



Improving Data Warehouse and Business Information Quality

Methods for Reducing Costs
and Increasing Profits

Larry P. English

Wiley Computer Publishing



John Wiley & Sons, Inc.

NEW YORK • CHICHESTER • WEINHEIM • BRISBANE • SINGAPORE • TORONTO

**Early Reviews for Larry P. English's
*Improving Data Warehouse and Business
Information Quality***

“The Information Quality Bible for the Information Age!

“Practical and useful. . . this book has it all in one package:
‘concept book, textbook, reference book, practitioner’s guide.’

“English’s sense of humor is reflected throughout. The rewards
from the implementation of his methods should be as
enjoyable as the reading”

- *Masaaki Imai*
Founder, Kaizen Institute
- *Bud H. Cox*
Managing Director, Kaizen Institute of Japan
(*Kaizen* is a Japanese word that connotes
“continuous improvement involving everyone”)

“This book is a must for every business bookshelf. Larry English
has been on the forefront of the Data Quality issue from the
outset. . . [and] has some real wisdom on this vital issue.”

- *John Zachman*
Zachman International,
*Creator of the Framework for Enterprise
Architecture*

“This book is long overdue. As a leading expert on Quality in the world today, Larry English shows the impact that data and information quality directly have on costs and on profitability—not just for data warehouses but also for business information. His examples are clear, and vital for management to read.

“This book will maximize your chances for success. No Data Warehousing project and no IT Department should be without it. I predict that it will become the ‘Bible’ of Quality success.”

— *Clive Finkelstein*
Information Engineering Services Pty Ltd,
“Father of Information Engineering”

“Everywhere we go. . . we see the results of data quality problems. In this book, Larry English not only turns up the heat by discussing the sources and nature of data quality problems, he also sheds real light through a practical approach to addressing data quality improvement. Time spent understanding and applying the principles and tips Larry offers will be well worth the investment.”

— *Vaughan Merlyn*
The Concours Group

“Very lively reading. The book belongs on the bookshelf of every manager and technician.”

— *Bill Inmon*
Pine Cone Systems,
“Father of Data Warehousing”

Publisher: Robert Ipsen
Editor: Robert M. Elliott
Assistant Editor: Pam Sobotka
Managing Editor: Marnie Wielage
Text Design & Composition: Benchmark Productions, Inc.

Designations used by companies to distinguish their products are often claimed as trademarks. In all instances where John Wiley & Sons, Inc., is aware of a claim, the product names appear in initial capital or ALL CAPITAL LETTERS. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

TQdM® is a registered trademark of INFORMATION IMPACT International, Inc.

This book is printed on acid-free paper. ∞
Copyright © 1999 by Larry P. English. All rights reserved.
Published by John Wiley & Sons, Inc.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4744. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0012, (212) 850-6011, fax (212) 850-6008, E-Mail: PERMREQ@WILEY.COM.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold with the understanding that the publisher is not engaged in professional services. If professional advice or other expert assistance is required, the services of a competent professional person should be sought.

Library of Congress Cataloging-in-Publication Data:

English, Larry, 1947-

Improving data warehouse and business information quality :
methods for reducing costs and increasing profits / Larry English.

p. cm.

Includes bibliographical references and index.

ISBN 0-471-25383-9 (pbk. : alk. paper)

1. Industrial management--Data processing. 2. Business--Data
processing. 3. Data warehousing. I. Title.

HD30.2.E54 1999

658.4'038'0285574--dc21

98-33167

CIP

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1



Introduction

“The best effect of any book is that it excites the reader to self activity.”

—THOMAS CARLYLE

The state of information quality today is worse than it was five years ago, and it is getting worse day by day. In fact, the quality of information in many organizations is enterprise-threatening. Consider some of my recent experiences:

Having just returned from chairing my fourth Data Quality conference, the first in London in November 1998, my renewed excitement about information quality improvement as a trend was brought back to reality. The next week I keynoted conferences in Phoenix, Arizona, and Orlando, Florida. When I checked into the Sheraton Mesa Hotel in Phoenix, the registration clerk asked if I was checking in. I replied, “Yes, my name is English.” She entered that into the computer, and asked if my first name was Ron. “No,” I replied, “it is Larry.” “I’m sorry, but we have no reservation for a Larry English,” was her reply. “But that is okay, we have rooms.” My confirmation number validated that the reservation was indeed for me, but under the name “Ron.” The clerk replied, “that’s no big deal, we can change it.” “It *is* a big deal,” I contended. “No it isn’t,” she insisted, probably thinking about the ease of making the change in the database, but not about the customer service aspect of the event. Seeing the history of the record, she asked if the last time I had stayed there was the 22nd of last month. No, I had not. When she gave me the printed copy of the registration form to sign, I discovered Arizona State University had just been “relocated” to my address in Brentwood, Tennessee!

I flew from there to Orlando, where registration went smoothly. I had a letter and fax waiting for me at the mail desk at the Omni Rosen hotel. I was pleased with the information that allowed me to pick them up then, rather than having to return after getting to my room and finding I had to go back to the mail desk. After I got to the room, I opened the fax, noting that only 8 of the 10 sent pages were stapled together and sealed in an envelope to

protect its privacy. The sealing was good; the quality control was not. There is an important reason why you document page counts on faxes: to assure all pages go through. Customer service is more than taking the pages and stuffing them in an envelope. Customer service is what The Hyatt in Phoenix did a month later when they received only 6 pages of an 11 page fax. They called my office, allowing them to re-send the missing pages immediately. This is information quality, comparing the information to reality and taking action to assure information customers have what they need when they are supposed to without having to chase it down. In Orlando, I had checked in after my office had closed, causing me to have to wait for the last 2 (most important) pages until the next day.

This book is not about information quality from an esoteric or theoretical standpoint. It is a practical book about using information quality as a business management tool for reducing the business and systems costs resulting from poor information quality. More importantly, this book is about increasing business profits and business effectiveness as a result of having higher-quality information and the customer satisfaction it generates.

The world is now experiencing a phenomenon that future historians may classify as The Golden Age of Information. The Oxford Dictionary defines *golden age* as a period during which commerce, the arts, and so forth, flourish. “Flourish” means that something is successful, very active, or widespread. With this litmus test, we are indeed living in the Golden Age of Information. Data warehouses that now contain multiple terabytes of data have forever altered the information processing landscape. The explosion of information on the Internet over the past few years confirms the importance and value of accessible information.¹

With this proliferation of information, the challenge of managing data and providing quality information has never been more important or more complex.

The premise of this book is that:

- Information and data are strategic enterprise resources.
- Quality information enables competitive advantage and business effectiveness.
- Information quality is not an isolated “function”; it is an inherent and integral part of business management.
- Everyone in the organization has a stewardship role for information quality.
- Without information the enterprise will fail, even process-driven organizations. For example, Coca-Cola has its formulas (information) required to produce its products. For the manufacturing firm Optical Fibres, “the

¹L. P. English, “The Golden Age of Information,” *DM Review*, January 1997, p. 20.

key is in the chemicals and the formula for turning sand (silicon di-oxide) into optical fiber capable of transmitting laser signals for thousands of miles without losing information.”

- Without *quality* information the enterprise will be suboptimized. In fact some will—and already have—fail.
- The costs of poor-quality information are high. Poor-quality information causes process failure and information scrap and rework that wastes people, money, materials, and facilities resources. The most significant problems caused by poor-quality information, however, is that it frustrates the most important resources of the enterprise—its people resources—keeping them from effectively performing their jobs, and it alienates its customer resources through wrong information about them and to them. Because there is a direct correlation between customer complaints and customer defection, the real cost of poor-quality information is in lost customer lifetime value, profits, and shareholder value.
- Information quality is free. When people ask, “what is the business case for making an investment in information quality improvement?” the answer is, “what is the business case for all the information scrap and rework caused by *not* having quality information?” It is the poor-quality information that costs money. The investment in improving both information product and information process quality is recouped multiple times in decreased costs and increased value of information to accomplish strategic business objectives.

QUALITY IS FREE

“Quality is free. It’s not a gift, but it is free. What costs money are the unquality things—all the actions that involve not doing jobs right the first time.

“Quality is not only free, it is an honest-to-everything profit maker. Every penny you don’t spend on doing things wrong, over, or instead, becomes half a penny right on the bottom line. If you concentrate on making quality certain you can probably increase your profit by an amount equal to 5 to 10 percent of your sales. That is a lot of money for free.”²

Who Should Read This Book

This book is not for everyone. It is for people who care about their customers and their information customers. This book is for people who do not like to see people and money resources wasted on information scrap and rework when they could be doing things that add value. This book is for people who seriously

²Philip B. Crosby, *Quality Is Free*, New York: Penguin Group, 1979, p. 1.

want to see shareholder value increase on a long-term basis, not merely from quarterly statement to quarterly statement.

If you do not want to rock the boat, make waves, or change your own behavior, please do not buy this book. This book is for people who are discontent with the status quo of their organization's practices in information management. If you are *reactive*, not *proactive*, this book is not for you. This book is for people ready to be change agents. If you are looking for a silver bullet or a magic panacea to solve your information quality problems, skip this book. It is for people ready to roll up their sleeves and make information quality happen.

This book is not just for companies in the private sector that face competitive pressures just to survive. It is for government and other not-for-profit organizations that desire to truly serve their constituents—or customers—and accomplish their mission. This book is for those who desire to provide quality services at the lowest cost.

You are a candidate to receive value from this book if:

- You recognize that information is an important business resource and you want to maximize its value.
- You care about your customers, both internal and external, and desire to maintain accurate information about them and for them.
- You are fed up with the high costs of low-quality information and the resulting problems, and are asking, “is there a better way?”
- You are a business person who requires quality information or who creates information and you don't just want to do your job, you want to do the *right* job, efficiently and effectively.
- You are an information systems professional and you don't just want to build applications or databases, you want to build applications and databases that *add value* to the business and to the end customers.

Who should read this book:

- Information quality managers and staff responsible for information quality processes.
- Data resource managers and staff, and those responsible for developing data models and databases that represent and house the enterprise knowledge resources.
- Data warehouse managers and staff responsible for data architecture, data acquisition, and cleansing, transforming and loading data into the enterprise's strategic knowledge base.
- CIOs and information systems managers responsible for application development processes who are responsible for creating and managing

the *information* infrastructure—not just the information *technology* infrastructure—for the enterprise.

- Systems analysts and designers who desire to add value to the business—not just create technology “solutions.”
- Business information stewards who are responsible for care taking of parts of the information resources for the enterprise.
- Business managers who are owners of processes that create information used by others outside of their business area. This book will be of special value to business managers of information-intensive business areas such as customer relationship management, marketing, sales, order entry, claims processing, customer service, accounting, accounts receivable, human resources, account management.
- Business personnel, such as business analysts, actuaries, and other knowledge workers who are intensive information customers.
- Senior management who are concerned about the high costs and low success of IT and who desire to deliver shareholder value. It is senior management who must understand the absolutes of quality as a management tool and who must establish a management environment that enables information quality to increase business performance.

Why *You* Should Read This Book

In the *Harvard Business Review*, Schaffe and Thompson cite a survey showing 63 percent of companies that had embarked on TQM-based programs had failed to improve quality defects in products by even as little as 10 percent.³ This book aims to help you understand how to avoid the pitfalls when conducting information quality improvements and when implementing an effective information quality environment.

The Gartner Group states that most reengineering initiatives will fail because of lack of attention to information quality. Experience is revealing that more than half of data warehouses built fail to meet expectations because of poor information quality. This book seeks to help you be successful in all information-related projects by addressing and solving the real problems and causes of poor quality information.

Making any kind of change to the status quo requires effort and work. Information quality is neither automatic nor easy. If it was, there would be minimal information quality problems today. Most of us need guidance in applying new skills. This book seeks to provide that guidance to minimize your risk in making information quality happen.

³*Harvard Business Review*, 1/92 volume.

Organization of This Book

This book is organized into four sections:

Part One, “Principles of Information Quality Improvement,” deals with the fundamental principles of quality and of improving information quality.

Part Two, “Processes for Improving Information Quality,” describes how to measure and improve information quality.

Part Three, “Establishing the Information Quality Environment,” outlines how to implement an information quality environment.

Part Four, “Appendixes,” provides an extensive glossary, recommended reading, and bibliography.

Part One: Principles of Information Quality Improvement

Part One describes the fundamental principles of information quality. They are not theory—they 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. 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, “The High Costs of Low-Quality Data,” outlines the business case for information quality improvement. It describes why data that appears to be of satisfactory quality is, in fact, not. Examples highlight the high costs of low-quality data. Failure to solve information quality problems can be fatal to organizations.

Chapter 2, “Defining Information Quality,” defines information quality, what it is and is not. Information quality is not a soft measure. It in fact can be quantified in bottom-line terms.

Chapter 3, “Applying Quality Management Principles to Information,” describes the principles of quality in general: customer focus, continuous process improvement, and the use of scientific methods. It describes the concept of information as a product, and knowledge workers as information customers. We outline who has accountability for information quality. The answer may surprise you.

Part Two: Processes for Improving Information Quality

Part Two is the guide and road map of the processes to assess and improve information quality. It defines the processes of information quality improvement as a management tool for business performance excellence.

Chapter 4, “An Overview of Total Quality data Management (TQdM),” provides an overview of the TQdM (Total Quality data Management) methodology. It provides a thumbnail sketch of information measurement, assessment, and improvement processes. It further outlines a methodology for guidance in the data warehouse context.

Chapter 5, “Assessing Data Definition and Information Architecture Quality,” outlines how to measure and assess the quality of data definition and information architecture. This represents the product “specification” of the information product. Without quality of information architectures that store the enterprise’s knowledge resources, information quality will be much more difficult to achieve.

Chapter 6, “Information Quality Assessment,” describes how to measure, analyze, and report information quality in databases, data warehouses, or produced by the business processes.

Chapter 7, “Measuring Nonquality Information Costs,” describes the process of analyzing and quantifying the costs of poor information quality in business terms. It provides a road map for measuring the devastating impact poor-quality information has on business operations, mission accomplishment, customer satisfaction, and profits.

Chapter 8, “Information *Product* Improvement: Data Reengineering and Cleansing,” describes the process of information product improvement; that is, reengineering and cleansing. It describes how to audit and control the extract, transformation and cleansing, and load processes for data warehousing.

Chapter 9, “Improving Information *Process* Quality: Data Defect Prevention,” outlines how to improve the quality of the information product through business process improvement. It describes how to identify root causes of information quality problems, and how to plan and implement permanent information quality improvements.

Chapter 10, “Information Quality Tools and Techniques,” describes the various categories of information quality tools and techniques that support the processes described in this section.

Part Three: Establishing the Information Quality Environment

Information quality improvement is not simply “scrubbing” data to put it into the data warehouse. Information quality is not simply auditing data to measure it. Information quality improvement *means* fundamental changes in how the information systems organization defines, develops, and delivers its products and services, and fundamental changes in how the enterprise plans, organizes, manages, and performs its business processes, and measures its business performance.

Sustainable information quality improvement will be accompanied by a change in the way people think about their information products and information “customers.” Part Three describes the culture shift required to create a sustainable information quality environment.

Chapter 11, “The 14 Points of Information Quality,” outlines Deming’s 14 Points of Quality along with their direct ramifications for information quality improvement.

Chapter 12, “Information Stewardship: Accountability for Information Quality,” describes Information Stewardship, the people roles and accountabilities for information products.

Chapter 13, “Implementing an Information Quality Improvement Environment,” describes the steps to implement an information quality environment. It begins with how to conduct an information quality management maturity gap analysis and describes a set of steps to take from where you are.

Chapter 14, “Epilogue: Reaping the Benefits of Quality Information,” concludes with an epilogue rather than a conclusion. The information quality journey will never bring you to a final destination. Rather, it will bring you incredible adventures and joys as you bring your organization into the *realized* Information Age.

Part Four: Appendixes

Appendix A is an extensive glossary that defines terms from the information management information, general quality, and statistical analysis domains.

Appendix B contains an extensive bibliography for further reading, beginning with a recommended starter set.

Internet Resources Available in Conjunction with This Book

Because of currency of information and space limitations, there are resources about information quality products, techniques, and information quality best practice case studies available for book holders at www.infoimpact.com under *Information Quality Resources*.

How to Use This Book

This book is a concept book, a textbook, a reference book, and a practitioner’s guide. Depending on your work and interests, your use of this book may vary.

Do not assume that you must implement every step described in this book to implement information quality. There are many steps and activities that could be performed. You must identify your priority needs and concentrate on those processes and process steps (see Table I.1). Be eclectic and pragmatic.

There are many ways to use this book; however, shelfware is not one of them.

Table I.1 How to Use This Book

AREA OF INTEREST:	SEE BOOK SECTION:
What is information quality?	Part One
Principles of information quality.	Chapter 3 and Chapter 11
Cost justification for information quality.	Chapter 1, states it; Chapter 7 describes how to justify it
How to determine what level of maturity your enterprise has in information quality practices.	Chapter 13, with an example information quality management maturity grid, along with guidance in how to conduct a maturity assessment
How much is poor quality information costing your enterprise?	Chapter 7
Developing a data warehouse with focus on identifying authoritative source databases, and cleansing and transforming data.	Chapter 8
How to perform an information quality assessment.	Chapter 6
You know you have information quality problems, but need to know how to eliminate them once and for all.	Chapter 9
What tools are available to assist in information quality and how to evaluate them.	Chapter 10, and Internet at www.infoimpact.com : <i>Information Quality Resources</i>
Developing a data model and creating a quality model.	Chapter 5
How to set up an information or data stewardship program.	Chapter 12
Best practices in information quality.	Chapter 9, last section
How to implement an environment for sustainable information quality.	Part Three
Case studies or organizations making information quality happen.	Throughout the book, especially in Chapter 9, Chapter 13, and on the Internet at www.infoimpact.com : <i>Information Quality Resources</i>

The Maturing of the Information Age

Every economic era, from the Agricultural Age, to the Industrial Age, to the Information Age, has its paradigms and principles, along with its technologies. The Agricultural Age had a paradigm of “managed” land and crops with principles of cultivation. Its technology was cultivation tools. The Industrial Age paradigm was “managed” work with principles of mass production and specialized labor. Its technology was power and machines applied to work. The Information Age paradigm is one of “managed” information with principles of resource management and collaborative work. The full power of information technology cannot be fully realized until business and information systems management comprehend and *apply* principles of resource management and collaborative work to information.

Every economic era sees maturation of its processes through experimentation, trial and error, and then formalized process improvement. Data warehousing has focused new attention on information quality. This signals the beginning of a new phase of the Information Age: the Awakening. It is in this phase that organizations will challenge the Industrial Age paradigms and replace them with Information Age paradigms and principles. Business process reengineering initiatives are now replacing the vertical, functional management paradigm with horizontal, process, or value-chain management principles. The notion of data as a byproduct of business processes is giving way to the eureka of information as a direct product that has value beyond its immediate processes. Organizations that fail to recognize and manage information as a strategic business resource will fail in the realized Information Age.

As a product, information has processes that create and maintain it. Information has processes that use it. Information has customers, those knowledge workers who use information to perform their work. Information likewise has suppliers, those information producers who originate and add value to data through their work processes.

Information quality problems occur when data as supplied by the information producers does not meet the expectations of the knowledge workers as “information customers” or “information consumers.”

Data warehousing has exposed horrific information quality problems that become apparent when an organization attempts to integrate disparate data. Its quality *may* satisfy operational knowledge workers and functional processes, but fail miserably to satisfy the downstream knowledge workers and enterprise processes. The fact that these information quality problems have not surfaced until now is not because information quality has suddenly gone downhill. Information quality problems have been masked by bad business and systems practices over the years. To be fair, the earliest computing technology did not enable data sharing. As a result, all early application development methodologies were created around that technology limitation. That limitation resulted in building

isolated applications and islands of information independent of or only loosely interfaced with other applications and databases.

Information quality problems have been masked by layer upon layer of interface programs in which inconsistent data of one application is “transformed” into usable data structure and values required by another application area. Organizations have accepted this as necessary. However, the validity of this approach is seriously challenged as the weight of those layers of interfaces consume the time and resources of information systems organizations to maintain them. One organization discovered that the equivalent of 120 to 160 of its 250 highly skilled application developers spend their careers maintaining programs that simply copy data from one database, transform it, and put it into another database. Is this a valuable use of developer skills? Hardly, especially considering the people are a *consumable* resource and data is a *reusable* source.

Any new process takes time and experimentation to mature. Agricultural processes continue to be improved to provide greater quality of crops. For example, genetic alteration in foods like tomatoes have extended shelf life and increased taste. The Japanese began improving manufacturing process quality in the early 1950s when Dr. W. Edwards Deming introduced Statistical Quality Control techniques. This led to the dramatic turnaround of the postwar Japanese economy and the quality revolution and maturing of the Industrial Age manufacturing processes. It was not until the 1980s and the Japanese quality invasion of America that American business began transforming and improving American industry’s manufacturing processes.

Fifty years into the Information Age we are now seeing the same quality improvement techniques being applied to information as the product of business processes. We are now seeing the maturation of the information management processes and the dawn of the *realized* Information Age. In the Realized Information Age, quality information indeed becomes the new economic currency and the competitive differentiator.

Information Quality Improvement: Beyond Data Cleansing

This entire book addresses not just the techniques for cleansing data for the data warehouse. It addresses a complete set of processes to attack information quality *problems*. Before an organization can significantly improve its information quality, it must understand the paradigms of information as a business resource and as a product. Information quality improvement seeks to measure information quality, both data definition (data specification) and data content; analyze and identify root cause of data defects; and improve processes to prevent defective data. The sole reason for improving information quality is to improve business efficiency and effectiveness and end-customer satisfaction

by eliminating the problems caused by nonquality data. This book addresses the components of mature information management processes and organization culture that embody a customer satisfaction mind set to provide quality information. We will also identify the fundamental changes that are required to move an organization from data as a neglected and proprietary resource to information as a strategic and open business resource for competitive advantage and to achieve information maturity.

The information quality movement signals the beginning of the maturing of the Information Age. A word of warning is due the reader: Information quality improvement does not mean “more of the same” way of doing business. After all, we have been successful in the past. It does not mean building the same kinds of systems faster. It does not mean building bigger databases faster. Information quality improvement will force management to rethink the way it builds (or buys) applications and databases. It will force management to rethink the relationship of business processes and information. It will force management to rethink how it performs work. It will finally force management to rethink its performance measures and accountabilities for the resources of the enterprise.

Author's Warranty

If you are not able to apply ideas contained in this book to achieve value to your organization worth multiple times the cost of the book, I will personally refund to you the purchase price you paid for this book, no questions asked. Simply contact me at Larry.English@infoimpact.com for refund instructions as to where to send the book. All I ask is for you to give me a copy of your sales receipt along with a statement of what you tried that did not work, as well as your assessment of why it failed to result in value. No further questions asked.



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.

The High Costs of Low-Quality Data

“Quality is Free. . . . What costs money are the unquality things—all the actions that involve not doing jobs right the first time.”

—PHILIP CROSBY

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:

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."³

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).

Continues

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

⁴D. Bartholomew, "The Price Is Wrong," *Information Week*, September 14, 1992, pp. 26–36.

⁵R. Beck, "Item-pricing nice, but not for retailers," *The Tennessean*, August 24, 1997.

⁶Tara Eck, "Health care companies renew compliance focus," *Nashville Business Journal*, September 1–5, 1997, p. 24.

⁷Maggie Urry, "Book errors figure in Salomon \$770m pretax loss," *The Financial Times*, February 3, 1995. As cited in Madhavan K. Nayar, "Framework for Achieving Information Integrity," *IS Audit & Control Journal*, Vol. II, 1996, p. 31.

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.

Continues

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.

THE HIGH COSTS OF LOW-QUALITY DATA (CONTINUED)

- In 1994, an American bank employee transposed some digits on a bid for a bond from an Italian bank that resulted in a \$4-million loss to the bank. The Italian bank refused to return the money.
- In March 1997, a U.K. bank discovered it lost around £90 million (\$145 million) due to data errors in a computer model that caused inaccurate valuation of a risk manager's investment positions.
- A Catholic school sent out invitations to 5-year-old children to come for consideration to attend their elite school's kindergarten. In attendance was a woman born in 1888, age 105.

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

⁹B. Inmon, "Data Warehouse—Into the '90's," presentation given at the *All About IRM '92* conference, Beaver Creek, CO, July 21, 1992, p. 6.

ONE HUGE INFORMATION QUALITY PROBLEM THAT CAN'T BE LATE

The Gartner Group estimates the worldwide costs to modify the software and change databases to fix the Year 2000 problem to be from U.S. \$400–\$600 billion. T. Capers Jones says this estimate is low. He expects the costs to “fix” the Year 2000 problem to be around U.S. \$1.5 trillion, including lawsuits that will arise.

To restate the Year 2000 problem, the costs to fix this single, pervasive information quality problem represents an amount equivalent to nearly one-third of the U.S. federal deficit. Yes, this is an information quality problem. When the systems analysts and data analysts designed these databases with only two digits because of processing speed, reducing data storage costs, or because of ignorance, they inadvertently condemned their organizations, or customers, to an expensive fix.

To look at the Year 2000 information quality problem from another perspective, the 50 most profitable companies in the world earned a combined \$178 billion in profits in 1996. If the entire 1996 profits of these companies were dedicated only to the Year 2000 problem, the companies would cover only 12 percent of the total costs.

Fixing the Year 2000 problem will hurt the U.S. economy, reducing the growth rate by 0.3 percentage points in 1999 and 0.5 percentage points in the years 2000 and early 2001, according to a study by Standard & Poor's DRI.¹⁰

An even worse tragedy exists for many unenlightened organizations that are “solving” this problem the wrong way. Treating the Year 2000 as a programming problem rather than a data problem, they are attempting to change the date comparison algorithm rather than convert the fields to support a four-digit year value. This *programming* “solution” allows one to define a new century breakpoint such as the year '35. Two digit dates of 35 and above are considered to have a “19” century prefix. Dates 00–34 are considered to be dates with a “20” century prefix. This merely postpones when the Year 2000 problem affects the business and requires the problem to be solved—and *paid* for—again.

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

¹⁰Michael J. Mandel, P. Coy, and P.C. Judge, “ZAP! How the Year 2000 Bug Will Hurt the Economy (It's worse than you think),” *Business Week*, March 2, 1998, p. 47.

¹¹Philip B. Crosby, *Quality Is Free*, New York: Penguin Group, 1979, p. 15.

¹²J. M. Juran, *Juran on Planning for Quality*, New York: The Free Press, 1988, p. 1.

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.

¹³Samuel Boyle, *Quality, Speed, Customer Involvement & the New Look of Organizations* seminar, Excel, 1992, p. 17.

¹⁴“Burger King Dropping Beef Supplier,” *New York Times* News Service, reported in *The Tennessean*, August, 24, 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.