Principles of Information Quality Improvement

“Back of every noble life there are principles that have fashioned it.” 
—GEORGE HORACE LORIMER

The three chapters of Part One describe the fundamental principles of information quality. This is not theory. These are very real and practical principles, even though they are foreign to many organizations. They provide the basis for understanding the background to information quality improvement as a management tool. They are the basis for the processes of information quality improvement described in Part Two, “Processes for Improving Information Quality.” Without understanding the principles of quality improvement, implementing the processes may be a hollow and empty exercise that performs the actions but lacks the soul. This may result in loss of motivation for any information improvement initiative, no matter how well intentioned.

Chapter 1 describes the business case for information quality improvement. The bottom line is that poor data quality is just too expensive for organizations in a competitive or tight economy. It describes why information initiatives, such as data warehouses, so often fail.

Information systems organizations are in crisis today, a crisis caused by using information technology in ways that add complexity to information processing and information management based on industrial-age paradigms. This compounds information quality problems by creating redundant databases.
Chapter 1 presents many examples of the high cost of low-quality data. The result is that failure to solve information problems can be fatal to organizations.

Chapter 2 defines information quality. It first defines what quality is and is not. In order to understand information quality, data and information must be defined. The chapter then defines knowledge and wisdom, which is where information impacts business performance.

In defining information quality, we differentiate between inherent and pragmatic information quality. Essentially, inherent quality is the correctness of facts, and pragmatic quality is the correctness of the right facts. Chapter 2 defines the three components required for information quality: data definition and information architecture quality, data content quality, and data presentation quality.

Chapter 3 describes the principles of quality in general: customer focus, continuous process improvement, and use of scientific methods. It briefly outlines several quality approaches to illustrate the common themes of quality principles. Included in this discussion are encapsulations of Deming's 14 Points of Quality, the Juran Trilogy, Ishikawa's quality control as a movement, Kaizen, Quality Function Deployment, Crosby's 14 Steps, ISO 9000 quality management system standards, and the Baldrige Framework of Seven Categories for Business Performance Excellence.

Chapter 3 then describes how these quality principles apply to information as a product, and knowledge workers as information customers. The stewardship roles in information quality are discussed briefly. Everyone in the enterprise has accountability for quality of information. Chapter 3 then describes the notion of “customer service” of information products in the information value chain, and concludes with a list of the fundamental principles of information quality.
In this chapter I describe the reason why an organization is—or should be—interested in information quality. It can be summed up in one word: profit. Profit, however, is only a byproduct. Profits come when we know and focus on customers’ needs and provide quality products that meet those needs. When information products fail to meet customers’ needs, profits go down. Information systems and data warehouses fail, squandering the investments.

We describe why data that appears to be of satisfactory quality is, in fact, not. We illustrate the huge costs incurred as a result of low-quality data. We illustrate examples of the costs, including enterprise failure.

There is and must be only one purpose for improving information quality: to improve customer and stakeholder satisfaction by increasing the effectiveness and efficiency of the business process. Information quality is a business concern, and information quality improvement is a business issue. Information quality improvement actually reduces business costs by eliminating costly scrap and rework caused by defective data. It increases business profits by providing more reliable information products that result in more usage, better decisions, and increased exploitation of business opportunities.

Unfortunately, the state of information quality in most organizations’ databases is so abominable that if the same level of quality existed in their products and services, they would go out of business. “Justify that statement!” you say. You can do it yourself by answering two questions:

"Quality is Free. . . . What costs money are the unquality things—all the actions that involve not doing jobs right the first time."
—Philip Crosby
How many private, proprietary databases and files that reside on personal computers (in spreadsheets, PC databases, and even in word processor files) in your enterprise include information contained in corporate databases or files that are not integrated with and synchronized to those corporate databases?

If the data in those corporate databases is high quality, why is there a need for those redundant, private databases? After all, data is the only business resource that is completely reusable without being used up.

Why Data Warehouses Fail

Many see data warehousing as the silver bullet out of the operational data abyss. Not! If data warehousing is approached with the same information and (mis)management principles that have produced the disintegrated islands of automation legacy environment, it will fail. It will fail spectacularly. In fact it will deserve to fail.

Data warehousing projects fail for many reasons, all of which can be traced to a single cause: nonquality. Poor data architecture, inconsistently defined departmental data, inability to relate data from different data sources, missing and inaccurate data values, inconsistent use of data fields, unacceptable query performance (timeliness of information), lack of business sponsor (no data warehouse customer), and so forth, are all components of nonquality.

With all of the emphasis on data warehousing technologies, it will serve you well to remember two things:

The product of the data warehouse is information.

The customers of the data warehouse are the knowledge workers who must make increasingly important decisions faster than ever before.

If the data warehouse does not deliver reliable information that supports the customers' decisions and strategic processes to their satisfaction, history will repeat itself.

The Information Quality Crisis

If the state of quality of your company's products and services was the same level of quality as the data in its databases, would your company survive or go out of business? One insurance company had a list of 12 "sacred data elements" that were considered so important that if the data was wrong, the company could fail. When it did a data inventory, it discovered that this data element was maintained in 43 separate databases by 43 independent applications, with data entered by 43 different information producers.
One manufacturing firm had 92 Part files, many defined with different primary identifiers so that the same part in different files could not even be cross-referenced.

A major bank had 256 different Customer files that it had to analyze just to answer the question, “Who is our best customer?”

A consumer goods company discovered it had over 400 Brand files containing product information.

Topping the list, however, is a Telecommunications provider that is the ultimate information schizophrenic with over 800 Customer files.

If the data in those corporate databases is high quality, why is there a need for the redundant, disparate databases that seem to multiply like rabbits? Data is the only business resource that is completely reusable. All other resources, when used, are used up; for example, money can be spent once, employees can perform only one task at a time, raw materials can be used once in the production of a finished good, and facilities can be used for only one purpose at a given point in time.

Yet data, the only nonconsumable resource, is the only resource where high redundancy is accepted as a “legitimate” cost of doing business. The insurance company with 43 different databases and applications capturing the same facts is the information equivalent of accounts payable paying a single invoice 43 times, or Human Resources hiring 43 people to perform the same task 43 different times, or building 43 buildings when only one is needed. Is this the legacy that Information Systems (IS) should provide its enterprise?

The dark side of the business case for data warehousing is the failure of Information Systems to provide for effective data management of the business-critical information resource across its operational applications—and the enterprise is paying for this dearly.

But Our Information Quality Is Not So Bad . . .

After all, the operational processes are running well. That may be, with an emphasis on may. The truth of the matter is that the tactical and strategic process requirements of data warehouse data are completely different from the operational process requirements of data. Consider the following scenario.

An insurance company downloaded claims data to its data warehouse to analyze its risks based upon Medical Diagnosis Code for which claims were paid. The data revealed that 80 percent of the claims paid out of one claims processing center were paid for a diagnosis of “broken leg.” “What is happening here?” was the concern. “Are we in a really rough neighborhood?” No. The claims processors were paid for how fast they paid claims, so they let the system default to “broken leg.” The information quality was good enough to pay a claim
because all the claims payment system needed was a valid diagnosis code. However, that same data was totally useless for risk analysis.

But worse than this is the fact that over the years, the archaic legacy data structures have failed to keep up with the information requirements of even the operational knowledge workers. As a result, because they require more and more information to do their jobs, knowledge workers have been forced to create their own data workarounds and store the data they need in creative ways that differ from the original file structure. The cause of this problem is simple. The Information Systems staff is busy maintaining, on average, a ten-fold redundant databases and the redundant applications or interface programs that recreate or move data. They don't have time to keep a single sharable database current to meet knowledge workers' needs. This represents only the beginning of the information quality challenges facing the data warehouse team.

Why have these issues not been seriously addressed until now? Two reasons: First, information quality is not a sexy topic. After all, who wants to work at the sewage treatment plant when they could be building factories (that create pollution)!? The second reason is insidious: Management has either deemed the costs of the status quo and the current level of low-quality data as acceptable and normal costs of doing business or they are unaware of the real costs of nonquality data.

The Incredible Costs of Nonquality Data

Most organizations have come to accept the level of nonquality data as normal and usual or they are totally unaware of its costs—after all, we are profitable, aren't we? As long as the level of information quality is relatively the same among the competition, the competitive battle lines are drawn in other areas. However, when someone redefines the role of information quality, as the Japanese did with automobile quality, the rules of the game change. The U.S. auto industry's Big Three (GM, Ford, and Chrysler) have been losing ground over the past decade. Their domestic car market share fell from 76 percent in 1984 to 62.5 percent in 1996, an all-time annual low. January 1997 started out worse with the Big Three's domestic auto market share dipping to 59.3 percent, according to industry tracker Autodata.1

General Motors lost a whopping $4.5 billion in 1991 and followed that with an incomprehensible $23.5 billion loss the next year before it got its act together. While GM has regained profitability, with a record $6.9 billion in 1995, its combined profits from the four years 1993–1996 have not erased the loss of 1992, and its market share in the United States continued to slide, from 34 percent in 1992 to 31.6 percent in 1996.2

---

1Micheline Maynard, “Buyers taking a pass on Detroit's passenger cars,” USA Today, February 2, 1997, p. 1B.
2Micheline Maynard, “GM’s report card barely surpasses expectations,” USA Today, November 11, 1996, p. 17B.
GM stockholders can only speculate what their stock value might be today if the American auto manufacturers had not been oblivious to the quality revolution.

The quality revolution has redefined quality from an optional characteristic to a basic requirement for both goods and services. It is no longer sufficient to compete on price alone. Customer satisfaction is the key driver for long-term financial and organizational success today. GM’s new CEO, Jack Smith, admonished employees in October 1996, “We cannot afford the luxury of complacency. Continuous improvement is the name of the game if we want to assure our jobs and the future of this great company.”

The same kind of revolution will happen with information quality, and it will change the economic landscape. Continuous improvement of information products and services will become the name of the game if information professionals want to assure their jobs and the futures of their organizations. Those oblivious to its imminence will suffer; the only question is, “How much?”

Management can no longer afford the luxury of the excessive costs of non-quality data. In the Information Age a quality, shared knowledge resource will differentiate the successful enterprise. Information quality is to the next decade what product quality was to the 1980s.

---

**THE HIGH COSTS OF LOW-QUALITY DATA**

The high costs of low-quality data are ubiquitous. They negatively impact all areas of our lives, personally as well as in our work. Anyone could fill a book with their own personal experiences in which nonquality data has cost them time, money, or bodily injury. Some are dramatic. Consider the following:

- Some Metro Nashville city pensioners were overpaid $2.3 million from 1987 to 1995, while another set of pensioners was underpaid $2.6 million as a result of incorrect pension calculations, according to The Tennessean, March 21, 1998.
- “Two 20-year-old ‘calculation errors’ . . . socked Los Angeles County’s . . . pension systems with $1.2 billion in unforeseen liabilities, and will probably force cash-strapped county officials to spend an additional $25 million a year to make up for insufficient contributions to the fund,” according to the Los Angeles Times, April 8, 1998.
- Trans Union Corp. was ordered by a jury to pay $25 million because of a clerical error that released names of several hundred First National Bank of Omaha customers to other credit card issuers, in breach of a confidentiality agreement (Washington Post, March 10, 1998).
- Ninety-two percent of claims Medicare paid to community health centers over one year's time were “improper or highly questionable,” according to an investigation conducted by the inspector general of the Department of Health and Human Services (Washington Post story reported in The Tennessean, October 7, 1998).

---

3Micheline Maynard, “GM’s report card barely surpasses expectations,” USA Today, November 11, 1996, p. 17B.
THE HIGH COSTS OF LOW-QUALITY DATA (CONTINUED)

■ Wrong price data in retail databases may cost American consumers as much as $2.5 billion in overcharges annually. Data audits showed four out of five errors in database prices read by bar-code scanners are overcharges from the published price of goods.4

■ Four years later, information quality had not significantly improved, with 1 out of 20 items scanned incorrectly according to a Federal Trade Commission study of 17,000 items. As a result, some state and local governments have passed laws requiring stores to put price stickers on items, or face substantial fines (up to $25,000 in Michigan). One Michigan retailer spends $2.4 million a year—11 percent of its payroll—to affix price tags on items.5

■ The U.S. Attorney General’s office has stated that “approximately $23 billion, or 14 percent of the health care dollar, is wasted in fraud or inaccurate billing.”6

■ According to The Financial Times, information quality problems were a factor in a $770-million pretax loss suffered by an investment firm in 1994, causing the company to write off $217 million in 1994 as a result of “bookkeeping errors.”7

■ Inaccurate data about one constituent cost a municipality a $2.5 million lawsuit.

■ A suspect in a kidnapping/homicide incident was accidentally released after posting a low bond for misdemeanor charges, because it was not known he was wanted for the kidnapping and shooting death of an 18-year-old. For whatever reason, only the hold order for the lesser charge followed the suspect in transferring from one jurisdiction to another. The sheriff’s department, police department, and warrants officials are now working together “to improve the computer database system and communications with other jurisdictions” (The Tennessean, June 19, 1998).

■ A physician in Florida amputated the wrong leg of a patient. The original order had been changed as to which leg was to be amputated, but the doctor, while following standard procedures before performing an amputation, followed the old order. The nurse who was aware of the changed order had left the operating room before the amputation, but assumed the doctor was aware of the change. Three years later, the same doctor failed to verify the name on a patient’s wrist-band and performed a risky procedure on the intended patient’s roommate! His license was suspended (The Tennessean, July 12, 1998, p. 7A).

■ A European company discovered through a data audit that it was not invoicing 4 percent of its orders. For a company with $2 billion in revenues, this meant that $80 million in orders went unpaid.

■ A petroleum exploration company drilling a new well in the North Sea drilled through the well shaft of a neighboring well because of flawed data that misidentified the well shaft’s exact location. Fortunately, the well was no longer producing oil. Had it

---

been, the pressure from the oil in the ruptured pipe would have gushed up the drilling well’s shaft, blowing the $500-million drilling investment to smithereens, and surely causing fatality to the crew.

- In 1992, 96,000 IRS tax refund checks were returned as undeliverable due to bad addresses.
- No fewer than one out of six U.S. registered voters on voter registration lists have either moved or are deceased, according to an audit comparing voter registration lists with the U.S. Post Office change-of-address list.
- Until January 1998, when new information quality processes were put in place, the State of Tennessee Department of Safety routinely sent out 200,000–300,000 motor vehicle registration renewal notices, with 20 percent (40,000–60,000) not getting to the intended owner because of incorrect addresses (The Tennessean, January 1998).
- Electronic data audits reveal that invalid data values in the typical customer database averages around 15 to 20 percent. Physical data audits suggest that actual data errors, even though the values may be valid, may be 25 to 30 percent or more in those same databases. The cost of this nonquality data takes its toll on the business’ bottom line in the form of wasted communication costs to its customers. The most significant real cost, however, is lost customer lifetime value as a result of missed or late communication or the aggravation factor. The aggravation factor is the nuisance caused to customers as a result of nonquality information such as incorrect invoices or having to change address information multiple times. Lost or missed customer lifetime value as the result of poor information quality can be significantly greater than the money wasted on duplicate and wrong address mailings.8 Wasted mailout costs of $10,000 may actually result in millions of dollars in lost customer lifetime value.
- A U.S. manufacturing company stock lost 20 percent of its value (dropping 4.5 points to 20) due to a discrepancy in actual inventory and automated inventory reports in December 1995.
- A U.K. engineering company stock lost 13 percent of its value in April 1997 because a data error caused profits to be overstated. Some costs that had been written off as they were incurred continued to be carried in the balance sheet.
- Barbra Streisand pulled her investment account from her investment bank because it misspelled her name as “Barbara.”
- When we wrote an $8,000 check against our home equity loan, the money was paid from someone else’s account because the printer had printed the wrong account number on the checks. The bank branch manager called us personally to inform us about the mistake, told us to destroy those checks, and new checks were printed for us.
- A $29,000 wire transfer due to me in Brentwood, Tennessee, ended up in someone else’s bank account in Seattle, Washington. The $3,000 wire transfer that was supposed to be deposited to that Seattle account ended up in my account. The payer’s reply, “Oops!” The Seattle account owner’s reply, “Wow!” My reply, unprintable.

---

8Customer lifetime value is the net present value of the profit and/or revenue of a typical customer over the life of their relationship with the organization. Chapter 7, "Measuring Nonquality Information Costs," describes how to calculate customer lifetime value.
Information Quality and the Bottom Line

Information quality problems hamper virtually every area of a business, from the mailroom to the executive office. Every hour the business spends hunting for missing data, correcting inaccurate data, working around data problems, scrambling to assemble information across disintegrated databases, resolving data-related customer complaints, and so on, is an hour of cost only, passed on in higher prices to the customer. That hour is not available for value-adding work. Senior executives at one large mail-order company personally spend the equivalent of one full-time employee (senior executive) in reconciling conflicting departmental reports before submitting them to the Chief Executive Officer. This means there is the equivalent of one senior executive’s time is wasted because of redundant and inconsistent (nonquality) data!

Bill Inmon observes that 80 to 90 percent of the human efforts in building a data warehouse are expended handling the interface between operational and data warehouse environments.9

This effort is caused by not having an integrated data environment. This requires data warehouse professionals to have to map undefined and unintegrated data from many disparate and redundant databases and files, standardize, remove redundant occurrences of data both within single files and across redundant files, and integrate and consolidate data and format it into an integrated data warehouse data architecture. Well over half of these costs are attributable directly to nonquality data and nonquality data management and systems development practices.

Even worse, because of the complexity and content, the temptation is great to quickly produce “90-day wonder” data marts, thrown together quickly without addressing the data integration issues. This only exacerbates the already huge problem of nonquality data and increases the costs of solving the right

---

problem later on. For data warehousing projects to be successful, the organization must address the problem of nonintegrated data head on.

The bottom line is that information quality problems hurt the bottom line. Quality experts agree that the costs of nonquality are significant. Quality consultant Philip Crosby, author of *Quality Is Free*, identifies the cost of nonquality to manufacturing as 15 to 20 percent of revenue.11

Joseph M. Juran is one of the world’s pioneering experts in quality. He is the recipient of the “Second Class of the Order of the Sacred Treasure,” the highest decoration presented to a non-Japanese citizen. Juran pegs the costs of poor quality at 20 to 40 percent of sales, including costs of “customer complaints, product liability lawsuits, redoing defective work, [and] products scrapped.”12

A.T. Kearney CEO Fred Steingraber confirms that “we have learned the hard way that the cost of poor quality is extremely high. We have learned that in

---

manufacturing it is 25 to 30 percent of sales dollars and as much as 40 percent in the worst companies. Moreover, the service industry is not immune, as poor quality can amount to an increase of 40 percent of operating costs."\(^{13}\)

But what about the costs of nonquality data? If early data assessments are an indicator, the business costs of nonquality data, including irrecoverable costs, rework of products and services, workarounds, and lost and missed revenue may be as high as 10 to 25 percent of revenue or total budget of an organization. Furthermore, as much as 40 to 50 percent or more of the typical IT budget may actually be spent in “information scrap and rework,” a concept well known in manufacturing. Chapter 7, “Measuring Nonquality Information Costs,” describes in detail how to analyze the costs of information and the costs of poor-quality data.

### POOR INFORMATION QUALITY CAUSES BUSINESS FAILURE

**Oxford Health Plans Inc.:** In 1997, Oxford Health Plans disclosed that computer snafus in trying to convert to a new computer system and resulting inaccurate data caused it to overestimate revenues and underestimate medical costs. Other information quality problems caused overbilling of its customers at the same time. Estimating a third-quarter loss of up to $69.3 million, its stock dropped 62 percent—the actual loss was even greater. The New York State Insurance Department fined Oxford $3 million for violations of insurance laws and regulations and ordered Oxford to pay $500,000 to customers that it had overcharged, according to the Wall Street Journal, December 24, 1997. Oxford is struggling for survival. Its stock price as of October 8, 1998 was around $8—only 9 percent of its all-time high value of around $89. Oxford will lose money in 1998, and the consensus of stock market analysts is that it will lose money in 1999 as well.

**Hudson Foods:** In August 1997, Hudson Foods lost its largest customer, Burger King, due to E. coli bacteria contamination that caused several illnesses. While the plant was one of the most modern, and was clean and generally well run, it had two problematic practices: “poor record-keeping and the mixing of one day’s leftover hamburger into the next day’s production.”\(^{14}\)

The information quality problem of not knowing which batches were mixed caused the largest meat recall in U. S. history: 25 million pounds. Accurate information would have probably limited the size of recall significantly. Without its largest customer, Hudson Foods was not able to be profitable, and not only was that plant subsequently sold to IPB Inc., but the rest of Hudson Foods was acquired by Tyson Foods.

**National Westminster Bank:** The British bank had to dispose of its equities businesses in February 1998, taking a pretax loss of around $1.77 billion (£1.01 billion), according to The Financial Times, February 25, 1998. The failure stemmed out of losses of over $150 million (£90 million) caused by incorrectly pricing its derivatives over a two-year period according to The Times, March 14, 1997.


Why Care about Information Quality?

Because the high costs of low-quality data threatens the enterprise. There is and must be only one purpose for improving information quality: to improve customer and stakeholder satisfaction by increasing the efficiency and effectiveness of the business processes. This in turn increases profits and shareholder value. Information quality is a business issue, and information quality improvement is a business necessity.

For organizations in a competitive environment, information quality is a matter of survival, and then of competitive advantage. For organizations in the public and not-for-profit sectors, information quality is a matter of survival, and then of stewardship of stakeholder (taxpayer or contributor) resources.