1

INTRODUCTION

Without even realizing it, everyone is affected by poor data quality. Some are affected directly in annoying ways, such as receiving two or three identical mailings from the same sales organization in the same week. Some are affected in less direct ways, such as the 20-minute wait on hold for a customer service department. Some are affected more malevolently through deliberate fraud, such as identity theft. But whenever poor data quality, inconsistencies, and errors bloat both companies and government agencies and hamper their ability to provide the best possible service, everyone suffers.

Data quality seems to be a hazy concept, but the lack of data quality severely hampers the ability of organizations to effectively accumulate and manage enterprise-wide knowledge. The goal of this book is to demonstrate that data quality is not an esoteric notion but something that can be quantified, measured, and improved, all with a strict focus on return on investment. Our approach is that knowledge management is a pillar that must stand securely on a pedestal of data quality, and by the end of this book, the reader should be able to build that pedestal.

This book covers these areas.

- Data ownership paradigms
- The definition of data quality
- An economic framework for data quality, including steps in building a return on investment model to justify the costs of a data quality program
- The dimensions of data quality
- Using statistical process control as a tool for measurement

- Data domains and mappings between those domains
- Data quality rules and business rules
- Measurement and current state assessment
- Data quality requirements analysis
- Metadata and policy
- Rules-based processing
- Discovery of metadata and data quality and business rules
- Data cleansing
- Root cause analysis and supplier management
- Data enhancement
- Putting it all into practice

The end of the book summarizes the processes discussed and the steps to building a data quality practice.

Before we dive into the technical components, however, it is worthwhile to spend some time looking at some real-world examples for motivation. In the next section, you will see some examples of "data quality horror stories" — tales of adverse effects of poor data quality.

1.1 DATA QUALITY HORROR STORIES

1.1.1 Bank Deposit?

In November of 1998, it was reported by the Associated Press that a New York man allegedly brought a dead deer into a bank in Stamford, Connecticut, because he was upset with the bank's service. Police say the 70-year-old argued with a teller over a clerical mistake with his checking account. Because he was apparently unhappy with the teller, he went home, got the deer carcass and brought it back to the branch office.

1.1.2 CD Mail Fraud

Here is a news story taken from the Associated Press newswire. The text is printed with permission.

> Newark — For four years a Middlesex County man fooled the computer fraud programs at two music-by-mail clubs, using 1,630 aliases to buy music CDs at rates offered only to first-time buyers.

> David Russo, 33, of Sayerville, NJ, admitted yesterday that he received 22,260 CDs by *making each address — even if it listed the same post office box — different enough to evade fraud-detection computer programs.*
>
> *Among his methods: adding fictitious apartment numbers, unneeded direction abbreviations and extra punctuation marks.* (Emphasis mine)
>
> The scam is believed to be the largest of its kind in the nation, said Assistant U.S. Attorney Scott S. Christie, who prosecuted the case.
>
> The introductory offers typically provided nine free CDs with the purchase of one CD at the regular price, plus shipping and handling. Other CDs then had to be purchased later to fulfill club requirements. Russo paid about $56,000 for CDs, said Paul B. Brickfield, his lawyer, or an average of $2.50 each. He then sold the CDs at flea markets for about $10 each, Brickfield said. Russo pleaded guilty to a single count of mail fraud. He faces about 12 to 18 months in prison and a fine of up to $250,000.

1.1.3 Mars *Orbiter*

The Mars *Climate Orbiter,* a key part of NASA's program to explore the planet Mars, vanished in September 1999 after rockets were fired to bring it into orbit of the planet. It was later discovered by an investigative board that NASA engineers failed to convert English measures of rocket thrusts to newtons, a metric system measuring rocket force, and that was the root cause of the loss of the spacecraft. The orbiter smashed into the planet instead of reaching a safe orbit.

This discrepancy between the two measures, which was relatively small, caused the orbiter to approach Mars at too low an altitude. The result was the loss of a $125 million spacecraft and a significant setback in NASA's ability to explore Mars.

1.1.4 Credit Card Woes

After having been a loyal credit card customer for a number of years, I had mistakenly missed a payment when the bill was lost during the move to our new house. I called the customer service department and explained the omission, and they were happy to remove the service charge, provided that I sent in my payment right away, which I did.

A few months later, I received a letter indicating that "immediate action" was required. Evidently, I had a balance due of $0.00, and because of that, the company had decided to revoke my charging privileges! Not only that, I was being reported to credit agencies as being delinquent.

Needless to say, this was ridiculous, and after some intense conversations with a number of people in the customer service department, they agreed to mark my account as being paid in full. They notified the credit reporting agencies that I was not, and never had been, delinquent on the account (see Figure 1.1).

1.1.5 Open or Closed Account?

Three months after canceling my cellular telephone service, I continue to receive bills from my former service provider indicating that I was being billed for $0.00 — "Do not remit."

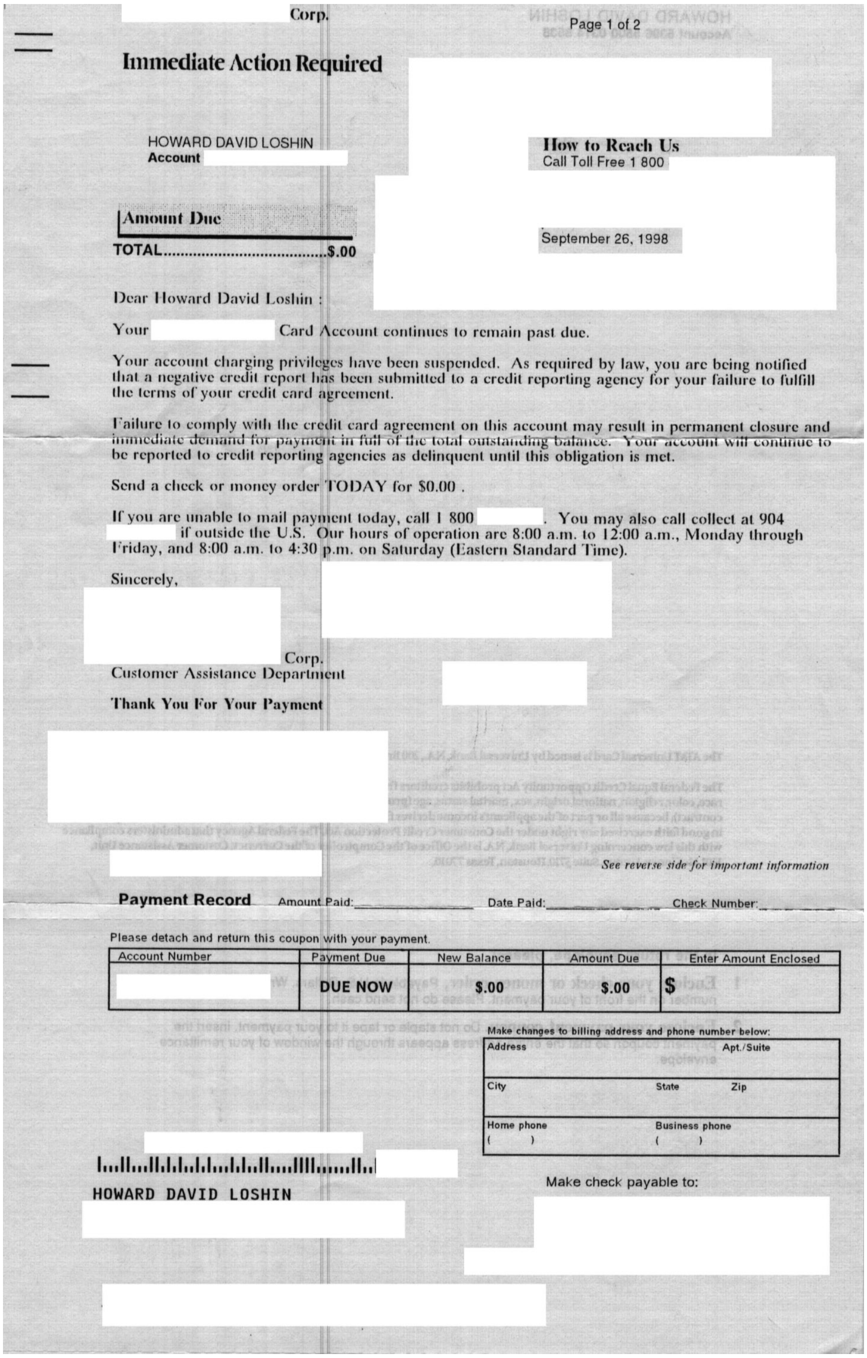
1.1.6 Business Credit Card

A friend of mine is the president of a small home-based business. He received an offer from a major charge card company for a corporate charge card with no annual fee. He accepted, and a short time later, he received his card in the mail. Not long after that, he began to receive the same offer from the same company, but those offers were addressed differently. Evidently, his name had been misspelled on one of his magazine subscriptions, and that version had been submitted to the credit card company as a different individual. Not only that, his wife started to receive offers too.

Six months later, this man still gets four or five mail offers per week in the mail from the same company, which evidently not only cannot figure out who he is but also can't recognize that he is already a customer!

1.1.7 Direct Marketing

One would imagine that if any business might have the issue of data quality on top of its list, it would be the direct marketing industry. Yet, I



[redacted] Corp.

Page 1 of 2

Immediate Action Required

[redacted]

HOWARD DAVID LOSHIN
Account [redacted]

How to Reach Us
Call Toll Free 1 800 [redacted]

September 26, 1998

Amount Due
TOTAL.....................................$.00

Dear Howard David Loshin :

Your [redacted] Card Account continues to remain past due.

Your account charging privileges have been suspended. As required by law, you are being notified that a negative credit report has been submitted to a credit reporting agency for your failure to fulfill the terms of your credit card agreement.

Failure to comply with the credit card agreement on this account may result in permanent closure and immediate demand for payment in full of the total outstanding balance. Your account will continue to be reported to credit reporting agencies as delinquent until this obligation is met.

Send a check or money order TODAY for $0.00 .

If you are unable to mail payment today, call 1 800 [redacted]. You may also call collect at 904 [redacted] if outside the U.S. Our hours of operation are 8:00 a.m. to 12:00 a.m., Monday through Friday, and 8:00 a.m. to 4:30 p.m. on Saturday (Eastern Standard Time).

Sincerely,

[redacted]

[redacted] Corp.
Customer Assistance Department

[redacted]

Thank You For Your Payment

[redacted]

[redacted]

See reverse side for important information

Payment Record Amount Paid:__________ Date Paid:__________ Check Number:________

Please detach and return this coupon with your payment.

Account Number	Payment Due	New Balance	Amount Due	Enter Amount Enclosed
[redacted]	**DUE NOW**	**$.00**	**$.00**	**$**

Make changes to billing address and phone number below:

Address	Apt./Suite
City	State Zip
Home phone ( )	Business phone ( )

[redacted]

[redacted]

HOWARD DAVID LOSHIN

[redacted]

Make check payable to:

[redacted]

[redacted]

Figure 1.1 Mysterious bill

recently received two identical pieces of mail the same day from the local chapter of an association for the direct marketing industry. One was addressed this way.

David Loshin
123 Main Street
Anytown, NY 11787

Dear David, . . .

The other was addressed like this

Loshin David
123 Main Street
Anytown, NY 11787

Dear Loshin, . . .

1.1.8 Tracking Backward

I recently ordered some computer equipment, and I was given a tracking number to follow the package's progress from the source to my house. If you look at the example in Table 1.1 (which has been slightly modified from the original), you will see that the package was scanned at the exit hub location in a specific state on June 26, was (evidently) scanned in Nassau county, NY, at 12:30 A.M. the following day but was scanned as a departure from the airport in the same state as the exit hub at 1:43 P.M., which is practically 11 hours later. The rest of the tracking makes sense — from the XX airport to an airport local to my area, then onto my locality, and finally to the delivery point.

Obviously, the June 27, 12:30 A.M. scan in Nassau has either the incorrect location or the incorrect time. It is most likely the incorrect time, since packages are scanned on entry to a location and on exit, and this scan appears between the location scan at EXIT HUB and the departure scan at ANYTOWN INTL, same state.

1.1.9 Conclusions?

These are just a few stories culled from personal experience, interactions with colleagues, or reading the newspaper. Yet, who has not been subject to some kind of annoyance that can be traced to a data quality problem?

TABLE 1
Tracking history for the equipment I ordered.

PACKAGE PROGRESS			
Date	Time	Location	Activity
June 28, 2000	5:25 P.M.	NASSAU-HICKSVILLE, NY US	DELIVERED
	3:42 A.M.	NASSAU, NY	DESTINATION
	3:31 A.M.	NASSAU, NY US	LOCATION SCAN
June 27, 2000	11:21 P.M.	NEWARK INTL, NJ	DEPARTURE SCAN
	4:45 P.M.	NEWARK INTL, NJ	ARRIVAL
	1:43 P.M.	ANYTOWN INTL, XX	DEPARTURE SCAN
	12:30 A.M.	NASSAU, NY US	ARRIVAL
June 26, 2000	11:29 A.M.	EXIT HUB, XX US	LOCATION SCAN
June 23, 2000	9:11 P.M.	ADDISON, IL US	LOCATION SCAN
	1:38 P.M.		SHIPMENT DATA RECEIVED

1.2 KNOWLEDGE MANAGEMENT AND DATA QUALITY

Over the past 30 years, advances in data collection and database technology have led to massive legacy databases controlled by legacy software. The implicit programming paradigm encompasses both business policies and data validation policies as application code. Yet, most legacy applications are maintained by second- and third-generation engineers, and it is rare to find any staff members with firsthand experience in either the design or implementation of the original system. As a result, organizations maintain significant ongoing investments in daily operations and maintenance of the information processing plant, while mostly ignoring the tremendous potential of the intellectual capital that is captured within the data assets.

1.2.1 What Is Knowledge Management?

An organization's data collection is a valuable business resource that until now has been largely underutilized. In the past, when data were mostly locked up in databases, ferociously guarded by organizational overlords, the ability to share and benefit from enterprise knowledge was limited. Today, as technology evolves to unlock and distribute these databases, a procedural methodology has evolved in tandem to help integrate the technical, organizational, and behavioral issues associated with enterprise knowledge. This methodology is referred to as "knowledge management."

According to Karl Erik Sveiby, knowledge management is "The art of creating value by leveraging the intangible assets." The Gartner Group states it in a more down-to-earth way: "Knowledge management is a discipline that promotes an integrated approach to identifying, managing, and sharing all of an enterprise's information assets. These information assets may include databases, documents, policies and procedures as well as previously unarticulated expertise and experience resident in individual workers."[1]

In other words, knowledge management is a strategic process meant to capture the ways that an organization integrates its information assets with the processes and policies that govern the manipulation of those intellectual assets. The desired goal of knowledge management is the determination and the harnessing of the value of the information resources within the enterprise.

While knowledge management encompasses many disciplines, such as document management or e-mail technology, our goal is to focus on the embedded knowledge in data sets that can be expressed as a set of business rules. Having expressed these rules, we can then validate our expectations of the information that we use by testing it against the business rules.

1.2.2 What Are Business Rules?

A business rule is an assertion about the state of a business process. All business processes and data sets have rules. Unfortunately, these rules are frequently expressed as "lore" instead of being properly documented, passed from one "generation" of managers or technicians to another by word of mouth. Even worse, sometimes these rules are forgotten over time, having been implemented in software and then left to happily chug away until a crisis strikes.

[1] http://cestec1.mty.itesm.mx~laava/sdsites/cursos/pqg_base/definicion1.htm

Business rules are likely to be expressed in a natural language. An example of a business rule for an employee database might be "No employee makes more than five times the salary of the lowest-paid employee." An example for an electronic data interchange system might be "If the header of an incoming message includes the symbol #, then the message is routed to the accounts payable office." A business rule might govern the way operations proceed. An example for a sales office might be "No customer is pitched for a new sales transaction within 10 days after the customer's last sales transaction."

When business rules are undocumented, the chances are high that the meanings or implications of these rules will be lost within a short period of time. Knowledge is lost when employees leave the company or change positions internally, when managers have too much control over information, and when there is entropy in communication protocols. When business rules are lost, the opportunity to take advantage of the information resources is squandered as well.

1.2.3 Why Data Quality Is the Pivot Point for Knowledge Management

The opportunity to take advantage of the data and information resource can only be enabled if there is an understanding of the structure and knowledge about the collections of information. The critical point is that a formal method is needed for collecting, documenting, and validating business rules. This methodology revolves around ensuring that the information that is present in the system meets or beats the expectations of what is in the system. Ensuring data quality is a process of stating information requirements followed by a process of validating that those requirements are being met.

A significant effort is made today to bring data out of the transactional framework and into an operational data store that can be used for analytical processing. These operational data stores are embodied in data marts and data warehouses, which are useful knowledge management tools when used effectively. A major component of the data warehouse process is the extraction and transformation of data from source and legacy data systems into the target data warehouse. This extraction process is an inflection point at which business rules can both be discovered and used for ensuring enterprise-wide data quality.

Based on the growing awareness of data quality as an enterprise responsibility, there is a burgeoning need for inline data quality policy management. According to the Gartner Group, "It is critical for enterprises to develop a data quality program and ensure that it is carried out. Key to this effort is identifying data stewards[2] in end-user areas where data ownership is clearly defined. . . . Enterprises can minimize these data inconsistencies by better understanding the parameters governing the meaning and movement of data."[3]

The use of data quality management as a tool for knowledge management along with these definitions are the initial motivators for this book. But we won't stop at that! The next section gives nine reasons for caring about data quality.

1.3 REASONS FOR CARING ABOUT DATA QUALITY

The data quality problem is pervasive in organizations across all industries. Bad data cost money and reduce productivity — time spent diagnosing and fixing erroneous data is time not spent productively. Low data quality eventually leads to reduced customer satisfaction. For example, customers exposed to incorrect reports or statements are less likely to trust the organization providing those reports. Finally, strategic decisions based on untrustworthy information are likely to result in poor decisions.

1.3.1 Low Data Quality Leads to Operational Inefficiency

The use of a manufacturing chain assumes that multiple stages are associated with the final product, and at each stage there is an expectation that the partially completed product meets some set of standards. Information processing is also a manufacturing chain — pieces of information flow into and out of processing stages where some set of operations are performed using the data.

To continue this analogy, when a product developed on a manufacturing chain does not fit the standards required at a specific stage, either the product must be thrown away or fixed before it can continue down the manufacturing line. Information is the same way: When a data

[2] Explored later in this book.

[3] J. Hill, and S. Laufer, *Data Transformation: Key to Information Sharing,* Gartner Group Strategic Analysis Report, September 29, 1998.

record is found to be incorrect, the record needs to be deleted or fixed before the processing can continue. Sometimes this "break" means the delay of the entire processing stream, although it is more likely that the records will be shunted aside and the stream continued on the next set of records, with the erroneous records being dealt with at a later time.

As the level of data quality decreases, the more frequent the breaks in operation. As more employees are allocated to fixing and reconciling incorrect records, the rate at which information is processed decreases — in other words, operational inefficiency. This inefficiency is manifested as error detection, error correction, and rework. We will look at these issues more closely in Chapter 4.

1.3.2 Low-Quality Data Constrains Decision Making

Information production can be used for either operational processing or analytical processing. The same information can be used for both purposes. Yet, if the data are used for analytical processing or decision support, the quality of the data can affect the analysis. If senior managers rely on the results of the analysis, they may rely on conclusions drawn from faulty assumptions. If these same managers are aware of the low quality of the input data, they may choose to delay making decisions until better information can be collected or until the same information can be improved.

The same factors that cause delays in decision making can also produce constraints. When one unit of an organization depends on the results of analytical processing or decision support analysis from another unit, the level of quality of the source information will affect the first unit's ability to take action on the analysis and may cause an erosion of trust between the two units.

For example, an airline may use an online reservation system as an operational process as well as the basis for analytical processing. The information may be targeted for use in determining what kinds of frequent flyer promotions to create and to which sets of customers to make these offers. If the data is of low quality, the conclusions presented by an analysis may not be trustworthy, and managers making any decisions about the new promotions may be hesitant if they cannot rely on those conclusions.

1.3.3 Good Data Enhances Data Warehouse Utility

According to an article presented by the Data Warehousing Institute,[4] a survey conducted by survey.com ("Database Solutions III") estimates that the worldwide data warehousing market is growing at the rate of 43% a year, and will reach $143 billion by 2003, with sharp growth outside of North America. Additionally, it is commonly recognized (and attributed to Data Warehousing pioneer Bill Inmon) that 80% of the effort of a data warehousing project is spent in extracting, cleansing, and loading data.

Considering this significant investment as well as the fact that data warehouses are used for analytical processing, anything that improves the customer's ability to analyze data increases the value of the data warehouse. A major component of the data warehouse solution process is the extraction and transformation of data from a legacy source before the warehouse is populated. If the information in the warehouse is of poor quality, a significant amount of time is spent in tracking and removing errors. And since many warehouses are populated on a short period basis (such as completely reloaded daily, with hourly sweeps), if the error correction time frame exceeds the refresh period, the warehouse would be nothing but a white elephant.

This stage of data population is the best opportunity to include data quality validation and standardization, since good data in the warehouse enables its use. We will focus on the data warehouse certification process that uses a data quality rules engine in Chapters 7, 8, and 12.

1.3.4 Bad Data Leads to Incorrect Conclusions

Just as a house built on a weak foundation cannot stand, conclusions based on incorrect input will not withstand scrutiny. As more and more data warehouses are being built for critical business analytical applications, the criticality of ensuring data quality increases. Analytical results based on bad data will be bad results — period.

Not only that, operational decisions that rely on poor data can cause inefficiencies in application systems. For example, load-balancing accesses to a database based on a distribution of data in one attribute can lead to skewed balancing if half of the records referenced have an empty index field!

[4] http://www.dw-institute.com/whatworks9/Resources/warehousing/warehousing.html

1.3.5 Bad Data Lead to Customer Attrition

Have you ever been billed for service that you have not received? Or have you been threatened with being reported to a credit bureau for being late on a payment on a balance due of $0.00? Many people have some nightmare experience with which they can relate, always associated with some incorrect information that causes pain in the pocketbook. These stories always seem to end with the customer ending his or her relationship with the vendor or product provider over the matter.

These errors are typically due to some mistake on behalf of the service or product provider, whether it is in customer records, customer billing, product pricing, or during data processing. No matter what, the problem is worsened by the fact that it is apparent that the organization at fault has no evident means of proactive error detection in place. This conclusion may be drawn because it is the customer who is doing the error detection. We can claim that while a significant expense is made to acquire new customers, it is worthwhile to invest the time and money into improving the data collected on the current customers, since customer attrition may be tied directly to poor data quality.

1.3.6 Good Data Enhances New Customer Acquisition

Just as poor data foster mistrust among current customers, it also can cast doubt in the minds of potential customers. When a potential customer is presented with an offer backed by high-quality data, the image of the seller is enhanced, which can improve the opportunity to turn a potential customer into a real customer.

As an example, consider this real-life pitch we recently received in the mail. My wife and I recently took a trip with our 5-month-old baby. We purchased a seat for our child at a discount rate because of her young age. About a month after our trip, our baby received a letter from the airline and a major long-distance carrier, offering her 10,000 frequent flyer miles if she switched her long-distance service. Clearly, this cooperative sales pitch, cobbled together between the airline and the long-distance carrier (LDC), may be effective some of the time, but consider this: The airline knew that we had bought a discount seat for our baby, and typically babies don't have authority to make purchasing decisions in a household. In addition, both my wife and I have received the same pitch from the same airline-LDC combination — on the same day! Because we saw that the long-distance carrier was unable to keep their records straight about who we were, we weren't sure that we could trust them to handle our telephone records correctly either.

1.3.7 Poor Data Quality Leads to Breakdown in Organizational Confidence

On behalf of both customers and employees, when an organization displays an inability to manage simple information management issues, it sows a seed of doubt that the organization can manage critical processes. When customers lose faith in a provider's ability to provide, it leads to customer attrition. When employees no longer believe in the company's ability to do business, it leads to employee turnover and loss of strategic business knowledge.

1.3.8 Bad Data Restricts System and Data Migration Projects

Having been involved in some legacy migration projects, I can say from direct experience that the most frustrating component of a migration project is the inability to accumulate the right information about the data and systems that are being migrated. Usually, this is due to the tendency of implementers to program first, document later, if at all. But as systems age, they are modified, broken, fixed, or improved but without any updates to documentation.

This situation forces the integrators to become information archaeologists to discover what is going on within the system. Naturally, undirected discovery processes will increase costs and delay the actual implementation, and the amount of time needed to determine what is going on cannot be predicted ahead of time.

1.3.9 Good Data Increases ROI on IT Investment

When we have assessed the data quality requirements of our system and put in place the right kinds of processes to validate information as it passes through the system, we can limit the downtime and failed processes based on low data quality. In turn, this means that without having to diagnose and fix data quality problems, processing can proceed with a greater bandwidth, which subsequently allows for an increase in processing volume without an increase in resources. In a company whose business depends on increasing processing volume without increasing overhead (such as securities processing or order fulfillment), this can increase the return on information technology investment.

1.4 KNOWLEDGE MANAGEMENT AND BUSINESS RULES

Business operations are defined using a set of rules that are applied in everyday execution. When the business depends on the correct flow of information, there is an aspect of data quality that intersects the operational specification.

In essence, in an information business, **business rules are data quality rules.** This implies that data quality is an integral part of any operational specification, and organizations that recognize this from the start can streamline operations by applying data quality techniques to information while it is being processed or communicated. This in turn will prevent bad data from affecting the flow of business, and denying the entry of incorrect information into the system eliminates the need to detect and correct bad data. Because of this, a "data quality aware" operation can execute at lower cost and higher margin than the traditional company.

1.4.1 Competition in Knowledge-Intensive Industries

Businesses that traditionally rely on the leverage of timely information (retail securities sales, market data providers, analytical market analysis) suddenly have competition from players whose barrier to entry is low. For example, in the retail securities investment business, the combination of low-cost "electronic" securities trading combined with the widespread availability of financial and investment recommendations demonstrates that margins on securities transactions can be lowered, but businesses can still profit.

As the availability of online analysis increases, there will be a decrease in the need for a large sales staff as a distribution channel for standard financial products. In the future, the perceived value of the information provided by an investment analyst will not necessarily be based on "brand name," but instead, the business rules will focus on presentation, accuracy and timeliness, which are critical data quality dimensions.

1.4.2 Micromarketing

Businesses that traditionally rely on marketing analysis will find that "micromarketing" analyses will provide opportunities for micromarket penetration. Businesses that have not traditionally relied on marketing analysis are beginning to use information gleaned from online transactions. The rules for accumulating and merging these data sets are managed as business rules. The capability provided by desktop access to data sources (such as consumer databases) allows more information to enhance the analysis. This strategic analysis can only be trusted if the data on which it is based are of high quality.

1.4.3 Cooperative Marketing Arrangements

With increasing frequency, organizations from different industries join forces in cooperative or cross-linked marketing programs. A common example is an offer of frequent flyer miles coupled with a change in long-distance service. Because of the potential disparity between different organizations' data models and the currency of their stored data, these arrangements provide strong opportunities for data quality improvements. The rules for value-added information merging, which incorporates both known data and inferred information, can be encapsulated as business/data quality rules.

1.4.4 Deregulation in the Communications, Energy, and Financial Services Industries

The recent consolidation of the insurance and brokerage industries in anticipation of deregulation and the (expected) waves of deregulation in the communications and the energy industries pave the way for consolidation of multiple business entities. When two (or more) independent organizations merge their operations, their combined data resources need to be combined as well. A streamlined data merging, migration, and quality assurance methodology will uncover synergistic opportunities when combining data sets. Leverage in affecting the organizational bottom line can be obtained through the qualified merging of data resources by decreasing operational costs (for example, error detection and correction) and by increasing customer response and customer satisfaction.

1.4.5 The Internet as a Knowledge Transport Medium

The rapid ascension of the Internet as a business medium has created opportunities for industries based on information and knowledge. When the business relies on the use of information, the quality of that information suddenly takes on a much higher visibility. This is an entirely different dimension of applicability for the data quality business, since the growing boom of "e-commerce" cuts across all major industries.

The World Wide Web can be can be seen as the largest uncoordinated database service in the world. All Web sites can be characterized in terms of some database service. Some sites act primarily as data presentation systems, others act as data collection systems, and some act as database query systems. Success in the arena of electronic commerce will be largely dependent on customer-focused micromarketing and automated customer relationship management and sales force automation. We will provide guidelines for both the acquisition of valid data over the Internet, as well as guidelines for accurate, customer-focused data presentation.

1.5 STRUCTURE OF THIS BOOK

1.5.1 Chapter 1: Introduction

This chapter is an introduction to the ideas behind enterprise knowledge management and the importance of data quality in this endeavor. In this chapter, we enumerate the most important issues regarding information quality, how business is aversely affected by poor data quality, and how business can be improved when data quality is high.

1.5.2 Chapter 2: Who Owns Information?

Because the data quality problem ultimately belongs to the data consumer, it a good idea to start out by establishing ownership and boundaries. Chapter 2 focuses on the issues of data ownership. The chapter begins by discussing the data processing activity as a manufacture of information. The final product of this factory is knowledge that is owned by the data consumers in a business enterprise.

Who are the data producers and data consumers in an enterprise? We look at internal data producers (internal processes like account opening, billing, marketing) and external data producers ("lead lists," consumer research, corporate structure data). We also look at the enterprise data consumers, ranging from the operational (customer service, billing, resource planning), the tactical (middle management, scheduling), and strategic consumers (directional management, strategists).

There are complicating notions with respect to data ownership. The means of dissemination, collecting data from the public domain, as well as acquiring data from data providers confuse the issues. Therefore, a set of ownership paradigms are introduced, including decision makers, sellers, manipulators, guardians, and workers. These paradigms bound the "social turf" surrounding data ownership. Finally, Chapter 2 focuses on a finer granularity of ownership issues, including metadata ownership, governance of storage and repositories, and accountability for data policies.

1.5.3 Chapter 3: Data Quality in Practice

In Chapter 3, the notion of data quality is introduced, loosely defined through the term "fitness for use." Background in traditional quality systems is provided. Chapter 3 also provides some insight into information theory as it can be applied to data quality. Finally, statistical process control is discussed as critical to improving data quality.

1.5.4 Chapter 4: Economic Framework for Data Quality and the Value Proposition

The rationale for designing and building data quality systems seems logical, but an economic framework that can be used to measure the effects of poor data quality while highlighting its benefits is needed to demonstrate the value of improving poor data quality. Chapter 4 reviews the knowledge hierarchy and the knowledge manufacturing process, then defines a set of impact domains that can be used to model the costs and missed opportunities that result from bad data. The chapter focuses on strategic, tactical, and operational impacts, followed by a description of a way to model these impacts.

1.5.5 Chapter 5: Dimensions of Data Quality

In Chapter 5, a number of dimensions of data quality are discussed. These dimensions are grouped into quality of the data model, quality of data values, quality of data presentation, and other data quality dimensions.

1.5.6 Chapter 6: Statistical Process Control and the Improvement Cycle

In Chapter 6, we look in detail at the notion of control and how that idea can promote predictability of the quality of the product being produced. Statistical process control is a method for measuring the quality of the product while its process is in operation, instead of relying on measuring the product after it has been completed. SPC is a process of gathering measurements during processing in order to identify special causes of inconsistencies and variations that lead to a flawed product. We will focus in particular on the notion of control charts and how these charts are used to look for processes that are in control and processes that are out of control. In addition, we will look at how control charts are used for setting up an improvement cycle with respect to conformance with data quality specifications.

1.5.7 Chapter 7: Domains, Mappings, and Enterprise Reference Data

In this chapter, we begin to explore the ideas revolving around data types and how data types are related to the notion of sets. We then describe our definition of data domains, both descriptive and enumerated. Next, we discuss the relations between domains, how those relations exist in databases, and the power of abstracting these mappings as reference metadata. Finally, we propose a publish/subscribe model for the management and use of enterprise reference data.

1.5.8 Chapter 8: Data Quality Assertions and Business Rules

In this chapter, we really begin the investigation of the kinds of rules that can be applied to the measurement of data quality. With respect to any tabular data (such as database tables), we will look at these classes of rules: (1) data value rules, (2) attribute rules (rules for data values associated with a particular attribute), (3) domain rules, (4) tuple rules (rules associating different attributes within a tuple), and (5) table rules. We identify a number of rules that can be characterized by a measurable method.

1.5.9 Chapter 9: Measurement and Current State Assessment

We look at measuring conformance to the rules described in Chapter 8 in two different ways. The first way is a static measurement, which involves looking at the data after it has been delivered into its target location. The second way is in-process (dynamic) measurement, which is consistent with the notions of statistical process control as described in Chapter 6. This way involves integrating the measurement of the conformance to the data quality rules while the data is in transit.

The goal of the initial measurement is to gauge the degree of any data quality problem as well as use the economic framework to measure the effects of any poor data quality. The current state assessment is a report that focuses on the important dimensions of data quality, the degree on conformance, and how the lack of conformance affects the bottom line. The result of the CSA is a characterization of the state of data quality and what needs to be done to improve it.

1.5.10 Chapter 10: Data Quality Requirements

How do we determine data quality requirements? We use a technique borrowed from object-oriented design called use-case analysis, which is used to specify the system in terms of actors, use-cases, and triggers.

Actors represent the roles that users play, use-cases represent what the actors do with the system, and triggers represent events that initiate use cases. We then select from the list of data quality dimensions from Chapter 5 those dimensions that are of greatest importance to the actors and define data quality rules, as described in Chapter 8. We can choose thresholds for conformance to the data quality rules as a baseline for acceptance of the data. These baseline thresholds are defined so that when met, the data set consumers can be confident of the levels of data quality.

1.5.11 Chapter 11: Metadata, Guidelines, and Policy

When defining data quality rules and requirements, it is necessary to build a central repository for the collection of all metadata, which is information about the data. All data processing systems maintain some form of metadata, whether it is explicit using a DBMS system's data dictionary, or implicit, such as the methods and inheritance embedded in a C++ class definition.

The traditional view of metadata includes mostly static information about the data, such as data types, sizes, and perhaps some rudimentary rules such as enforcing a primary key or indexing. Metadata, though, could and should encompass much more, including data quality and business intelligence rules that govern the operational manufacture and use of data. This is not to say that this kind of metadata isn't being maintained now — because a lot of it is! This metadata, however, is maintained in a format unsuitable for easy access and understanding by the actual data consumer.

1.5.12 Chapter 12: Rules-Based Data Quality

Having defined an assertional framework in Chapter 8, how would data integrity assertions be defined and validated? The answer is through the use of rule-based systems that a user employs to both define and test rules. In Chapter 12, we describe rule-based systems: definitions of "rule" and rule-based system and how we incorporate our assertion system from Chapter 8 into an information integrity language. Last, we describe how a rules engine can use these validation specifications to provide a test framework for both data integrity validation and the data for statistical process control, all within the context of the data consumer's constraints.

1.5.13 Chapter 13: Metadata and Rule Discovery

Chapter 12 presents a rule system for defining and testing information validity. In Chapter 13 we discuss what we learn from the rule definition and rule-testing process. Included is domain discovery, association rules, map discovery, functional dependencies, overlapping and intersecting domains, and attribute splitting. These all form a link in a feedback chain in the metadata analysis and data quality improvement process.

1.5.14 Chapter 14: Data Cleansing

Understanding data quality rules, assertions, and validity checks is one thing; actually "fixing" the problems is another! Chapter 14 focuses on techniques for cleansing data to the point where it meets or exceeds the conformance thresholds and is therefore acceptable. We start with metadata cleansing — identifying and documenting what we know about our data. We look at using our assertions to form information filters that prevent bad data from entering the system.

When identifiability constraints specify uniqueness, it is worthwhile to eliminate duplicate entries. This takes on two faces: absolute duplicates and near-match duplicates. To address the latter, we discuss approximate searching and matching techniques. These techniques are also useful in householding, a special kind of duplicate elimination process that looks to accumulate information about individuals sharing a single residence.

When there is a standard form for an information set (as specified by a structured domain or such as the United States Postal Service standard), we use a technique called standardization to make our data to look the way we want it to. We will discuss issues surrounding the opportunity for cleansing data during a data migration process.

1.5.15 Chapter 15: Root Cause Analysis and Supplier Management

One technique not yet discussed is using the results of the information validity exercise to prevent the continuation of low data quality events.

When using the rule-based system for validity checking, we can use the information in the reports to look for root causes of the occurrences of bad data quality. This technique is the last link in our improvement chain, since fixing the sources of bad data will directly improve overall data quality.

1.5.16 Chapter 16: Data Enrichment and Enhancement

Data enrichment is a process of enhancing the value of a pool of information by combining data from multiple sources. The goal of enrichment is to provide a platform deriving more knowledge from collections of data. A simple enrichment example is the combination of customer sales data with customer credit data to build a data mart that can be used to make special offers to preferred customers (customers with high purchasing profiles and good credit profiles). Other examples would be merging health insurance claim information with professional billing information to search for insurance fraud and the enrichment of financial electronic data interchange messages to enable straight-through processing.

In any case, data quality is critical to successful enrichment. In Chapter 16, we discuss how data standardization, clustering techniques, and the use of data quality rules can be used to express enrichment directives. We also look at some enrichment examples and how they can be affected by poor data quality.

1.5.17 Chapter 17: Data Quality and Business Rules in Practice

In Chapter 17, we review our data quality rules in a context that demonstrates ways to actually make use of these rules. Specifically, we talk about the way these rules partition the data into two sets, the set of conforming records and the set of nonconforming records. There is already a means for specifying set partitions in databases — the query language SQL, which is not just used for performing operations on databases — but to show how to turn the rules into executable queries. The rest of this chapter is designed to show how the implementation of a data quality rules process can add significant value to different operating environments.

1.5.18 Chapter 18: Building the Data Quality Practice

Our final chapter is basically a “run book” for building a data quality practice. We walk step by step through the issues in building the practice, starting with problem recognition, followed by the way to gain senior-level support, adopting a methodology, and creating an education program as well as actual implementation.