studies yield error rates ranging from 1 to 75 percent. Direct mailers find that up to 20 percent of their flyers are returned as undeliverable; customers find billing errors; scanners report incorrect prices; etc.
- *Inconsistency:* Data values in two databases, "owned" by two organizations within a bank, disagree. The two organizations cannot determine who their common customers are.
- *Difficulties in interpretation:* A clothier may provide shirts in four sizes: small, medium, large, and extra large. In the computer system supporting manufacturing, these sizes are coded "1," "2," "3," and "4," respectively. Later, operators can't recall whether "1" means small or extra large.
- *Unmanageable redundancy:* Many enterprises have any number of copies of the same data. L. P. English (1998) cites a company that had 43 such redundant databases. In addition to adding complication and expense, redundancy makes it more difficult to maintain consistency.

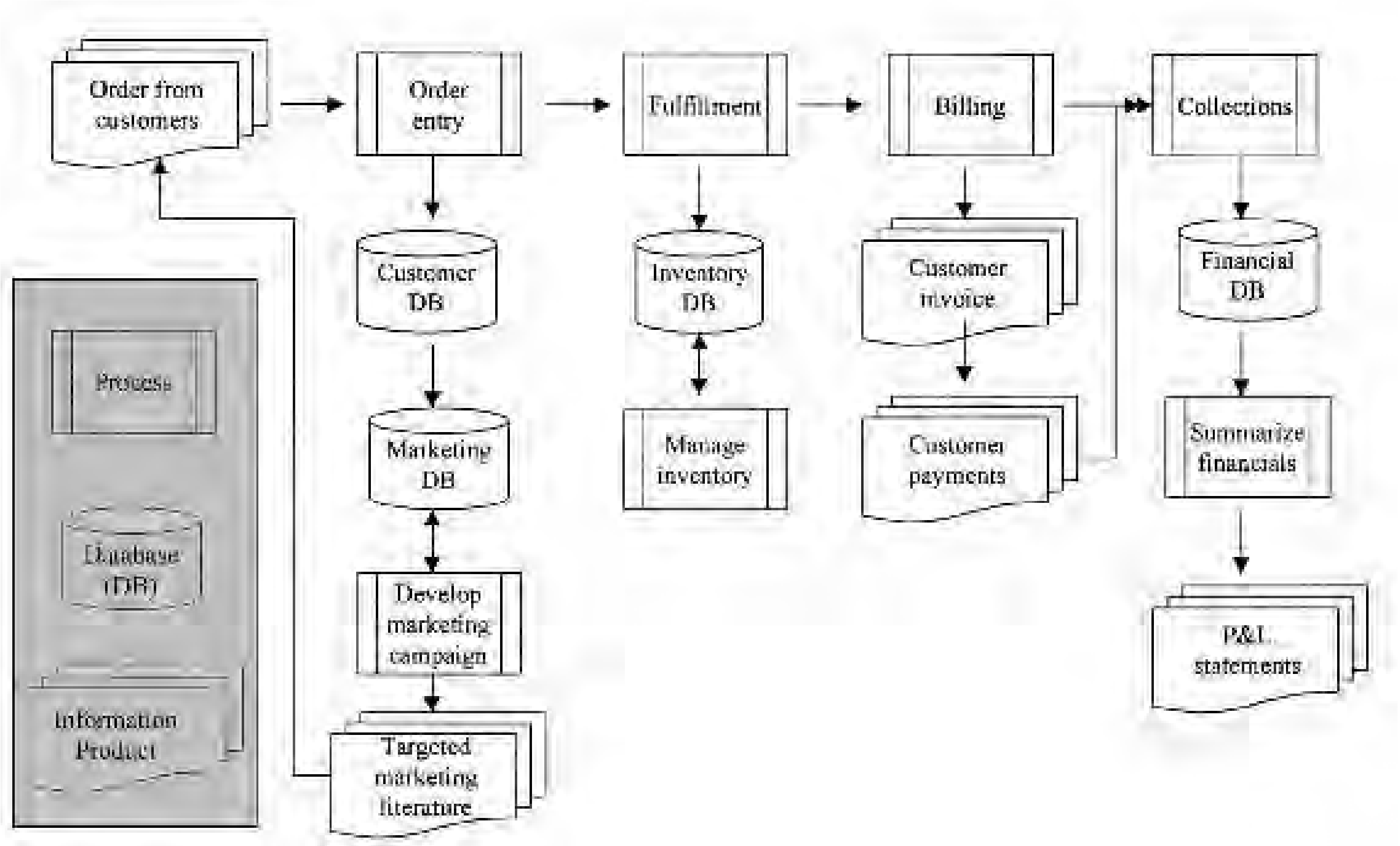

FIGURE 34.1 An example of an information chain (or chains depending on how the enterprise chooses to manage) and the numerous recodings in databases and information products produced.

Poor quality data seem to attack information chains like viruses—there is no way to predict exactly where they will strike next or the impact they will have.

Error Detection and Correction Yield Poor Results. Naturally many enterprises are not blind to poor data quality. So they undertake efforts to detect and correct errors. Direct inspection is one way of doing so. Many organizations review the data supplied to them by upstream organizations to identify data they "know are wrong" because the values are outside accepted domains. Others search for inconsistencies in a collection (or collections) of data. Thus, if a customer's telephone area code = 999, the value cannot be correct. Or if the telephone area code = 212 (New York) and address zip code = 90210 (California), then at least one value must be in error. The search for inconsistencies can be quite sophisticated, involving numerous fields, several databases, and sophisticated error logic. These searches may be computerized and a number of good software programs are available. Once errors have been found, they must be corrected and in some cases, error correction logic helps do so. But correcting errors is often more difficult than finding inconsistencies. In some cases the company must make reference to the real world to determine correct values.

Even conscientiously applied efforts to detect and correct errors do not yield satisfactory results. The litigation alone from noncompliance in the Y2K problem may cost $1 trillion in the United States (Thompson 1997). Of course, "everyday" problems have costs also. The total costs across the typical enterprise are summarized in Table 34.1.

Some costs are stated in quantitative financial terms. Those affecting customers or employees are not so easily quantified but may be even more important. First, is the impact of poor data on customer satisfaction. Customers receive much data as a byproduct of the product or service they receive. The invoice for product sent is a good example. Many customers are remarkably

Table 34.1 The Costs of Poor Data Quality to the Typical Enterprise

Typical problems
Inaccurate data
Inconsistencies across databases
Inappropriate formats
Difficulties in interpretation
Unmanaged redundancy
Typical costs*
Operational costs:
Lowered customer satisfaction
Increased cost: 8–12% of revenue in the few, carefully studied cases; for service organizations, 40–60% of expense
Lowered employee satisfaction
Tactical costs:
Poorer decision making: Poorer decisions that take longer to make
More difficult to implement data warehouses
More difficult to “mine” data and re-engineer
Increased organizational mistrust
Strategic costs:
More difficult to set and execute strategy
Contribute to issues of data ownership
Compromise ability to align organizations
Divert management attention

*Operations, tactics, and strategy represent a loose hierarchy of work performed. Operations are the day-to-day tasks such as order entry, customer support, and billing. Tactics are the decisions of short- and mid-term consequence and work to sup port them usually made by middle managers. Strategy involves long-term direction.

Source: After Redman (1998).

unforgiving of billing (and other data) errors, reasoning that any company that can’t bill them properly simply can’t provide a good product or service.

Second is the impact on employees. One cannot expect the hotel clerk, dealing with irate travelers whose reservations have been lost, to have high job satisfaction. And many enterprises have dozens, even hundreds, of such “customer care” jobs. This issue occurs at the organizational level as well. When one organization depends on another for input data and those data are frequently wrong, it is natural for the former to develop a poor opinion of the latter. And their ability to work together in the future is compromised.

The Importance of Data Will Continue to Grow. Data have always been important to commerce and their importance continues to grow rapidly. They are the fuels for economic growth in the Information Age.

Increasingly, important data are computerized. This trend has accelerated for several decades, driven by impressive advances in database, networking, and communications technologies. It has made data available to new, unsophisticated users and exacerbated issues of poor data quality. Many people have a tendency to assume that “if it is in the computer, it must be right.” Such people are more easily victimized by poor quality data. Even sophisticated users cannot be expected to have familiarity with the nuances of all the data they encounter. The Internet is taking these phenomena in directions that are not fully understood. Specifically, companies are making their data, heretofore closely held, directly available to Internet users, a remarkably diverse group with disparate needs that are likely to be quite different from those of internal users. Thus, data quality issues too will become both more critical and difficult as a result.

Second-Generation Data Quality Systems. In recent years superior approaches that focus not on individual errors, but rather on identifying and eliminating root causes of entire categories of errors have been developed. Companies find they can make order-of-magnitude improvements, often by eliminating some rather simple (once they have been identified) problems. After Ishikawa (1990), we call techniques aimed at error detection and correction "first-generation techniques" and those aimed at eliminating root causes "second-generation techniques."

By a *"data quality system"* we mean the totality of an enterprise's efforts to manage, control, and improve data quality. The system includes:

- Activities aimed at understanding customer needs
- Activities aimed at detecting and correcting errors and activities aimed at preventing future errors (including control and improvement)
- Activities to build management infrastructure to do so effectively and efficiently

Second-generation data quality systems are those that emphasize prevention of future errors over error detection and correction and build management infrastructure to do so effectively and efficiently.

Practitioners have learned that second-generation data quality systems cost far less and produce much better results. Consciously applied, these systems have helped many organizations reduce error rates by factors of 10 to 100, cut cycle time in half, and reduce costs by up to 75 percent. Customer service is better and employees feel much greater involvement in their jobs, so morale improves. These results, summarized in Table 34.2, stand in marked contrast to the baseline of Table 34.1.

DATA DEFINED AND DIMENSIONS OF DATA QUALITY

Data Defined. As used here, "data" per se (or a "data collection," "data set," etc.) consist of two interrelated components, "data models" and "data values" (Fox and Redman 1994). Data models are the definitions of entities, attributes, and relationships among them that enterprises use to structure their view of the real world. Enterprises often model a given portion of the world

Table 34.2 Typical Results from Implemenation of a Second-Generation Data Quality System

Typical improvements
Accuracy improved by 1 to 2 orders of magnitude
Easy-to-read formats
Redundancy and inconsistency minimized
Typical benefits
Operational benefits:
A primary source of customer dissatisfaction eliminated
Decreased cost: Two-thirds to three-quarters of cost of error detection and correction eliminated
Important cycle times cut in half
Employees feel empowered
Tactical benefits:
More confident decision making
Re-engineering opportunities suggested
New technology implementation easier
Organizations build trust by working together
Strategic benefits:
Issues of data ownership eased
Easier to set and execute strategy

differently. For example, an employer may be interested in a person as an *employee,* the IRS as a *taxpayer.* The person is the same real-world entity in both cases, the employer and IRS are interested in different attributes. Data values are the specific realizations of an attribute of the data model for specified entities. The "123-45-6789" in taxpayer social security number = 123456789 is a data value.

"Data records," as distinct from data per se, are the physical realizations of data stored in paper files, in spreadsheets, in databases, and so forth and presented to users in ways that (one hopes) make them easy to store and use. Note that data (i.e., the models and values) are abstract, while data records are their tangible realizations.

It is evident that data are not just random facts and figures, but rather are quite structured. Especially for computerized data, several steps must be completed before they can be used. These steps, summarized in Figure 34.2, include model development, acquisition of the values, storage, selection of what the user wants to see, and presentation to the user. All can impact customer satisfaction, though some steps are more pertinent to the computer systems, than to data. And many of them are carried out by separate organizations. Thus data models are the responsibility of those who develop databases, variously called Information Technology, Information Systems, or Information Management Systems (all abbreviated IT herein). Data values are created within or obtained from external suppliers by line organizations or business units in the course of everyday business. IT is also usually responsible for the manner in which data are recorded and presented to users.

The most important common dimensions of data quality are discussed below.

Dimensions of Data Quality. "Dimensions" of data quality stem from user needs. The "properties" described previously are intrinsic to data. Customers naturally have needs or requirements that bear on each major constituent of data (model, values, and records). It is important to distinguish closely related needs, such as accessibility, that bear more on supporting technology, from those directly pertinent to data. Table 34.3 lists the most important "dimensions of data quality" [The definitions of many of these dimensions are quite technical and the interested readers are referred to Fox and Redman (1994) and Levitin and Redman (1995).]

To summarize, high-quality data are data that are fit for use in their intended operational, decision-making, planning, and strategic roles. As Figure 34.3 depicts, data models and presentation define the features of high-quality data. And data values must be free of defects. Supporting technologies must make data secure yet accessible and ensure privacy.

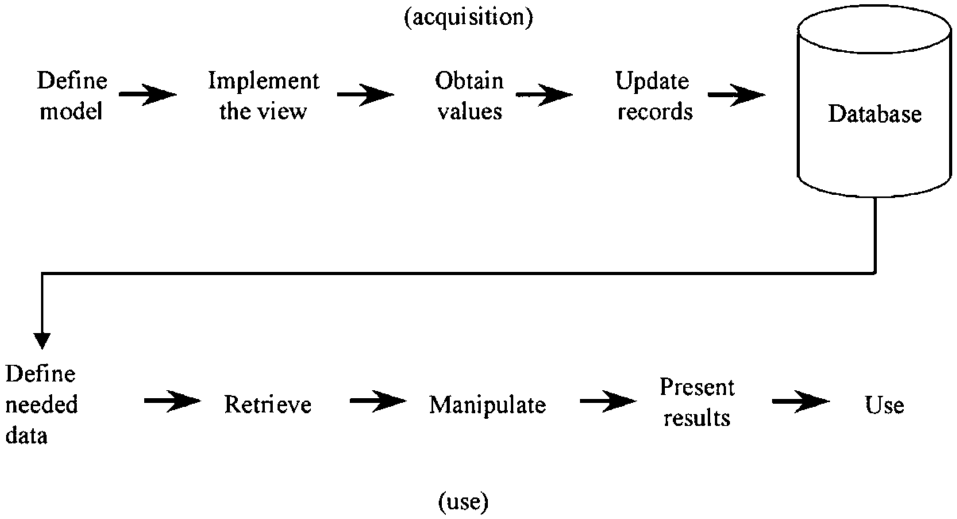

FIGURE 34.2 The data life-cycle model features activities needed to define a data model, obtain values, store and process data, and present the user with what he/she wants.

Table 34.3 Dimensions of Data Quality*

Quality dimensions of a data model		
Scope	*Comprehensiveness*	*Essentialness*
Level of detail	*Attribute granularity*	*Precision of domains*
Composition	*Naturalness*	*Identifiability*
	Homogeneity	*Simplicity*
Content	*Relevance*	*Obtainability*
	Clarity of definition	
View consistency	*Semantic consistency*	*Structural consistency*
Reaction to change	*Robustness*	*Flexibility*
Quality dimensions of data values		
	Accuracy	*Completeness (entities and attributes)*
	Consistency	*Currency (cycle time)*
Quality dimensions of data records and presentation		
Formats	*Appropriateness*	*Format precision*
	Use of storage	*Correct interpretation*
	Flexibility	*Portability*
	Represent null values	
Physical instances	*Representation consistency*	
Other dimensions often associated with data		
	Accessibility	*Appropriate use*
	Privacy	*Redundancy*
	Security	

Source: Fox et al. (1994) and Levitin and Redman (1995). For an alternative formulation, see Wang and Strong (1996).

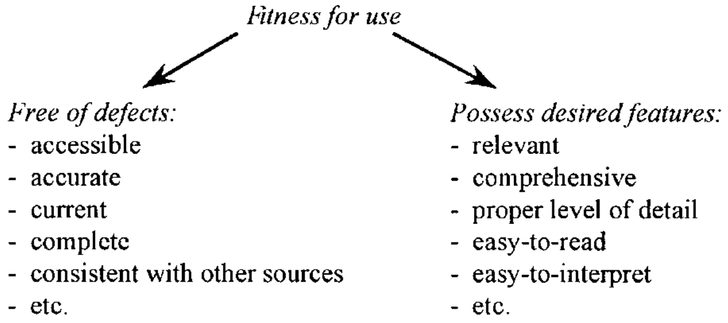

FIGURE 34.3 Data are of high quality if they are "fit for use" in their intended operational, decision-making, and other roles. Fitness implies both freedom from defects and possession of desired features. Most users associate dimensions associated with data models as desired features and the lack of other features as defects.

SECOND-GENERATION DATA QUALITY SYSTEMS

The thrust of second-generation quality systems is to consistently identify and prevent the most important root causes of future errors. This requires both proper technique and management infrastructure. This section describes each element of a second-generation data quality system. (See Table 34.4 for a list.) Each enterprise must craft and evolve its own data quality system. Four elements, senior leadership and support, quality planning, quality control, and quality improvement are required. Other elements are selected based on the enterprise's opportunities and challenges, its culture and organization, and the origin of its most important data.

Table 34.4 Elements of Second Generation Data Quality Systems

Management infrastructure	Process description
Senior leadership and support	Measurement
Data quality vision	Quality planning
Data quality policy	Quality control
Supplier management	Quality improvement
Process management	Process (re-)design
Change management	Inspection and test (data editing)
Database of record	Quality assurance
Strategic data quality management	Document assurance
Training and education	Rewards and recognition
Technical capabilities	Domain knowledge
Identification of information chains	Standards
Customer needs analysis	Quality handbook

Management Infrastructure

Senior Leadership and Support. As we have observed, data cross organizational boundaries in the blink of an eye and the politics associated with data can be brutal. So senior leadership, support, and intervention in the conflicts that are sure to arise are essential. These may be provided by a Data Council (or a Quality Council). Senior management is ultimately responsible for crafting the data quality system and leading its implementation. It ensures that data quality efforts are directed at the most pressing business problems and opportunities. To do so, it develops and deploys a vision and policy (see the following two paragraphs), sets quality goals, selects planning, control, and improvement "projects," and provides cross-functional coordination for projects that require it. It allocates funds for data quality efforts and ensures that the enterprise is properly trained.

Data Quality Vision. People need to know where the enterprise "is going." A vision is a "picture" of the enterprise's desired future state with respect to data and information, including a rationale for people to work to create that future state. Developing the vision forces the Data Council to think about the enterprise's long-term data and information needs to ensure business success. A vision, supported by broad communication, helps align the enterprise and motivates people to work together to achieve the desired future state.

Data Quality Policy. A data quality policy (or simply data policy) is management's statement of intent regarding long-term data and information quality. It serves as a "guide for managerial action." It should recognize data as "business assets" and, accordingly, specify improvement objectives and management accountabilities for achieving them. The best second-generation policies delineate accountabilities along information chains from those that create data models to creators of data values to data users (Redman 1995, 1996). Like the vision, formulating policy forces the Council to think broadly and deeply about management accountability for data. Widely communicated policies help align the enterprise and provide a basis for decision making by lower-level managers.

Supplier Management. Almost all enterprises receive data from outside. Some is purchased, while other data, such as an invoice, are the byproduct of other goods and services. The quality of these data is extremely important. Supplier management is the overall program for managing suppliers, including selecting them, ensuring that they understand what is expected, measuring performance against expectations, and negotiating improvements to close gaps.

Process Management. Well-managed information chains ensure that data values created throughout will be of high quality. Process management (see Section 6) provides the infrastructure and technique needed to do so. Experience confirms that most individual functions within a single organization work fairly well. But "problems" and/or opportunities occur when organizational boundaries are crossed (and as noted in Figure 34.1, data cross organizational lines in the blink of an eye). Process management focuses specifically on organizational interfaces. It provides a structured framework for utilizing many of the techniques presented in the next subsection and is a proven method for making and sustaining improvements to data.

Change Management. Most enterprises/organizations have first-generation data quality systems. Second-generation systems require them to think and act differently and change is always difficult. Experience shows that when change is managed, the risks can be reduced (Kotter 1996). Enterprises are advised to be conscious of these issues and actively address them.

Database of Record. Most enterprises have too much redundant data. But how to manage and reduce redundancy? It is natural to try to do so by recognizing an official "standard" data source to be used throughout the enterprise. But people have resisted what they perceive to be authoritarian directives about data, particularly when the designated source is difficult to use, inaccessible, out-of-date, or otherwise of low quality. A database of record addresses this issue directly. It provides a set of quality standards (for example, accuracy greater than 99.5 percent) and a manager ("custodian," "keeper," or "steward") charged with ensuring that the source satisfies them. Only then is the database designated as the "approved master source" for that data.

Strategic Data Management. In most enterprises, today's data do not meet the enterprise's current needs. And all expectations are that the typical enterprise's data needs will grow exponentially in the future. Strategic data management aims to ensure that the enterprise's top-line business strategy is "data-enabled," that the enterprise has the data and information assets (especially data sources, information chains, and the ability to exploit them) to effect its strategy.

Training and Education. Management must ensure that all involved have the knowledge, skills, and tools needed to improve data quality.

Technical Capabilities. We now turn our attention to the technical capabilities needed to focus improvement efforts, make and sustain gains, and ensure continuity.

Identification of Information Chains. Not all data are created equal. The enterprise should define and execute a process (either formal or informal) of identifying the data and information chains most critical to the enterprise. The data quality program should be focused on these assets.

Customer Needs Analysis. The customer is the final arbiter of quality, so understanding who they are and their needs, prioritizing those needs, and communicating to those who need to know is essential. Good second-generation systems focus incessantly on users of data, formally document and keep customer needs, and keep them current.

Process Description. Although data cross organizational boundaries in remarkably confused paths, managers need to understand these paths. Process description is the means of acquiring and documenting this understanding. There are many ways to conduct this process. Good descriptions of information chains include suppliers, the steps taken to produce information products, customers, and all other essential aspects (including data, organization, supporting technologies, etc.) that may impact performance. The simple act of describing what is actually happening often brings incongruities to the surface, and these incongruities become opportunities for improvement.

Measurement. Management of data quality proceeds on the basis of fact (see also Section 9). And facts are needed at many levels of the enterprise. At a low level, measurement is the process of quantifying information chain performance, including aligning measures to customer needs, specifying the measurement protocol, collecting data, and presenting results. Good second-generation practice emphasizes measurement *within information chains* at the points of new data creation. "Data tracking" is one method to do so (see Redman 1995, 1996). It also helps solve the intangibility issue noted previously. At the enterprise level, measurement also refers to the overall system or collection of measurement processes and management summaries to track overall progress. One important overall measurement may be the cost of poor data quality.

Quality Planning. At the enterprise level, planning is the regular (e.g., annual) process of setting quality goals or targets and/or improvement and putting in place the means to achieve those goals. At the "project" level, planning is a team-oriented process that creates or replans new information products, information chains, or controls to meet specific customer needs. The first steps of a re-engineering project also constitute quality planning.

Quality Control. Quality control is the process of evaluating (quality) performance, comparing that performance with standards or goals, and acting on the difference (see Section 4). Establishing and maintaining control is essential because it provides the basis for predicting that errors will not occur in the future.

Quality Improvement. Good second-generation practice calls for use of a structured team process for reducing errors and other deficiencies in information chains and information products. This process involves identifying and selecting improvement opportunities (projects), chartering teams to make improvements, completing those projects, and "holding the gains" (see Section 5). Like the other elements of the Juran trilogy, quality improvement is essential to second-generation data quality systems. Participating in and completing projects is the vehicle by which most people are inculcated into the "quality culture."

It is important to recognize the proper role of computer, networking, and database technologies in data quality improvement. These technologies have been quite effective in enabling well-established and well-managed information chains to perform faster and cheaper. But technology alone [see also Landauer (1995) and Strassman (1997)] is not the key to improving data quality. Indeed an overreliance on technology appears to exacerbate inadequacies. Information chains must be put into reasonable working order before applying the newest technology. For years, quality gurus have advised against automating ineffective factories. So this prescription simply extends that time-tested maxim to data.

Process (Re-)design. Some information chains are fated to yield poor results because they are poorly designed. A chain in which raw data must be manually rekeyed into several computers is a good example. It simply won't work effectively or efficiently. The planned "blueprint" of an information chain (including suppliers, the sequence of work activities, interfaces, supporting technologies, management accountabilities, and information products delivered to customers) should incorporate principles of good design. For example, the tools needed by the process owner (i.e., measurement and control) should be incorporated into a process design. Other best practices are given in Redman (1996, Chapter 14).

Inspection and Test (Data Editing). Data editing is the process of determining if data values satisfy consistency criteria. Editing is usually applied in data clean-ups and, as such, is more properly a first-generation technique. Edits may be employed in second-generation systems at the points of data creation and entry to prevent defective data from proceeding further. Failures (and corrections) must be counted and classified and used in control and improvement projects. Editing may also be employed as part of a supplier program and is sometimes necessary for complex information products.

Quality Assurance. A quality assurance program consists of audits to determine the degree to which the data quality system, as designed, is deployed and functioning.

Document Assurance. Procedures are needed to control documents and data/information that relate to requirements of the data quality system. In particular, a master list of all current policies, procedures, and results should be accessible to those that need it.

Rewards and Recognition. As noted previously, second-generation data quality systems require people to think and act differently. Rewards and recognition that provide the reinforcement and feedback of superior performance should be part of the enterprise's merit/compensation rating system. Specifically, compensation increases and promotion decisions should take employees' contributions to data quality into account.

Domain Knowledge. The enterprise should take care to learn more about data as a resource (see Levitin and Redman 1998), the data they use, the way data create value, and techniques to manage them. As we have noted, data are different from other resources. Intimate knowledge of their properties is critical.

Standards. Standards are accepted definitions, rules, and bases of comparison, usually developed and agreed-upon by a body with authority to do so. Standards can advance a data quality program. For example, a standard definition of common terms such as "customer" could help reduce the number of partially redundant and disparate databases. But standards have been very difficult to effect in most enterprises. The various organizations just can't seem to agree on a common definition. The root cause of the issue seems to be that the various organizations think differently about their customers and their thoughts are captured in their data models. They fear that a standard definition would compromise their relationships with customers.

There is no easy resolution to this issue. As a practical matter, it is usually best to postpone work on standards until other elements of the enterprise's second-generation data quality system are in place. And the Data Council should first gain some experience with relatively less contentious standards.

Quality Handbook. As a quality system matures, it is appropriate to codify it into a published "book" containing an enterprise's quality policy, important concepts and definitions, and procedures. Ideally the handbook is customized to the enterprise, general enough for widespread use, yet specific enough to help focus the enterprise's efforts.

CONCLUSIONS

An important analogy likens a database to a lake. Water represent the data and the streams feeding the lake represent information chains. Animals who drink from the lake and others who enjoy the lake represent users. Presented with a polluted lake, the community has three choices:

- It can suffer the consequences.
- It can filter the lake water.
- It can eliminate sources of pollution upstream.

Experience confirms that the best results are obtained with the third approach. For data, this analogy explains precisely why second-generation data quality systems are so successful.

ACKNOWLEDGEMENTS

The author wishes to thank Bill Barnard of the Juran Institute, Bob Pautke of AT&T Labs, and Charles and Joan Redman of Nashville, Tennessee, for their insights, suggestions, and improvements to this section. The author also wishes to thank his wife Nancy for carefully proofreading drafts.

REFERENCES

Davenport, T. H., Eccles, R. G., and Prusak, L. (1992). "Information Politics." *Sloan Management Review,* vol. 34, no.1, pp. 53–65.

English, L. P. (1998). "The High Costs of Low Quality Data." *DM Review,* January.

Fox, C., Levitin, A., and Redman, T. (1994). "The Notion of Data and Its Quality Dimensions." *Information Processing and Management,* vol. 30, no. 1, pp. 9–19.

Ishikawa, K. (1990). *Introduction to Quality Control.* 3A Corporation, Tokyo.

Kotter, J. P. (1996). *Leading Change.* Harvard Business School Press, Cambridge, MA.

Landauer, T. K. (1995). *The Trouble with Computers: Usefulness, Usability, and Productivity.* The MIT Press, Cambridge, MA.

Levitin, A. V., and Redman, T. C. (1993, 1995). "A Model of Data (Life) Cycles with Application to Quality." *Information and Software Technology,* vol. 35, no. 4, April, pp. 217–223.

Levitin, A. V., and Redman, T. C. (1995). "Quality Dimensions of a Conceptual View." *Information Processing and Management,* vol. 31, no. 1, January, pp. 81–88.

Levitin, A. V., and Redman, T. C. (1998). "Data vs. Traditional Resources: Properties, Implications, and Prescriptions for Management." submitted to *Sloan Management Review.*

Redman, T. C. (1995). "Opinion: Improve Data Quality for Competitive Advantage." *Sloan Management Review,* vol. 36, no. 2, winter, pp. 99–107.

Redman, T. C. (1996). *Data Quality for the Information Age.* Artech House, Norwood, MA.

Redman, T. C. (1998). "The Impact of Poor Data Quality on the Typical Enterprise." *Communications of the ACM,* vol. 41, no. 2, February, pp. 79–82.

Strassman, P. (1994). *The Politics of Information Management.* The Information Economics Press, New Canaan, CT.

Strassman, P. (1997). *The Squandered Computer: Evaluating the Business Alignment of Information Technologies.* The Information Economics Press, New Canaan, CT.

Thompson, R. J. (1997). "The Unspeakably High Cost of Noncompliance." *InformationWeek,* June 30.

Wang, R. Y., and Strong, D. M. (1996). "Beyond Accuracy: What Data Quality Means to Data Consumers." *Journal of Management Information Systems,* vol. 14, no. 4, spring, pp. 5–34.