Zero Defect Data

Tackling the Corporate Data Quality Problem

by

Mark David Hansen

A.B., Mathematics
Cornell University
(1986)
M.S., Mathematics
University of Chicago
(1987)
Ph.D., Applied Mathematics
Massachusetts Institute of Technology
(1991)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN MANAGEMENT
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

January 1991

Signature of Author.......
....
Sloan School of Management
January 18, 1991

Certified by..
Dr. Robert P. Goldberg
Thesis Supervisor

Accepted by..
Jeffrey A. Barks
Associate Dean, Master's and Bachelor's Program

Zero Defect Data

Tackling the Corporate Data Quality Problem

by

Mark David Hansen

Submitted to the Sloan School of Management
in partial fulfillment of the
requirements for the degree of
Master of Science in Management

Abstract

This thesis represents a detailed and comprehensive study of data quality management from an information systems perspective. In the manufacturing world, managers and academics universally recognize the strategic importance of quality management for improving productivity and securing competitive advantage. Likewise, the central hypothesis of this thesis states that managing data quality should be equally critical for improving information systems productivity and strategic management. To evaluate this hypothesis, and offer some practical guidelines to managers on the topic of data quality management, five principle goals were established for this research:

- Formulate a working definition of data quality based on field work.

- Compile a significant body of anecdotal and statistical evidence to document the extent data quality problems in American business.

- Estimate the economic value of data quality management.

- Propose a methodology for implementing a data quality improvement program within an information systems organization.

- Evaluate the potential contribution of existing and emerging technologies for data quality management.

To achieve these research goals, two primary methods have been employed: (1) a review of the literature covering manufacturing quality management; (2)

extensive field work involving 37 information systems organizations contacted during the fall and winter of 1990. The bulk of the research effort, however, focussed on collecting data from the participating information systems organizations. A combination of interviews and surveys was used to identify the business impact of data quality problems and formulate the critical success factors for managing those problems. As a result of this work, our primary findings can be summarized as follows:

- Data quality can be defined along four parameters: accuracy, availability, interpretability, and timeliness.

- Data quality constitutes a significant and expensive problem for most IS organizations. Conditions are so bad that a majority of the IS executives surveyed admit their own departments produce corporate information which is less than 95% accurate.

- Improvements in data quality management have the potential to create significant productivity improvements within information systems departments.

- Information systems organizations need to implement data quality improvement programs analogous to the manufacturing quality improvement programs which many companies have established over the last ten years.

- Emerging technologies will play a critical role in facilitating data quality management. Presently, the technologies which most IS departments are planning to implement include data modeling and repositories. In the future, expert systems, intelligent user interfaces, data auditing tools, and statistical quality control methods could all make significant contributions to data quality management.

The author believes that during the 1990s corporate America will devote significant financial resources to solving data quality problems. The information contained in this thesis forms the basis for a solid understanding of these problems and a step toward their solution.

Thesis Supervisor:	Dr. Robert Goldberg
Title:	Visiting Scholar

Table of Contents

Acknowledgments..6

Dedication 7

Introduction..8

The Quality Revolution..13

Review of Manufacturing Quality..14

1.1 The Relevance of Manufacturing Quality to Data Quality..14

1.2 Quality: A Management Philosophy......................................16

1.3 Costs of Quality...18

1.4 Measuring Quality..19

1.5 Improving Quality is an Iterative Process...............................23

1.6 Organizational Responsibility for Quality...............................25

1.7 Adopting Quality Improvement as a Management Philosophy...27

Information Systems Quality..31

2.1 The Quality Movement in Information Systems....................31

2.2 The Economic Importance of IS Quality.................................33

2.3 Blueprint for a Successful IS Quality Plan..............................38

2.4 Examples of IS Quality Plans...40

Data Quality...48

Zero Defect Data: Defining Data Quality..49

3.1 A Methodology for Understanding Data Quality....................49

3.2 The Four Parameters of Data Quality......................................51

Documenting the Problem: Survey Results...56

4.1 Target Organizations and Goals for the Survey.......................56

4.2 Summary of Survey Results..59

The Economic Costs of Poor Data Quality..69

5.1 How Data Quality Impacts the Business...................................69

5.2 Estimating the Costs of Data Quality..74

Achieving Zero Defect Data...80

Implementing a Data Quality Improvement Plan......................................81

6.1 Five Categories of Change..81

6.2 Critical Success Factors for Achieving Zero Defect Data 90

6.3 Obstacles to Achieving Zero Defect Data 102

Technologies for Improving Data Quality 106

7.1 The Role of Technology in Achieving Zero Defect Data 107
7.2 Statistical Quality Control 108
7.3 CASE Tools: Standardizing Applications Development 114
7.4 Data Modeling and Repositories 118
7.5 Intelligent User Interfaces 121
7.6 Data Warehouses 123
7.7 Expert Systems Based Measurement Tools 127
7.8 Auditability Tools 130

Directions for Further Research 133

Appendix A 135
Appendix B 137
Appendix C 139
Appendix D 140

Bibliography 143

Acknowledgments

In order for academic research to serve a meaningful role in the mission of a business school, that research should be a collaborative effort between scholars and businessmen. However, mixing the rigors of academia with the practical necessities of business can be a difficult and frustrating challenge. In this respect, I am fortunate to have worked with a thesis advisor who is both a talented businessman and a thoughtful scholar. Dr. Robert Goldberg has been instrumental in helping me to uncover business problems which can be profitably studied in an academic setting. He introduced me to the topic of data quality and is responsible for pointing out its significance as a business issue. Dr. Goldberg's combination of sharp intelligence, entrepreneurial spirit, and sincere interest in teaching has contributed substantially to my practical education at MIT as well as to the work contained in this thesis.

In addition, I would also like to thank Professor Richard Wang for serving as my reader. His encouragement and insightful comments have been extremely helpful throughout the process of writing this thesis.

Finally, I owe a great deal of thanks to Eileen Glovsky and Sarah King. Eileen and Sarah were my co-authors on a 15.565 term project which grew into this thesis. In this capacity, they were instrumental in conducting the field study and contributing ideas during the early stages of my research.

To Lorraine

for patience and love
now and in the future

Introduction

Data Quality: A Vital Economic Issue

Most academics, economists, and businessmen recognize that we live in an information age where data represents an economic resource rivalled only by energy in importance. For banks, insurance companies, consumer marketing firms, retail stores, and health care organizations databases containing customer and market information represent their most important strategic assets. In the public sector, government data describing population distribution, incomes, and economic statistics drive monetary and political policies which affect the redistribution of hundreds of billions of dollars in wealth each year. As a result, problems with data quality can have tremendous consequences on both the macro and micro economic levels.

One can hardly pick up a newspaper today without running across a lead story describing the consequences of a data quality problem. The exhibit appearing at the end of this introduction contains two such articles which appeared on the front page of the New York Times during December 1990. The first describes bank overcharges to Americans holding adjustable-rate mortgages. In this case, a large number of banks were applying inaccurate mortgage rate data, resulting from careless miscalculations, to homeowners accounts. Government regulators estimate that this widespread incompetence in data quality management has cost American homeowners $8 billion in overcharges. The second article describes pervasive dissatisfaction with the accuracy of the 1990 US census. In this example, critics point to archaic statistical methods employed by the Census Bureau resulting in widespread undercounting in urban areas. As a result, many prominent politicians are calling for a statistical adjustment of the population count. At stake are billions of dollars in federal aid to cities and states.

These articles are dramatic, and serve to illustrate the tremendous impact which poor data quality can have on our economy. More importantly for many managers, however, is the fact that, at the microeconomic level, poor data quality management can cost firms staggering sums of money. In a well documented example, incorrect data produced by American Airline's reservation system cost the firm $50 million dollars in lost bookings during one summer[1]. Another example involves a leading investment bank where inaccurate data concerning customer trades results in one percent of all stock transactions being executed incorrectly. Trades which are executed incorrectly wind up in an error account which the bank owns. Although 1% may seem small, the net result is that over $10,000,000 of the firm's capital is continuously tied up in the error account and exposed to significant volatility risk.

Finally, and perhaps most importantly for the information systems professional, the direct link between quality and productivity in the manufacturing world implies similar consequences for the data center. It is widely believed that 40% of the information systems costs result from quality

[1] Computerworld, September 19, 1988, pg. 2.

related problems[2]. For many large organizations, this adds up to hundreds of millions of dollars each year wasted on unproductive reruns, downtime, redundant data entry, and inspection. Effective data quality management offers the potential to dramatically reduce those costs.

Organization of the Thesis

The focus of this thesis falls on data quality management within the information systems organization. As a result, this document primarily addresses the microeconomic consequences of data quality and methods which individual managers can implement to improve data quality. These consequences and methods are covered in three parts.

Part I of the thesis contains two chapters which establish the context for the study of data quality. To date, much of the research regarding the importance of quality management for corporate productivity and competitive advantage has examined the manufacturing environment. As a result, a great deal is known about manufacturing quality. This body of knowledge is summarized in Chapter 1 and serves as a framework for much of work presented in later chapters[3]. The reader who is already familiar with manufacturing quality is encouraged to skip this chapter and proceed directly to Chapter 2. Chapter 2 presents a summary of quality management within the information systems organization. It examines the principles which organizations involved in

[2] Merrill Lynch, Information Systems Department, personal communication, October 1990.

[3] To compile this chapter, the author relied heavily on the information and references presented in the work of Fine and Bridge [FB87].

our study have developed to manage the quality of service provided by the IS department.

The three chapters comprising Part II focus on the specifics of data quality management and present the results of the author's field research regarding the definition and economic impact of data quality. Chapter 3 presents the methodology used to extract data quality information from the 37 organizations participating in this research. The definition of data quality along four parameters (accuracy, availability, interpretability, and timeliness) is presented in this chapter. Each parameter is explained and illustrated with an example from the field study. Chapter 4 presents the results of the data quality survey in detail and describes the significance of these results. Finally, Chapter 5 estimates the potential cost savings which could be achieved by IS organizations through improved productivity resulting from better data quality management. This is accomplished by extrapolating from manufacturing quality studies using the distribution of survey data along the four data quality parameters.

Finally, Part III contains two chapters which offer practical advice to managers regarding the management of data quality. Chapter 6 presents a framework for organizational change within information systems departments. It goes on to identify the critical success factors for data quality management and discuss some of the principle obstacles to achieving superior data quality. Chapter 7 evaluates seven different technologies for improving data quality with which IS managers should become familiar. These are all technologies which could have a significant impact on data quality management over the next ten years.

The New York Times

VOL.CXL . . . . No. 48,447

Many Bills Are Found Incorrect On Adjustable-Rate Mortgages

Mistakes Discovered in Audits of Failed S.& L.'s

By IVER PETERSON

As many as one of every four Americans who bought homes with adjustable-rate mortgages may be receiving incorrect billings each month because of bank errors in calculating their interest rates, according to a report by the General Accounting Office. Some people are paying too much while others are paying too little.

The miscalculations were discovered in routine audits of mortgages issued by failed savings and loan associations that have been taken over by Federal regulators. But bankers concede that some adjustable loans, which have complex terms, are also being incorrectly adjusted by solvent institutions.

Overcharges Put at $8 Billion

The extent of the problem is disputed. John Geddes, a former Federal mortgage banking auditor who has been the chief whistle blower on the errors, estimates that 30 to 35 percent of the country's 12 million outstanding adjustable-rate mortgages, which include home equity loans, are billed incorrectly, with half the borrowers charged too much and the rest too little.

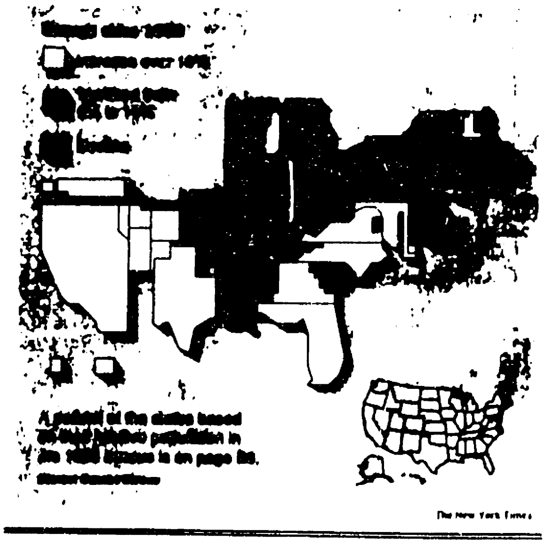
He says the net overcharges are about $8 billion, but says he has not estimated the undercharges. It is also unknown how many of these loans have already been paid off. Mr. Geddes's estimate is based on his review of 7,000 adjustable mortgages taken over by the Government in the savings bailout.

Experts cited in an October report of the accounting office, the investigative arm of Congress, put the error rate at 20 to 25 percent of these mortgages, but some banking industry officials contend that the rate is considerably less than that. Chase, for instance, audited its adjustable mortgages last year and found almost no errors.

VOL.CXL . . . . No. 48,462

NEW YORK, THURSD

Census Bureau Places Population at 249.6 Million



The New York Times

10.2% Rise in a Decade — Urban Officials Want Adjustment

By FELICITY BARRINGER

Special to The New York Times

WASHINGTON, Dec. 26 — After one of the most controversial head counts ever, the Census Bureau today put the population of the United States at 249,632,692, an increase of more than 23 million people, or 10.2 percent, over the official 1980 total.

The increase reflects almost a million overseas Federal workers, including military personnel and their dependents, who were included in the 1990 total under a new law. When they are eliminated from the count, as they were in 1980, the 1990 population comes to 248.7 million. The 1980 figure was 226.5 million.

While the Constitution requires that a population figure for each state be reported to the President by the end of the year for apportioning the number of representatives from each state, the Commerce Department can still decide that the survey was flawed and that it will statistically adjust the count. The department, the Census Bureau's parent agency, has said that it will determine by July 15 whether to do so

Part I

The Quality Revolution

Chapter 1

Review of Manufacturing Quality

The central problem of management in all its aspects in to understand better the meaning of variation. – W. Edwards Deming

1.1 The Relevance of Manufacturing Quality to Data Quality

A great deal of research has been published on the subject of quality in a manufacturing environment. This body of knowledge is relevant to the study of data quality because of the fundamental analogy between manufacturing systems and information systems. This analogy is illustrated in the following figure.

	Manufacturing	Information Technology
Inputs	Material	Data
Process	Line	Software

At the highest level, any manufacturing system can be viewed as a process acting on input material to produce output material. The input material could be parts, chemicals, subassemblies, or raw materials. Similarly, the output material could be parts, subassemblies, work-in-process, or finished product. The manufacturing process, frequently referred to as a line, consists of a combination of mechanical and chemical processes acting on the input material. In this manner, any information system can be viewed as a process acting on input data to produce output data. The input data can come from many sources (e.g. keypunch, database, optical scanner, output from another system). Similarly, the output data can take many forms (e.g., screen display, report, database entry, electronic mail message).

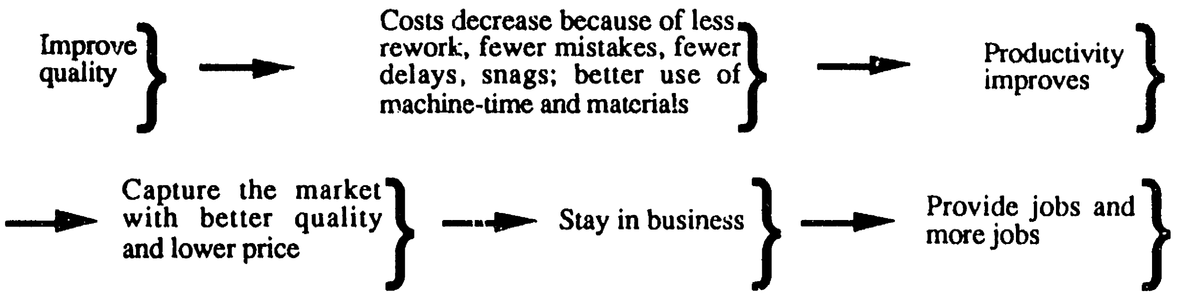
In manufacturing, quality management is the discipline of understanding and controlling process variation to minimize the number and variety of defects. This would suggest that at least one aspect of data quality management involves controlling the variation of software processes in order to minimize accuracy errors introduced into data. As discussed in Chapter 3, data quality management involves much more than simply controlling the number and variety of data accuracy errors. Nonetheless, regarding the management of data accuracy, there are many valuable lessons to be learned from the world of manufacturing.

Even more importantly, however, research in the manufacturing sector has a great deal to say about the link between productivity and quality management. In summary, this work indicates that investments which increase quality invariably more than pay for themselves with improvements in productivity. This suggests that many data center productivity problems may be related to poor quality management. In the very least, it indicates that manufacturing quality offers an excellent place to begin the study of data quality.

1.2 Quality: A Management Philosophy

The fundamental tenet of a quality centered management philosophy is that improvements in quality cause improvements in productivity. The relationship between quality and productivity was first formally studied and documented by Walter A. Shewhart [S31] in 1931. Shewhart's management principles were originally put into practice by Bell Laboratories' engineers working with American industry to improve productivity during World War II. After the war, however, Japanese companies became the foremost practitioners of quality management in the world. To a large extent the philosophies adopted by the Japanese manufacturers are those espoused by W. Edwards Deming. Deming has written that from July 1950 onward, the following chain reaction was on the blackboard of every meeting with top Japanese management that he attended.[1]

[1] W. Edwards Deming, *Out of the Crisis*, Massachusetts Institute of Technology, Center for Advanced Engineering Study, 1986, pp. 3.


Improve quality
Costs decrease because of less rework, fewer mistakes, fewer delays, snags; better use of machine-time and materials
Productivity improves
Capture the market with better quality and lower price
Stay in business
Provide jobs and more jobs

This chain reaction explains the motivation of companies which place quality improvement at the center of their management philosophy. Improving quality reduces manufacturing costs and improves efficiency. Better manufacturing facilitates a superior marketing position based on quality and price. Productivity gains allow a manufacturer to lower price and quality improvements create customer loyalty. The end result is increased market share and profitability.

Definition of Quality

Putting this chain reaction into practice requires a working definition of quality improvement. For many Japanese manufacturing companies the definition of quality improvement has been supplied by Deming:

> The central problem of management in all its aspects, including planning, procurement, manufacturing, research, sales, personnel, accounting, and law, is to understand better the meaning of variation, and to extract the information contained in variation. [2]

[2] W. Edwards Deming, *Out of the Crisis*, Massachusetts Institute of Technology, Center for Advanced Engineering Study, 1986, pp. 20.

Quality improvement, then, is the reduction in variation from some standard. This implies that the definition of quality is conformance to standards. In fact, "conformance to standards", has become widely accepted in the manufacturing world as the definition of quality.

Standards are defined from the top down through the manufacturing process[3]. At the top level, standards are defined in terms of the attributes a product must have in order to meet customer requirements. For an automobile, such standards might be expressed as "good acceleration", "responsive steering", "adequate gas mileage", etc. From these requirements, standards at the next level are defined (e.g., 0–60 mph in 6.0 seconds). This process continues, with the standards at each level being derived from the requirements set forth by the standards at the level above. Eventually, this process leads to measurable standards (e.g., weight, size, strength, durability) for parts and subassemblies.

1.3 Costs of Quality

In the manufacturing world, the costs of quality are well documented and understood. As we shall see, however, this does not mean that these costs are easy to measure. In this section, we summarize the costs of quality as described by two of the world's leading experts in manufacturing quality: Fiegenbaum (1983) and Juran (1964). These authors divide quality related costs into three categories: failure costs, appraisal costs, and prevention costs.

[3] For a more detailed description of this top down approach to defining quality standards, see the description of "quality function deployment" given by Hauser and Clausing in [HC88].

Failure Costs

Failure costs are the expenses which a company incurs as the result of a defective manufacturing process creating product which does not conform to standard. These costs can be divided into internal failures and external failures. Internal failure costs manifest themselves in terms of poor productivity. These are the costs of scrap, rework, retest, and manufacturing downtime. Costs related to external failures may be accounted for in terms of overhead and include returns and warranty expenses.

Appraisal Costs

Appraisal costs represent the expenses associated with inspecting the manufacturing process and measuring quality levels. Examples include: inspection of incoming materials, work in process, and finished goods; collecting statistical quality control data; and equipment inspection.

Prevention Costs

Prevention costs include the overhead expenses incurred to maintain quality improvement programs. Examples of these costs are: employee quality training programs, quality control data analysis, specific quality improvement projects, and quality planning. It is widely accepted that the cost of correcting quality problems rises exponentially the further downstream in the manufacturing process they are detected.

1.4 Measuring Quality

In order to manage quality, manufacturing organizations need measurable quality parameters which can guide their efforts. Commonly used parameters of manufacturing quality fall into two broad categories: product quality parameters and process quality parameters. Process quality can be further broken down into process effectiveness and process efficiency.

Product Quality

Product quality parameters include defect rates and conformance to specifications. These are measures of the quality of finished or intermediate manufactured products without regard to the fabrication process. For example, in a steel mill, product quality is frequently measured in terms of the defect rate. In this case, finished product can be visually or mechanically inspected and observed defects can be classified into various groups for statistical analysis. On the other hand, in semiconductor manufacturing facility, product quality is often defined in terms of conformance to specifications. For example, a particular memory chip may be designed for specified read and store response times. Finished memory chips can be tested and grouped according to their conformance with such response time specifications. In practice, chips which have superior response times are often grouped together and sold at a premium.

Process Effectiveness

Process effectiveness parameters include measurements of process variability, machine uptime, or late deliveries. These are measures of how well the manufacturing process is conforming to its design specifications. For example, in a chemical manufacturing plant, a particular reactor may have a specified temperature range for optimal operational effectiveness. The variability of temperature readings, taken at regular intervals, provides a measure of process quality.

Process Efficiency

Finally, process efficiency parameters gauge the productivity of a particular manufacturing process. These measurements include product throughput, rework, yield, and material waste. For example, in an injection molded plastics operations, a significant percentage of the finished product may have to be melted down and remolded because of poor quality. In this case, measurements of the yield and rework levels across time provide a means of interpreting process efficiency. Note that a process may be performing very effectively and still be quite inefficient. In this case, the process is simply poorly designed and no matter how effectively managed will never provide adequate productivity.

Statistical Quality Control

Statistical quality control (SQC) provides a common language for the measurement and interpretation of manufacturing quality parameters in all three categories. These methods were popularized by Deming in Japan during the 1950s and have recently enjoyed widespread acceptance in the United States and Europe as well. Here, we give a brief review of the principles of statistical process control which will be valuable in later discussions of techniques for managing data quality. This presentation is based primarily on the work of Ishikawa (1976). The reader is referred to Appendix D for an explanation of the principle SQC charting techniques.

Ishikawa identifies five principle purposes for using statistical quality control methods. These five purposes encompass the aspects of quality con'rol which are relevant in a manufacturing setting. They are: situational analysis, process parameterization, process control, cause and effect analysis, and acceptance or rejection.

Situational Analysis:

Situational analysis is the process of discovering which data quality problems a particular product or process is experiencing. It provides a means of determining which problems are most important so that they can be attacked first. The most commonly used charts[4] for situational analysis are the pareto diagram and histogram.

Process Parameterization:

Process parameterization consists of defining the normal limits of a particular manufacturing process with respect to a particular quality issue (e.g., the normal operating temperature range of a furnace). The method employed is usually a statistical analysis of a measured distribution. The most commonly used chart is the histogram.

Process Control:

Process control constitutes making effective use of process parameterization data in order to maintain a manufacturing environment within its normal operating limits. Control charts are typically employed to identify statistically unlikely events which signal an out of control process.

[4] See Appendix D for a description of all the charts discussed in this section.

Cause and Effect Analysis:

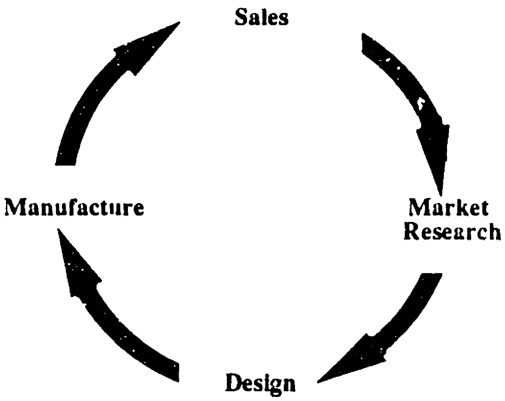
Cause and effect analysis involves tracking down the root cause of quality problems either by analytic reasoning or establishing statistical correlations between events. An example of analytic reasoning would be tracking down the cause of brake failures by exploding the parts diagram for brake˽ and examining each component. Ishikawa diagrams are the relevant chart for this activity. An example of establishing statistical correlations would be discovering a relationship between the temperature of a process and the tensile strength the product it produces. Scatter diagrams are most helpful for this activity.

Acceptance or Rejection:

Acceptance and rejection involves the inspection of product or incoming materials for defects. A statistically significant sample is taken for inspection and, based on the findings, an entire lot is either accepted or rejected.

1.5 Improving Quality is an Iterative Process

Within the manufacturing world there is almost complete agreement that the quality improvement process is iterative. This is not to say that organizations don't strive for breakthroughs in quality improvement, but rather that all improvements, whether dramatic or incremental, should be viewed as part of an iterative cycle of quality improvement. In general, this cycle is described as having four stages: design, manufacture, sales, and market research. Products are designed to meet quality specifications obtained through market research. From there, manufacturing processes are designed which guarantee that products meet those specifications. During the sales process, more market research is conducted to determine which quality improvements are demanded by the customers, and from there, the cycle repeats itself.



The Quality Improvement Cycle

Quality experts disagree about the techniques which should be employed at each stage. On the one hand, Deming argues that the goal at each stage is the same: reduce variability and uncertainty. Juran, on the other hand, maintains that specialized quality knowledge is required at each stage. In practice, most organizations seem to use a combination of techniques. Below we discuss two of the techniques which enjoy widespread use in the design and manufacturing points of the cycle.

Design for Manufacture

For example, during the design phase, a practice known as design for manufacture has gained widespread acceptance. A company practicing design for manufacture will strive to ensure that product designs facilitate a high level of quality during manufacture. An example of a design for manufacture goal would be to dramatically reduce the number of parts required to assemble a finished product.

Just in Time Manufacturing

During the manufacturing phase, many organizations have implemented a set of techniques which taken together are known as just in time manufacturing. The primary goal of just in time is to reduce inventories as a means of exposing quality problems in the manufacturing process. The logic behind this approach is as follows. High inventories of materials and work in process can be used to smooth variability and uncertainty in the manufacturing process. However, such inventories do nothing to improve the quality problems associated with such variability. One remedy is to remove the inventories in order to expose manufacturing variability. At first, operations may not run as smoothly, but this situation forces quality problems to be addressed, and in the long run a smooth, high quality manufacturing process is developed which maintains low levels of inventory.

1.6 Organizational Responsibility for Quality

In order to implement quality improvement programs, manufacturing companies must assign organizational responsibility for the various aspects of these programs. Among manufacturing quality experts, there is a general consensus that quality improvement breaks down into three major processes which can be assigned at different levels in the organization. (Deming 1986., Juran e.d., Schonberger 1982.) These processes can be expressed as maintenance, iterative improvement, and breakthroughs.

Maintenance

Maintenance involves ensuring that manufacturing processes remain stable and that quality does not deteriorate dramatically. The relevant SQC tools here are the control chart and histogram[5]. In general, quality maintenance is considered the primary responsibility of line workers.

Iterative Improvement

Iterative improvement involves targeting products and processes for evolutionary changes designed to reduce variability. Examples include rearranging production schedules to reduce the amount of retooling or instituting a preventative maintenance program designed to improve machine uptime. Pareto charts and Ishikawa diagrams are the most valuable SQC tools in this case because they enable the identification dissection of quality problems which are ripe for iterative improvement. Middle management is typically held responsible for iterative improvement.

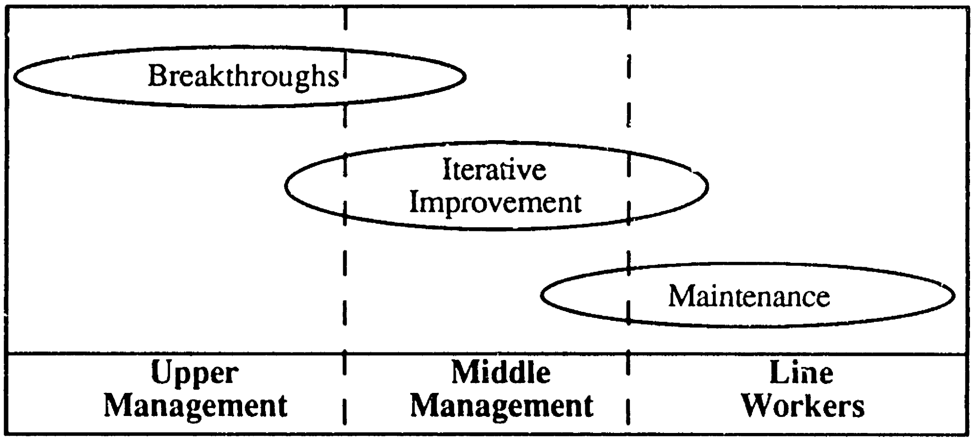
Breakthroughs

When a manufacturing process or product is totally redesigned in order to improve quality, this is considered a breakthrough. An example would be converting an entire production line to JIT, or redesigning a product to have 80% fewer moving components. Because systemic quality problems

[5] See Appendix D for a description of all the charts discussed in this section.

frequently reveal themselves in skewed defect distributions, histograms are the most valuable SQC tool for identifying products or processes which need major redesign. In general, upper management assumes most of the responsibility for quality breakthroughs.

The chart below illustrates the division of responsibility for these three distinct processes of quality improvement.

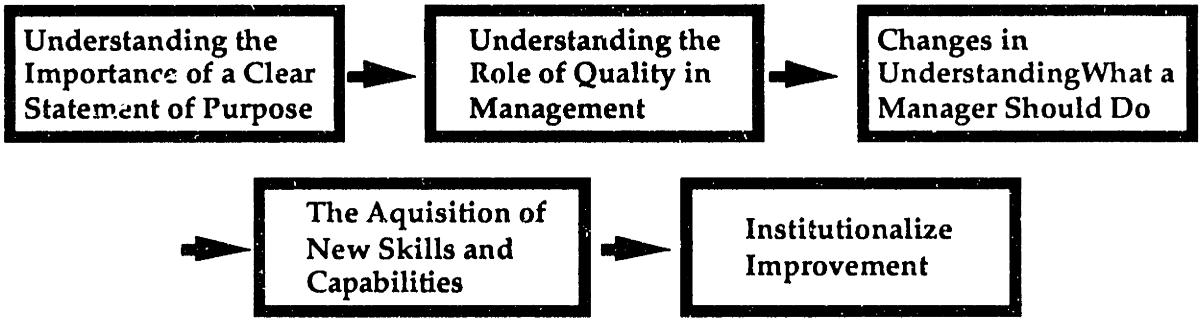


Allocation of Rcsponsibility for Manufacturing Quality Improvements

1.7 Adopting Quality Improvement as a Management Philosophy

The recent dominance of Japanese companies in manufacturing industries such as automobiles and consumer electronics has caused increasing concern in the United States. Many attribute the Japanese success to their early adoption and implementation of the quality improvement principles discussed in this section. More and more American businessmen are becoming convinced that the quality chain reaction discussed in Section 1.1 holds the key to international competitiveness. As as result, American companies and government officials are becoming increasingly interested in manufacturing quality.

The Center for Advanced Engineering Study (CAES) at the Massachusetts Institute of Technology provides a center of information to which such companies can turn in order to learn about quality improvement. As a result, CAES has worked with hundreds of organizations interested in implementing quality based management philosophies. Based on this experience, Tribus and Tsuda (1984) have identified five broad categories of change which companies need to implement in order to become quality focussed. Below appears the migration path encompassing these categories. Companies interested in becoming world class manufacturers can use this path as a guide to adopting quality improvement as a manufacturing philosophy.


Understanding the Importance of a Clear Statement of Purpose
Understanding the Role of Quality in Management
Changes in UnderstandingWhat a Manager Should Do
The Aquisition of New Skills and Capabilities
Institutionalize Improvement

Understanding the Importance of a Clear Statement of Purpose

Improving quality requires that cooperation of the entire organization. Such cooperation can only be obtained through leadership and clear principles that are well understood by all personnel. Because these principles must be universally understood and implemented, they must be expressed simply and their consequences should be clear. Top management must be careful to provide principles which give clear guidance regarding the importance of quality.

Understanding the Role of Quality in Management

Management must learn that quality is not an additional expense, but rather the solution to a wide range of productivity problems. In this manner, managers need to understand that increasing quality invariable causes productivity to increase and costs to decrease. This implies that quality control and improvement is a fundamental part of every managers job.

Changes in Understanding What a Manager Should Do

Tribus and Tsuda express the manager's role in a quality company as follows: "The people work in a system. The manager should work on the system. To improve it ... with their help." The managers primary responsibility is to work with his people to improve the design and operation of their systems to minimize variation and improve quality.

The Acquisition of New Skills and Capabilities

In order to continually improve systems design and operation, management and workers need to learn new skills and capabilities. Tribus and Tsuda group these into three categories: (1) The ability to check upon the performance of a system; (2) Cooperate with other people in generating methods to improve systems; (3) Understanding the meaning of statistical variation.

Institutionalize Improvement

Policies, procedures, and organizational structures need to be set up in order to institutionalize the process of continual quality improvement. This requires leadership from top management and the establishment of positions and job functions dedicated to quality improvement.

Chapter 2

Information Systems Quality

2.1 The Quality Movement in Information Systems

This part of the report sets the context for the study of data quality. In the course of this research, it became apparent that within real world organizations, the notion of data quality is always intimately linked with the broader concept of information systems quality. As a result, in order to understand data quality in its broadest sense, it is necessary to being with an examination of the principles which corporations and institutions have developed to manage the quality of service provided by the information systems organization. Toward this end, this section begins with an overview of the quality movement in information systems. It moves on to discuss the economic importance of IS quality and cost justification issues for IS quality initiatives. Finally, a generic blueprint for a successful IS quality plan is presented and three real world examples are given.

During the past decade, quality has been recognized as an integral component of business success for most corporations and large institutions. Quality Assurance (i.e., QA, the prevention of quality related business problems through careful design of procedures, processes, and products), has been a focus of attention primarily within manufacturing and engineering groups. Recently, however, management attention is turning toward the issue of information systems quality. Organizations are beginning to realize that significant business advantages and cost savings accrue to those who take a disciplined approached to building quality information systems.

In this manner, firms are placing more emphasis on data and information service quality assurance functions and corporate certification for their data products and services. Traditionally, as organizations recognized the importance of information technology to their businesses, formally defined positions such as *chief information officer* (CIO) and *database administrator* (DBA) were created to show a new level of corporate commitment to IS. Today, as firms come to grips with the importance of managing information systems quality, many are finding a need to create a new position: the *data administrator*. Typically, the data administrator reports directly to the CIO and is responsible for ensuring that the firm's information systems infrastructure and data are designed to serve the needs of the business. This position is often filled by a business person, rather than a technologist and has primary responsibility for managing information systems quality.

2.2 The Economic Importance of IS Quality

The primary reason that organizations are becoming increasingly interested in managing information systems quality is that the economic costs of poor quality are more widely appreciated today. In particular, the organizations participating in this study revealed three primary areas where information systems quality impacts their business: customer service, operations efficiency, and strategic competitive advantage.

Customer Service

Information technology is rapidly changing the quality of services which organizations can offer to customers. In many cases, the customer's impression of a company is formed at the point-of-service. Speed and quality of automated services have a direct impact on a customer's willingness to do business with a firm. For example, Merrill Lynch built a considerable consumer banking business by using information systems to provide an account with a single statement combing credit card, banking and brokerage information. Similarly, because of higher quality information systems technology, Federal Express has been able to offer better customer service in the area of package tracking. As a result, they are able to maintain a large and loyal express mail customer base. Winning the Baldridge Quality Award in

1990 represents another benefit of Federal Express' attention to IS quality[6]. They are the first service company to win this prestigious award.

Some information technologies provide both improved service to the customer and increased strategic advantage to the firm. As an example, consider the supermarket checkout. Owners greatly increased the speed and accuracy of product purchases with the introduction of bar code readers. As a result, customers can move through the checkout line much faster. However, an additional benefit accrues to the supermarket: the information collected has value as market data, and significantly improves the business's ability to track product sales and inventory.

Lower Operating Costs

Just as manufacturing companies have discovered that quality improvement programs can dramatically lower their operating costs, information systems organizations are discovering that implementing IS quality plans can improve their productivity. This is particularly true in the area of rework. It is a widely held belief among operations management experts, that service organizations spend 35%-40% of their budgets on rework[7]

. Since an IS department is a quintessential service organization, substantial savings should be achievable through an improvement in quality.

Strategic Competitive Advantage

[6] *MIS and the Pursuit of Quality,* Information Week, Jan. 7, 1990. pg. 36.

[7] Merrill Lynch, Information Systems Department, personal communication, Oct. 1990.

In the past, information systems were mostly used for the automation of record keeping, financial reporting, billing, and payroll. Management perceived IS as a service center to be included with the other overhead expenses. Today, a broader view suggests that information can serve as a corporate asset. In this manner, IS can provide a competitive edge and IS spending can be considered an investment. In particular, top management are learning that strategic management information systems (also called executive information systems or decision support systems) can provide many benefits such as:

- Detailed market analysis capabilities: Analyzing trends in customer buying patterns and preferences enables management to better segment the market and position a product.

- Improved product development: Detailed consumer information databases enable firms to tailor products to meet the specific needs of consumers.

- Improved performance measurement: Quickly understanding trends in sales and earnings allow companies to anticipate opportunities and threats ahead of the competition.

HMOs provide an excellent example of organizations realizing the strategic importance of information technology. The rising costs of health care and insurance are putting pressure on Health Maintenance Organizations (HMO's) to become more competitive. They must cut costs, manage with limited resources and high-risk expenses, while offering a wider range of coverage options and quality services to their members. As a result, Utilization Research and Planning departments have invested heavily in information systems to track trends in health service utilization.

For example, Health Care Company A regularly studies items such as patients' average length of stay in hospital, cost of stay, visits per member, visits per doctor in the plan, average cost of procedures (e.g. hip replacement, cardiac surgery), number of patients with certain diagnoses, admissions by age group. In this situation, systems can provide management with the information needed to make adjustments in the various plans, benefits packages, and rates offered to members which are critical to the organization's survival and growth. Marketing can also use the information to target new members (e.g. by community, through specific hospitals, etc.)

In this situation, IS quality is crucial in order to provide the flexibility and confidence necessary to take advantage of the available information. An HMO with branches in several communities needs to integrate data from many sources. In order for this data to be useful for decision making at the corporate level, it must be accurate and easily incorporated into integrated reports. This requires information systems quality planning to ensure standards of data accuracy and integration across the different branches.

In general, the avoidable costs related to poor information systems quality can be divided into three main areas: rework, internal customer service, and external customer service.

Rework

During a recent presentation at MIT, DuWayne Peterson, CIO of Merrill Lynch, estimated that 40% of their backoffice costs are attributable to rework. For example, a report running on a large database may take several hours. If an analyst fails to verify a data format or if the DBMS has poor checking tools, it may perform the entire search on an invalid field, only to report "no match found." Insurance Company B reports that rerunning a report which has failed in this manner can require six to eight hours of mainframe time and involve upwards of fifteen people.

Internal Customer Service

Internal customer service costs include a loss of service to internal users when the system crashes, as well as costs of lost opportunities when poor data leads to poor decisions. Repeated errors or a history of low quality and reliability will eventually force internal users to find other sources for information services.

For example, at Bank C, financial information concerning all publicly traded companies is stored on a large database. Standard applications have been implemented which allow investment banking analysts to automatically produce reports and run comparisons across companies based on this information. At any given time during the day, this system may have 50 or more users. When the system is not working properly, the users either become unproductive, or must resort to purchasing the information they need from expensive outside sources such as Compuserve. While the internal costs resulting form poor productivity amount to thousands of dollars for each hour of downtime, the external costs which may result from lost business are measure in the tens of millions of dollars.

External Customer Service

External customer service costs result from business which a firm loses because the level of service which it provides through its information systems is poor. An example with which we are all familiar is the automatic teller machine. Customers are more likely to keep bank accounts at institutions where the automatic teller machines are reliable, fast, and convenient to use. Another example is provided by brokerage houses. Investors are much more likely to place orders with firms which can process their securities transactions quickly and accurately.

2.3 Blueprint for a Successful IS Quality Plan

In many organizations the costs discussed above, if measurable, would be more than sufficient to justify an IS quality plan being put into place[8]. However, in order for the investment to pay for itself and provide an adequate return, the quality plan must be structured and executed properly. Among the organizations participating in this study, a informal consensus exists that a successful information systems quality plan must meet three important criteria:

- It must be aligned with a broader corporate quality plan.

[8] The problem for many organizations, however, is that these costs are very difficult to measure. See Section 6.3.

- It must place emphasis on meeting users' business needs.
- It must assign responsibility for qualit› within the organization.

Alignment with Corporate Quality Plan

For most manufacturing companies, the IS quality plan comprises one part of a broader quality plan for the entire organization. This is certainly true at Manufacturer Y, which is discussed below. The most important aspect of alignment with corporate quality plans involves having the commitment and leadership of the company's top management behind any IS quality initiative. Quality plans are implemented at all levels of the organization, but they must originate at the top because they require a very broad management perspective in order to be properly designed and executed. In this manner, it is essential that the CIO and CEO together be involved and committed to the IS quality plan.

Emphasis on User Needs

In addition, IS quality plans need to be driven by business needs, rather than the IS department's conception of what good quality is. Business needs include: generating useful and accurate reports without too much detail; access to data throughout the company; and up to date information about the business. Issues such as response time and support are also critical. In many instances, IS needs to evaluate itself as a service organization and approach quality from that perspective. For example, Merrill Lynch's IS Quality Plan includes a survey for users which asks them to rate IS on categories such as "Problem Response: Problems are resolved quickly"; "Problems are resolved correctly"; "Your contact is courteous and competent".

Assigning Responsibility within the Organization

Once organizations are aware that data quality is a problem, they must develop a strategy for assigning responsibility for improving quality. Many of our survey respondents believed that a corporate data administrator function needed to be implemented as the source of this responsibility. This approach is particularly prevalent in organizations that find interpretability of data difficult because of incompatible data standards across the company.

While we agree that a fairly high level IS executive should be given responsibility for data quality, proponents of the data administrator sometimes miss the importance of pushing responsibility for quality down in the organization. In particular, many leading firms believe that responsibility for data quality needs to be pushed back to the source of the data. This strategy results in a diffusion of responsibility across the entire organization which may increase the probability that the data entering the company databases will be accurate and further improve an organization's ability to trace data quality problems back to the source.

Commitment throughout the organization is becoming even more important with the trend toward distributed computing. Businesses moving from large centralized operations to networks of PC's will have less direct control over the quality of their data and systems. For such organizations, an increased awareness of the importance of IS and data quality is essential.

2.4 Examples of IS Quality Plans

In this section, three examples of IS quality plans are presented. These examples are drawn from banking, manufacturing, and healthcare, and serve to illustrate some of the principles outlined in Section 2.3.

Bank X [9]

Bank X recently completed a task force study of information systems quality. Part of the vision articulated in this study's recommendations asserts that "customer service and decision making at Bank X will be unconstrained by the availability, accessibility , or accuracy of data held in automated form on any strategic platform." In this manner, information systems quality is viewed as a means of improving customer service and decision making. In other words, quality is defined in terms of business needs rather than absolute measures. This vision is also aligned with the company's overall corporate quality goals of providing higher levels of customer service.

As part of the study, the task force established a set of business requirements for defining information systems quality. Of prime importance among these is the quick and flexible access to accurate information via decision support systems. Along with this requirement goes the need to standardize the meaning and presentation of reports. The goal here being that all reports should have a familiar format which communicates information effectively. Additionally, there are requirements related to the performance, integrity and security of systems. Finally, Bank X would like to move all of its data and systems development and maintenance to a repository-based environment.

[9] Throughout this thesis, the names of companies have been disguised whenever they are identified with specific, unpublished, examples.

As explained in Part III, a repository can form the cornerstone of an organization's efforts to standardize data definitions and ensure quality.

On the data quality side, Bank X has developed a number of operating principles to support these business requirements. Below are listed a few which they believe will move them toward their vision of information systems quality:

- Data should be defined consistently throughout the organization.
- Data should be entered into machine form only once, and this should be accomplished as close as possible to the point of origin of that data.
- Access to corporate data should be regulated to prevent contamination.
- Newly entered data should be subjected to automated edits, consistency checks, and audits as appropriate.
- Any uploading of data to the mainframe requires the same editing and consistency checks required of newly entered data.
- Minimize the number of products and technologies supported by the IS department.

To put these operating principles in perspective, it is necessary to briefly mention the primary information systems quality problems which the bank is trying to address. First and foremost among these is the incompatibility of data standards across the organization. Mission critical systems are currently running IMS, IDMS, DB2, Oracle, and flat file, based applications[10]. No effort

[10] For a description of these different database standards, please see: Date, C.J. 1990. *An Introduction to Database Systems,* Addison-Wesley.

has been made to standardize data definitions *within or across* these different platforms. This problem greatly impedes efforts to develop and deliver interdepartmental information systems. Of secondary importance are problems arising from a lack of centralized control over enduser manipulation of data. Data is frequently corrupted by enduser processing at the PC level. When this data is presented in management reports, it results in misinformed decision making. Even worse, if this data makes its way back into a corporate systems, it can permanently contaminate a corporate level database.

To meet the challenges posed by these problems, Bank X has outlined both organizational and technological initiatives. The organizational initiative centers around the creation of a data administrator position. The data administrator will be responsible for standardizing data definitions and insuring that information systems quality levels are high enough to meet the needs of the business as articulated in the task force's recommendations. The technological initiative centers around the conception of a *data delivery utility architecture* for storing corporate data[11]. This system's primary function is to serve as a regulated, central repository for data storage and standards enforcement. It will serve as the corporation's official source of data. Updating and accessing the information stored there will occur via a set of technologies designed to insure data quality.

Manufacturer Y

[11] See Section 7.6 for a more detailed discussion of the data delivery utility and data warehouses.

Manufacturer Y is in the process of incorporating an information systems quality plan into all aspects of its operations. Initially, management believed that their data was completely accurate. However, during the course of implementing the IS quality plan, two incidents were uncovered which served to change their minds. The first resulted from a request by the chairman of the board asking all the divisions to meet a 10% operating margin goal. At year end, all four divisions met the goal., but when the financial data was rolled up to the corporate level,operating margins slipped below 10%. This inconsistency resulted from the divisions developing differing definitions of revenue and expense.

Recently, there has been some standardization of terms to prevent this from happening again. In particular, the Director of Applications Technology & Planning, believes that Manufacturer Y will be moving to a data warehouse concept. Using this concept will entail corporate defining what data they want from the divisions, when they want it, and where is should reside. This may not insure accurate data, but it does address interpretability, availability and timeliness.

Manufacturer Y has been implementing a decision support system in one area for several of years now. One of the most time consuming aspects of this task, and one not foreseen by the planners, has been tracing the origin of the data that will be part of the DSS. This process took 8 months and served to educate the IS staff about the lack of integrity in the data.

Manufacturer Y has not opted for automated technologies to ensure data quality, but is attempting to push responsibility for accurate data back to the source. This is in line with their corporate quality goals to ensure quality at the source and not build inspection into a product or process.

Manufacturer Y is also attempting to simplify the methods by which they gather data to ensure data quality. One example of this is their system for maintaining employee addresses. Previously there were multiple inputs of an employee's work and home address. Maintenance of these systems was handled by a number of people, but not the obvious person -- the employee. Manufacturer Y redesigned its systems to consolidate this data gathering and made the employee responsible for maintaining the data. This helped to insure quality of the data at the source and reduced staff because there was no need for multiple records.

Like most organizations, Manufacturer Y believes that their financial data is the cleanest data that they have. Consequently, they see lack of data quality as having little impact on the bottom line. We believe, however, that as they move to using information as a strategic resource during the 1990's, they will encounter significant problems with data quality. Manufacturer Y is confident that its total quality program will assist them in correcting these problems, at the source, in an efficient manner. The real question is, are they sufficiently aware that data quality is a potential problem?

Hospital E

Hospital E has used the Integrated Computer System (ICS) for most operational activities since 1983. This system integrates clinical, administrative, and financial programs into one system of networked minicomputers. Most hospital systems have been developed piece by piece, with different departments using different hardware and software.

Services available include: on-line orders to Pharmacy and Lab, automatic lab follow-up, on-line results for almost all diagnostic tests, computerized planning and records of nursing care for patients, e-mail, various directories and schedules of medical staff, medical record requests, and patient care analysis.

This last function uses the PHASE system, a smaller database which stores detailed information about each patient, such as demographics, diagnosis, and itemized bills. This data is used to evaluate the efficiency and cost-effectiveness of patient care, provide information for clinical researchers, specific reports for administration and outside agencies like the Massachusetts Rate Setting Commission. Various departments also maintain small databases (e.g. Cancer/Tumor Registry, Operating Room, Labor and Delivery).

Management uses ICS to generate current budget and census screens . They also use PHASE as an "official" database; it is cleaner than ICS because data is put through additional edits once it is downloaded. PHASE is also more flexible in ad-hoc reporting.

A Data Quality Initiative was started in June 1990. Various department directors are involved in a task force to coordinate and enforce new standards, including a data dictionary and improved edit-checking procedures. That task force sees a need to place responsibility for data quality with the people who generate and consume that data, from data collectors to end-users to top management.

Current data quality techniques involve regularly run consistency checks, front-end edits with each weekly update, and manual verification. The biggest challenges involve documentation of key data elements, maintaining consistency in use of those data, implementing ongoing procedures for cleanup and maintenance, securing cooperation of people involved in data collection, and getting top management support for the resources necessary to accomplish these tasks.

IS programmers, analysts, and data administrators work together to provide user support, data access, report creation. Currently, many reports must be requested through the IS department (Direct access for certain users is being planned). Hospital E is also in the process of moving to a new platform of networked PC's and workstations for greater flexibility, power, speed, reliability and reduced costs.

IS quality is considered to be excellent in most areas. ICS is extremely reliable (unscheduled downtime was < 2 hours per year from 1983 to 1989). All levels of management use IS daily to monitor operations (budget, admissions reports, etc.) and strategic initiatives. Planning and Billing departments rely heavily on IS. Reimbursement from HMO's and government regulations also require a high level of data quality. IS quality closely reflects the hospital's mission: to provide the highest quality of care to its patients.

Part II

Data Quality

Chapter 3

Zero Defect Data: Defining Data Quality

3.1 A Methodology for Understanding Data Quality

In the previous chapter, we have concentrated on the broad issue of information systems quality. However, most organizations that are involved in information systems quality programs are unaware of the importance of data quality to these efforts. Unfortunately, most organizational quality efforts today seem to focus on the information system itself, overlooking the most important asset in the system -- the data. They seem to forget the old adage, "garbage in means garbage out."

In this chapter, the importance of data quality is considered in depth. In manufacturing, one would never dream of trying to improve the quality of production without placing primary emphasis on the quality of the materials input. Similarly, it is clear that as organizations begin to seriously address the issue of information systems quality, they will have to first deal with the quality of their data. The analysis provided in this chapter indicates the extent of this problem and difficulties associated with any solution.

One of the most difficult research tasks addressed in this thesis consists of simply defining data quality. Our approach in this matter has been to collect extensive data from information systems organizations and base the definition on our impressions from this field work. Over thirty-five organizations were contacted in connection with this study, and the difficulties encountered in trying to come up with an adequate definition of data quality stem from the broad diversity of opinion concerning the importance and relevance of its various aspects. In order to grapple with these difficulties, an iterative information gathering approach was taken. This approach consists of the following stages:

- initial interviews
- preliminary definition of data quality
- identification of business impact of data quality problems
- formulation of critical success factors for managing data quality
- second round of interviews using a data quality survey
- analysis of survey results
- formulation of recommendations for data quality management
- identification of critical technologies for data quality management

For a list of organizations which have provided input to the various stages in the methodology, see Appendix A.

3.2 The Four Parameters of Data Quality

In its broadest sense, data quality encompasses much more than simply the accuracy of the data stored in corporate files and databases. Analogously, we do not define the quality of a car simply in terms of its ability to move us from point a to point b. For a car, quality involves such issues as ease of use, expected maintenance, and durability. Likewise, data quality has many attributes. As revealed in the field research, organizations define data quality in terms of four parameters: accuracy, interpretability, availability, and timeliness. Examples of data quality problems along each of these four parameters appears below.

Data Quality Parameters			
Accuracy	**Interpretability**	**Availability**	**Timeliness**
30% of the addresses in an insurance company's customer database are incorrect.	A manufacturer cannot accurately measure its margins because each division defines them differently.	A bank cannot evaluate its real estate exposure because vital information is stored in incompatible data definitions.	Hospital capacity underutilized because of delays updating admissions information about bed availability.

Accuracy

Accuracy measures the correctness of the information stored in data. This is the concept which C.J. Date [D90] refers to as data integrity. In addition to correctness, it includes such data quality problems as inconsistency and redundancy.

Example:

Bank D executes upwards of 20,000 stock trades per day for its customers. Most of these orders come in over the phone and must be entered into the systems. Although the data entry process is 99% accurate, in this case that margin of error still leaves the firm exposed to substantial risk. Trades which are executed incorrectly are put into an error account which is owned by Bank D. The average value of a trade (across both institutional and retail) is at least $50,000. At a 1% error rate, roughly 200 trades per day are executed incorrectly resulting in $10,000,000 of the firm's capital tied up in the error account. Given the volatile nature of the stock market today, this is not an acceptable risk.

Interpretability

Interpretability measures how easy it is to extract understandable information from the data. Many factors such as data definitions, report formats, and information processing algorithms, impact the interpretability of data. As anyone who has every tried to sift through a one hundred page stack of computer printout knows, data can be accurate, but remain totally uninterpretable and therefore useless.

Example:

Insurance Company F is currently completing the development of a corporate level management information system which has been under development for the past two years. Like many organizations, they have experienced a great deal of difficulty presenting information to managers in a usable form. It remains easy for them to dump an enormous pile of reports on someone's desk, but information in this form is useless. The goal of the corporate MIS is to be able to summarize data into information which is easily absorbed and understood by management.

Another twist on the interpretability theme comes from the Manufacturer Y example discussed in Part I. When the CEO set a goal of 10% operating margins for each of the four divisions, all divisions were justifiably proud that they met the goal. However, the chairman was a little surprised that the company as a whole didn't meet the goal. This is especially perplexing when one realizes that there are no financial structures in between the divisions and the chairman. What was the problem? Each division had slightly different definitions of operating margins, and definitions that differed from the corporate ones. The CEO found out that while divisional autonomy is helpful in managing diverse businesses, the lack of communication regarding data definitions can hinder the interpretability of data.

Availability

Availability measures how quickly information stored in corporate data can be gathered by the people who need it. One very important aspect of availability addresses the ease with which information stored in systems and databases across an organization can be brought together for analysis and reference. This is often referred to as the data integration problem.

Example:

Like many regional banks, The Bank X is concerned about the financial health of their real estate portfolio. The government may also be interested, as auditors begin taking a closer look at real estate lending across the country. In order for IS to build a real estate portfolio system which could monitor the financial status of the bank's loans, data needed to be accessed from the commercial loan system (when lending information resides) and the real estate appraisal system (where current asset valuations reside). Unfortunately, these two systems use incompatible data definitions and, as a result, it remains very difficult to build the necessary real estate portfolio system.

Timeliness

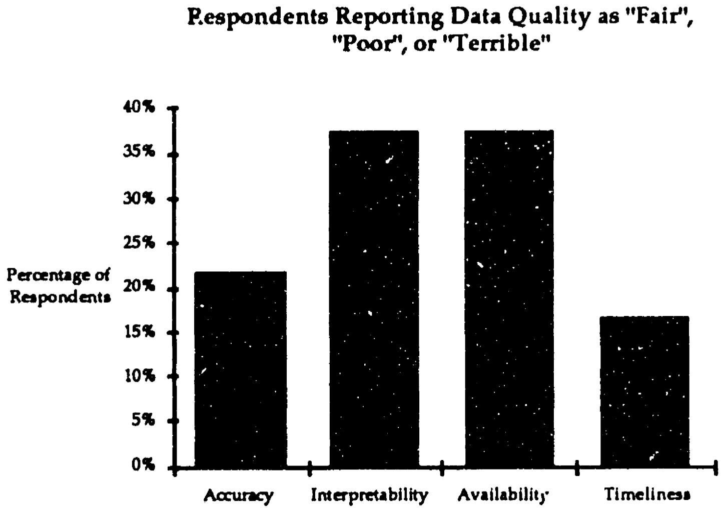
Timeliness measures how up to date the information stored in the data is. In particular situations, this can be more important than accuracy. For example, a 99% accurate, but up to date, mailing list is much more valuable than a 100% accurate, but five year old, mailing list.

Example:

Within hospitals, the timeliness of information reporting patient conditions and availability of beds is critical for effective and efficient administration of health care. Toward this end, Hospital E is developing a Bed Control System which will provide accurate, up-to-the-minute census of available beds, on which floors, which units, etc. This will greatly improve the staff's coordination of "patient flow" - admissions, discharges, transfers; allow doctors and nurses to easily locate patients, allow housekeeping and other services to schedule work more efficiently, etc. Currently , the hospital takes census every midnight and manually manages the daily use of beds.

Survey Data

Having established the four parameters defining data quality during the initial interview stage, a survey was prepared to measure organizations' perceptions of quality problems along these parameters. Below appears a graph which indicates the percentage of respondents perceiving significant data quality problems along each of the four parameters: accuracy, interpretability, availability, and timeliness. A more detailed discussion of the survey results is given in the final section of this part.

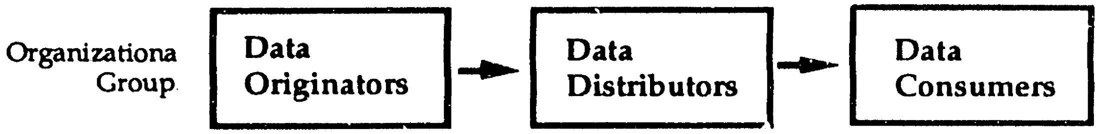


Chapter 4

Documenting the Problem: Survey Results

4.1 Target Organizations and Goals for the Survey

After the initial rounds of conversations with IS organizations, and formation of a framework for studying data quality, a survey was drawn up to measure the parameters along which data quality had been defined. During the course of the investigations, however, it became apparent that different organizations had widely different needs with respect to managing data quality. In general, these needs varied with the economic function being performed in the *data value chain*. This value chain for data appears below and represents a division of organizations into three groups with respect to their data function.



Examples:			
	Supermarkets	Information Resources	Marketing Organizations
	Retail Stores	Dow Jones	Banks
	Stock Exchanges	Reuters	Insurance Companies
	Credit Card Companies	TRW	
		Dun & Bradstreet	

The Value Chain for Data

Data originators are those organizations which generate data having value to other organizations. Supermakets which collect and resell point of sale data are the most common example. Data distributors purchase data from the originators and resell it to consuming organizations. Information Resources, Inc. (IRI) is an excellent example of a data distributor. They purchase point of sale data from supermarkets, analyze and process it, and then resell it to consumer marketing firms such as General Mills. Finally, consuming organizations are those which acquire data generated externally. Banks offer an example of data consumers because they buy credit data from distributors such as TRW and Dun & Bradstreet.

With the exception of distributors, most companies do not belong solely to one group or another. In fact, most organizations are very vertically integrated with respect to data. Frequently, different departments within an organization will perform different functions with respect to data. For example, marketing organizations consume data on customer buying habits from information compiled by the finance organization. In addition, this information is frequently supplemented with data purchased from a distributor. As a result of vertical integration, most IS organizations have responsibility for data origination, internal distribution, and consumption.

More and more, they also have responsibility for purchasing and integrating data from distributors.

Because the focus of this research is data quality management from an information systems perspective, it was decided that the survey should target organizations which are vertically integrated with respect to data. In this manner, the target audience for the survey was corporate level IS staff, preferably high level executives or DBAs.

Having established a target group, goals for the survey were developed along the following lines:

- Capture, with one direct question, the respondent's general impression of the accuracy of the information provided by the IS organization for corporate consumption.

- Measure respondents' perceptions of data quality along the parameters of accuracy, interpretability, availability, and timeliness.

- Test the hypothesis that organizations become particularly concerned about data quality when they consider data migration.

- Measure data availability in the context of integrating departmental data at the corporate level.

- Measure the adoption of a few critical user interface technologies.

- Measure consumers impressions of the quality of third party data.

- Understand the extent of the data auditability problem.

- Measure the adoption of a selected group of techniques for certifying data quality.

4.2 Summary of Survey Results

When interpreting any set of data, the biases of the source must be taken into consideration. In this case, the survey was aimed at IS managers. In many cases, truthful answers to the questions could reflect very poorly on the employers of the respondents. In this manner, the data is believed to represent a more optimistic vision of America's corporate data quality problem than an unbiased source might offer. Bearing this in mind, several observations about the data are worth mentioning here:

- Data quality is probably quite bad when the people who manage that data are willing to admit significant problems exist. Over 50% of the respondents rated the accuracy of the information produced by the IS department at 95% or less.

- Interpretability and availability were judged to be the parameters along which data quality problems were most significant.

- Inconsistent data standards across departments are broadly admitted to be a major problem facing IS organizations.

- On the issue of departmental data, where our survey respondents may have felt freer to be honest, there was considerable criticism. Almost 90% of the respondents maintained that departmental data was not of suitable quality to base important business decisions. Over 25% recommended not using departmental data for *anything* important unless it was checked twice.

- Assigning responsibility for data quality within the organization is a top priority of most IS departments.

- A majority of respondents expressed difficulty in tracking down the sources of their data quality problems.

A copy of the survey appears in Appendix B. In what follows, we present the survey results with respect to each of these goals. Since each of these goals corresponds to a particular survey question, the following sections each begin with that survey question presented in italics.

Frequency of Errors in Information Systems Products and Services

Consider the information products and services such as reports, decision support systems, accounting records, customer files, customer service, mailing lists, which are produced from corporate data. How accurate are these products?

□.....*100% Information products and services never contain errors.*

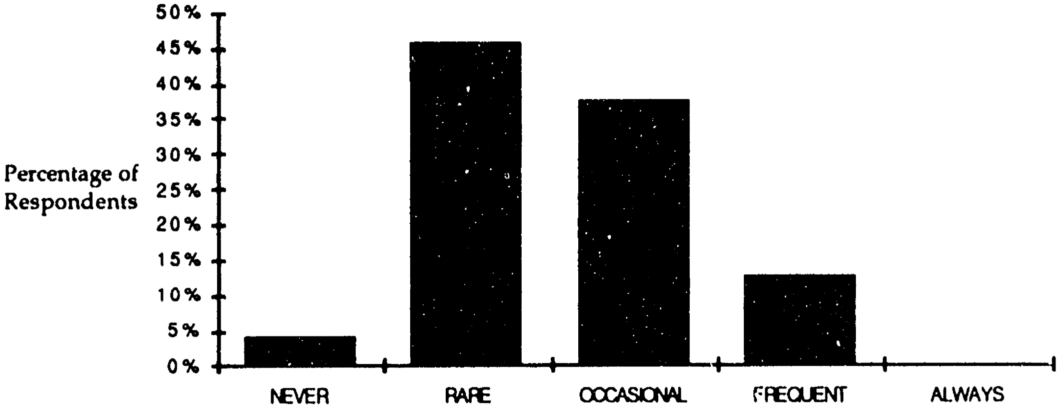
□.....*99% Information products and services rarely contain errors.*

□.....*95% Information products and services occasionally contain errors.*

□.....*9 0 % Information products and services frequently contain errors.*

□.....*below 90% Information products and services are plagued with errors.*

Estimate the frequency of errors in information produced by the IS department.



Responses to this question indicate that over half of the participating organizations rate their products and services at 95% accuracy or less.

Ratings of Accuracy, Interpretability, Availability, and Timeliness

Now consider the data underlying these information products and services. The quality of this data can be defined in terms of four parameters: accuracy, interpretability, availability, and timeliness. Accuracy measures the correctness of the information stored in data. Interpretability measures how easy it is to extract understandable information from the data. Availability measures how quickly information stored in corporate data can be gathered by the people who need it. Timeliness measures how up to date the information stored in the data is.

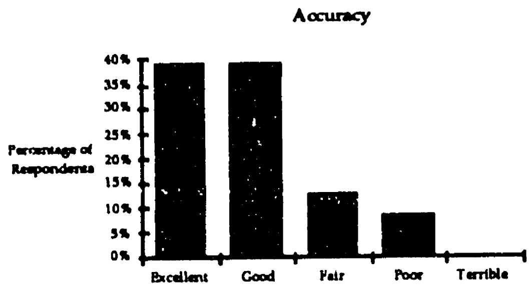
Please estimate the quality of your corporate data along these four parameters.

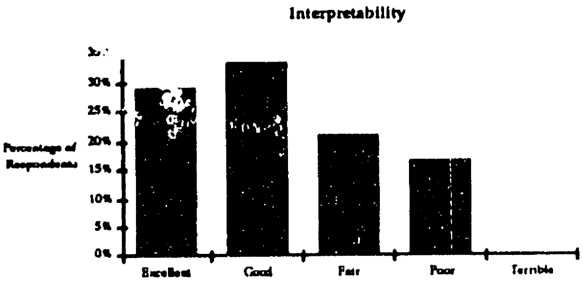
Accuracy: □ *Excellent.* □ *Good.* □ *Fair.* □ *Poor.* □ *Terrible.*

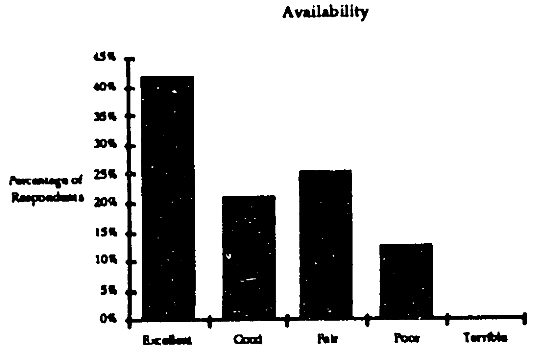
Interpretability: □ *Excellent.* □ *Good.* □ *Fair.* □ *Poor.* □ *Terrible.*

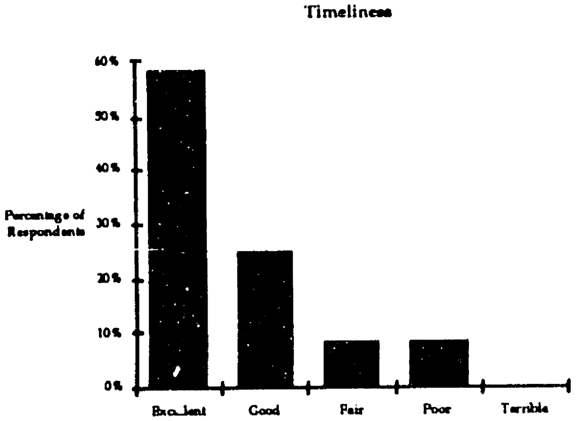
Availability: □ *Excellent.* □ *Good.* □ *Fair.* □ *Poor.* □ *Terrible.*

Timeliness: □ *Excellent.* □ *Good.* □ *Fair.* □ *Poor.* □ *Terrible.*









These responses indicate that IS organizations currently view interpretability and availability as the parameters along which data quality problems are most acute. It should be noted , however, that these two parameters are widely accepted by IS organizations as their direct responsibility. Hence, it is not surprising that our survey respondents are most sensitive to problems in these areas. In the author's opinion, data accuracy are potentially more severe than recognized by IS organizations. In this manner, as companies begin to assign organizational responsibility for data accuracy we may witness an increased awareness of the severity of accuracy related problems.

Major Data Quality Challenges Facing Your IS Organization

List the three major challenges your IS organization faces in maintaining the quality of corporate data.

This question clearly does not lend itself to graphical analysis, so the responses of the participating organizations have been summarized below. In general, the four most frequently cited challenges to maintaining data quality are:

- Assigning responsibility for data quality.
- Managing data in a decentralized environment (i.e., availability).
- Insuring the quality of data feeds (e.g., data entry or third party data).
- Converting data into usable information.

Comfort with Quality of Departmental Data

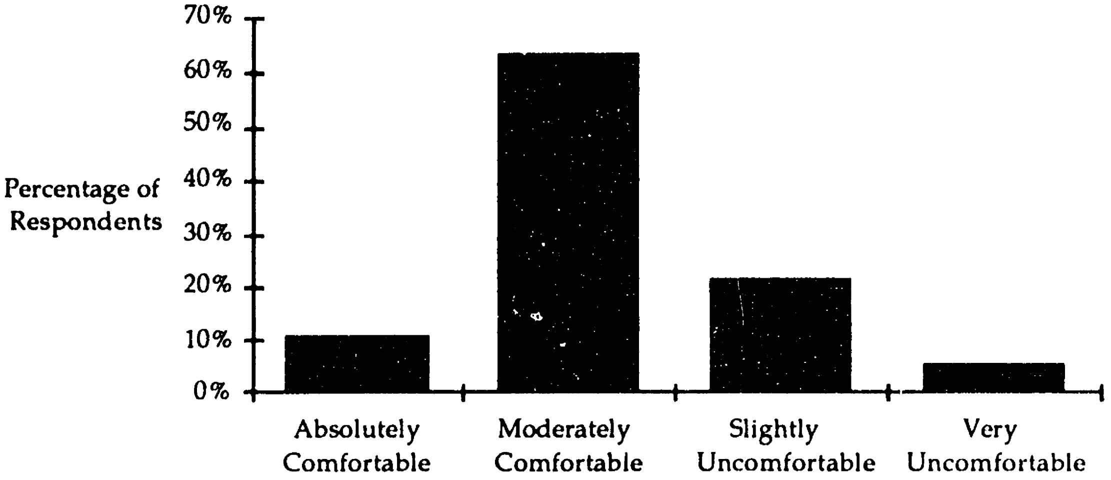
Consider the data stored and maintained within your company's various departments. How comfortable do you feel about using this data?

□.....*Absolutely comfortable. Important business decisions are based on this data.*

□.....*Moderately comfortable. Suitable for informal analysis and internal use.*

□.....*Slightly uncomfortable. Check twice before using it for anything important.*

□.....*Very Uncomfortable. The departmental data is almost unusable.*


How comfortable do you feel about using departmental data?

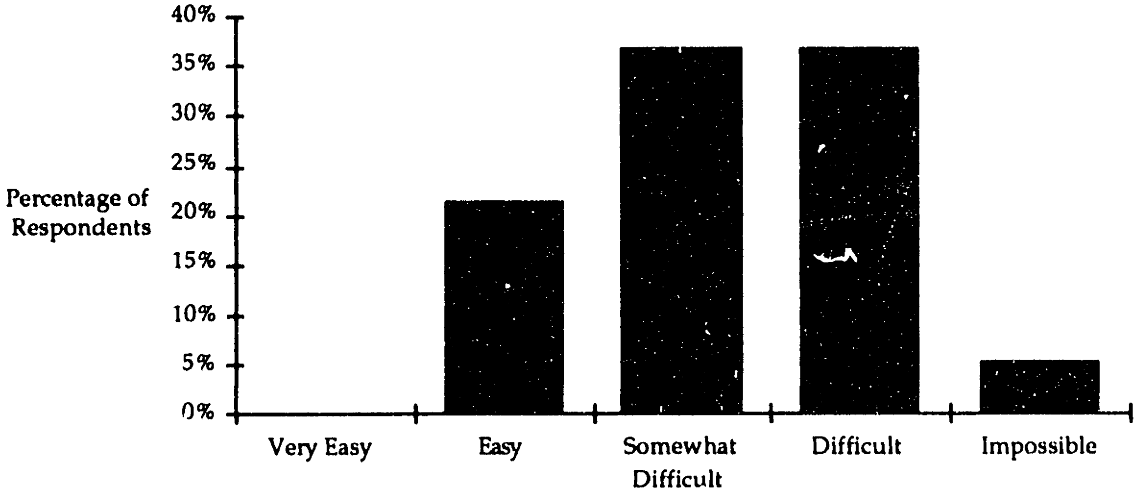
These results indicate that corporate IS views the quality of departmental data with some suspicion. Most expressed only moderate comfort, and more than 25% indicated some degree of discomfort, with the quality of the departmental data.

Ease of Integration for Departmental Data

Many IS organizations would like to be able to use the information stored in their departmental databases for corporate or inter-departmental purposes. (e.g., building inter-departmental applications or decision support systems) How easy is it for your IS organization to use data stored in departmental databases?

□ *Very easy.* □ *Easy.* □ *Somewhat difficult.* □ *Difficult.* □ *Impossible.*

How easy is it to integrate departmental data into corporate level applications?



The responses to this question clearly indicate significant availability problems with respect to departmental data. In fact, most IS organizations are implementing technologies and procedures designed to improve departmental data integration[1]

Obstacles to Using Departmental Data

List three obstacles your IS organization faces when trying to use departmental data.

Again, this question also does not lend itself to graphical analysis and a summary of participants' answers is given below. Specifically, the two most frequently cited obstacles to using departmental data are:

- Incompatible data standards.
- Unwillingness of different departments to share data.

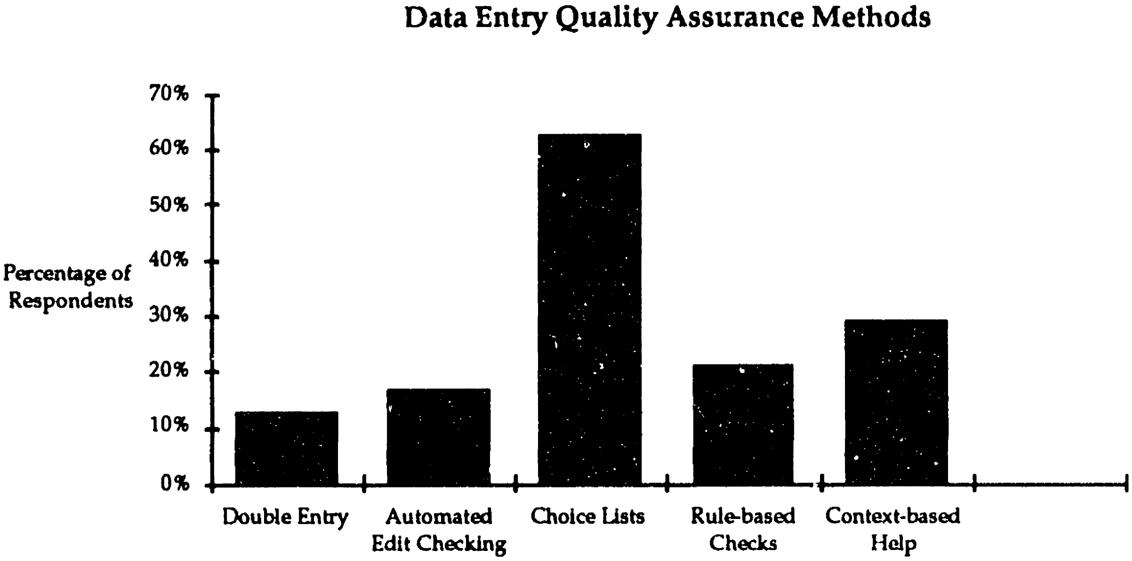
[1] See Sections 7.4 and 7.6.

Data Entry Quality Assurance Methods

A large portion of the data in corporate databases is generated internally through user interfaces to applications software. Which of the following technologies does you company use to ensure data quality at the user interface?

□ *Double entry.* □ *Choice lists.* □ *Context-based help.*

□ *Automated edit checking.* □ *Rule-based checks.*

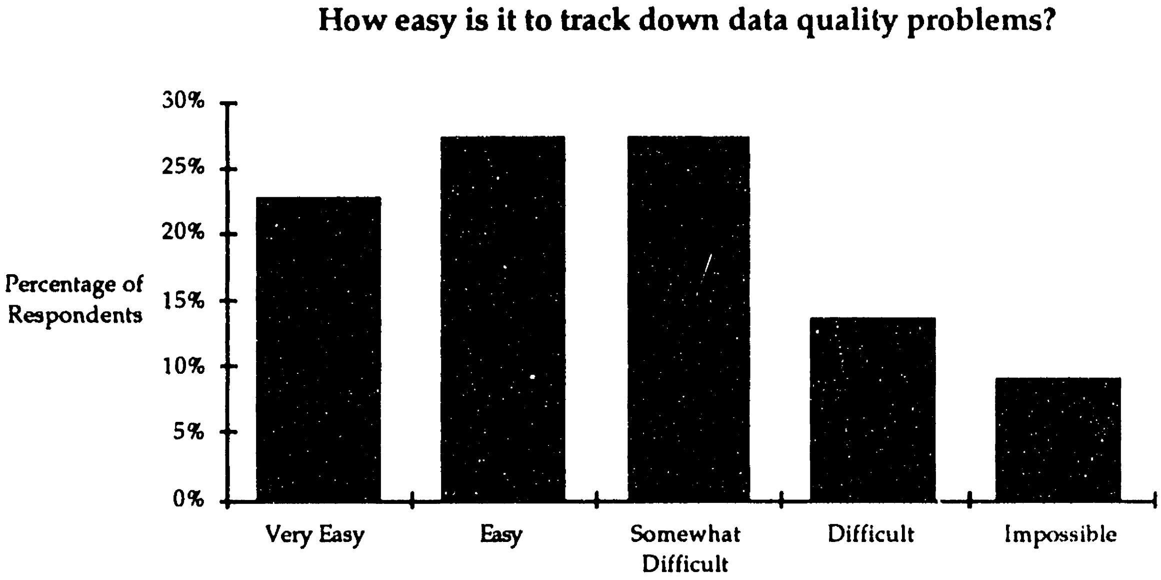


The significance of this result lies principally in the lack of attention that IS organizations surveyed have given to designing user interfaces. Although all four of the technologies listed can potentially have a significant impact on data quality at the source, only choice lists have been implemented by a significant number of the respondents.

Data Auditability

When data quality problems are discovered, how easy is it to track down the source?

□ *Very easy.* □ *Easy.* □ *Somewhat difficult.* □ *Difficult.* □ *Impossible.*



Responses to this question reveal that tracking down the sources of data quality can be difficult. This result points to an emerging need for some technology to address the issue of data auditability[2].

Use of Technology

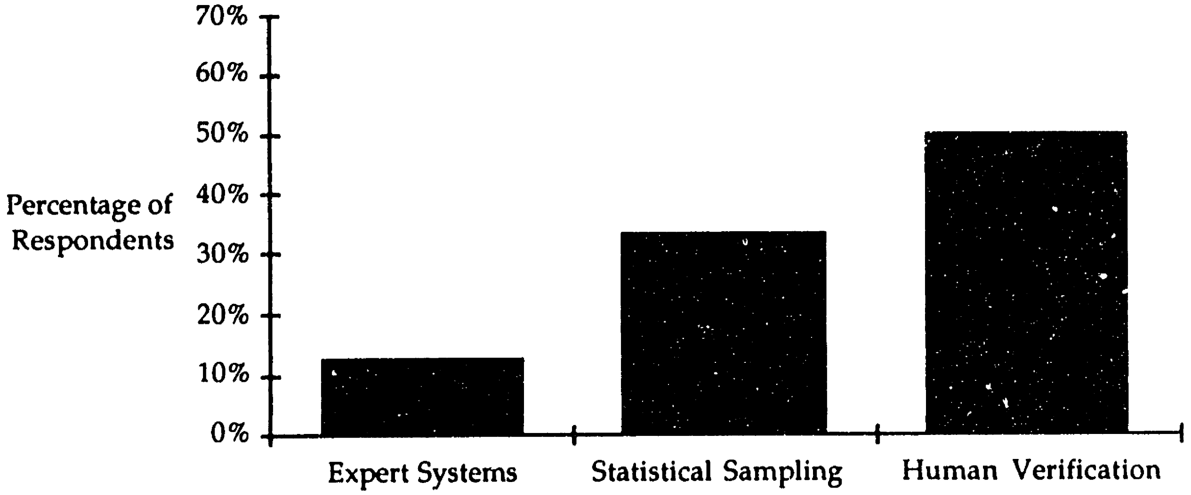
Which of the following technologies does your company use to ensure data quality?

□ *Expert Systems* □ *Statistical Sampling* □ *Human verification*

□ *Other* ____________________

[2] See Section 7.8.

Technologies Used to Ensure Data Quality



Responses to this question show that automation of the data quality inspection process is still in its infancy. The reliance on human inspection as a means of controlling defects is the hallmark of an immature quality management effort.

Chapter 5

The Economic Costs of Poor Data Quality

5.1 How Data Quality Impacts the Business

Having formulated a working definition of data quality, the research effort focussed on the impact of data quality on the business of an organization. Needless to say, this is a very broad area, with data quality issues touching virtually every aspect of business. In order to get a handle on this complexity and sort through the hundreds of anecdotal stories encountered during field research, three parameters were utilized to define the primary interfaces between business operations and data. These parameters are management decisions, customer service, and operations productivity. Field study revealed that most data quality problems cost organizations money by creating difficulties in one of these areas.

Management Decisions

Because it is strategic to an organization's success, the managem^nt decision making arena offers the most potential for data quality to impact the bottom line. With the advent of decision support systems, more and more of the information which top management relies on to guide its thinking originates within corporate databases. If quality problems occur, either with the accuracy, interpretability, availability, or timeliness of this data, severe consequences can result for the bottom line.

Example:

Most large investment banks have implemented risk management systems within the past few years. Ri.. management systems are intended to gather information documenting all of the securities positions to which the firm's equity is committed at any given moment. Ideally, the system should be continuously updated so that the firm can be continuously aware of the financial risks to which it is exposed. When functioning properly, the risk management system serves as a tool which executives can use to help them judge which securities positions the firm should stay in and which to close out.

While speaking at MIT in October 1990, the CIO of Merrill Lynch, DuWayne Peterson, related an incident which illustrates the importance of data quality to such systems. Merrill had been very successful trading a particular set of derivatives of mortgage backed securities known as IOs and POs. When held together in the proper proportions, a portfolio of IOs and POs is very stable.

However, each security held alone is extremely volatile and sensitive to interest rate fluctuations. Merrill Lynch was very active in the market for IOs and POs, and as a result held a large portfolio of these securities. Because of data availability and timeliness problems, however, the risk management system was unaware that the firm had closed out its IO positions and still held the POs for a brief period. During that time, interest rates changed dramatically causing the POs to drop rapidly in value, and the IOs to rise. Merrill's resulting net loss totalled in the hundreds of millions of dollars.

Customer Service

When poor data quality causes a firm to offer poor customer service, there can be a direct negative impact on the bottom line. Frequently, this impact is hard to measure because customers may stop buying from a firm without specifying any reasons. A banking customer, for example, may switch an account to another bank because the original bank's automated teller system experiences a high level of downtime. Sometimes, however, the cost of a data quality problem with respect to customer service can be directly measured.

Example:

Manufacturer Y is one of the largest providers of optical fiber in the world. If one has had the opportunity to look at this fiber, it all pretty much looks the same. As a result, Y uses an automated computer system to mark fiber in the lab with a code that designates its type. This system works very well unless data accuracy problems cause the fiber to be incorrectly labeled.

Earlier this year, Y was notified of a problem with a cable that contained fiber produced at this plant. This cable was under a lake in the state of Washington. Careful review of the problem led them to believe that some experimental fiber was in that cable. Y was forced to pay for the removal of the cable, replacement of the experimental fibers, rebundling of the cable, reinstallation of the cable. This added up to about $500,000 and is directly attributable to the fact that the data code on the fiber was incorrect. Certainly Y did everything it could to correct the problem, but we suspect the customer will think twice before purchasing Y's fibers in the future.

Operations Productivity

As mentioned previously, it is a widely held belief the 40% of the costs within any service organization are related to rework expenses. Since an IS organization is an excellent example of a service organization, there is good reason to believe that rework accounts for a significant percentage of the total budget. Data quality problems constitute a primary cause of rework in the IS department.

Example:

At Hospital E, the Admitting Department and Medical Records Department are especially concerned with data quality: Admitting enters the majority of a patient's registration information (e.g., patient demographics , admitting physician, insurer). Medical Records codes all records of patients discharged.

Both Admitting data and Medical Records data are used in a corporate level decision support system. Some automated filters and edit checks are used to examine this data as it is downloaded into the DSS. Data that fails the filters and edit checks is fed back to department that entered it for rework.

An employee in Admitting spends several hours each day maintaining the "Outside Referring Physician Directory" This directory keeps outside physicians informed of their patients who are admitted, documents new medical procedures at Hospital E, etc. It is important for generating hospital admissions. Using an *exception report* generated daily by the system, the employee tracks down all doctors entered yesterday as free text, checks existing lists, calls their offices to verify all the required information, updates the list, codes the name, updates medical records with that doctor, etc.

In addition to such "scheduled" data cleaning efforts, all personnel in Admitting and Medical Records are supposed to be responsible for making corrections as needed. Usually, these are obvious errors like misspellings or male admitted to Obstetrics. It is difficult to calculate how much time this responsibility requires, but it is not a trivial commitment.

Summary

Hopefully, the examples discussed above serve to illustrate the three areas where data quality impacts the bottom line. The following table summarizes this research and and the impact of data quality along the three parameters discussed above.

Business Issue	Goal	Role of Data Quality
Management Decisions	Make correct business decisions.	Correct decisions require data which is accurate, available, and interpretable.
Customer Service	Deliver a competitive level of service quality to the customer.	Competitive levels of service require customer account and product data which is accurate and available.
Operations Productivity	Run operations as efficiently and effectively as possible.	Efficient operations require accurate data in order to minimize rework.

Having identified and defined the impact of data quality on business operations, one goal of this research has been achieved: the need for managing the quality of corporate data has been firmly established. It should be clear from this section that data quality has a broad impact on the operating profitability of an organization. The scope of this impact will only increase as organizations become more and more automated.

5.2 Estimating the Costs of Data Quality

One goal of this research has been to estimate the impact, in dollar terms, which data quality improvements could have on an organization's bottom line. Toward this end, in this section we examine our twenty four survey participants and estimate the average value that data quality improvements could contribute to their bottom lines. Since the potential value of improved management decision making is nearly impossible to measure, we take into account only the savings which result from improvements in customer service and operations productivity.

Because the data gathering efforts undertaken for this research involved no direct cost measures of data quality, we are forced to draw conclusions from indirect measures and extrapolation. In this manner, this section should be regarded primarily as providing an indication of the significant costs attached to poor data quality management. The author regards the cost of data quality as an area where significant further research could and should be undertaken.

The principle hypothesis underlying this section can be stated as follows: In the manufacturing world, the potential for reducing quality related costs through proper management is greater than 80%[3]. By analogy, similar cost savings should be achievable in the information systems world. As mentioned many times in this thesis, a widely held belief states that 40% of IS costs result from rework[4]. A significant portion of this rework probably results from poor data quality management. Hence, an 80% reduction in data quality related costs translates into significant savings for IS organizations. The potential savings are large enough to free up substantial IS resources for important management, customer service, and revenue enhancing systems.

In an attempt to quantify the potential cost savings, we extrapolate from two bodies of research performed by authorities on quality in the manufacturing world: David A. Garvin and Philip B. Crosby.

Garvin's Estimates

[3] Garvin, D.A. 1983. "Quality on the Line." *Harvard Business Review*. Sept.-Oct. and Crosby, P.B. 1979. *Quality is Free.* New York: McGraw-Hill.

[4] Merrill Lynch, Information Systems Department, personal communication Oct. 1990.

In 1983, Garvin[5] published a thorough analysis of the comparative levels of manufacturing quality in the American and Japanese room air conditioner industries. In this research, the air conditioner manufacturers were divided into five groups based on quality. Garvin measured the total cost of quality (i.e., prevention costs, appraisal costs, and failure costs) as a percentage of sales for each company and published median values for each group. This data appears below.

Garvin

Total Cost of Quality as a Percentage of Sales				
Poorest US Plants	Average US Plants	Better US Plants	BestUS Plants	Japanese Plants
> 5.8%	3.9%	3.4%	2.8%	1.3%

Crosby's Estimates

In 1979, Crosby[6] published estimates of the total cost of quality as a percentage of sales across his five categories of quality management maturity. These categories represent descriptions of manufacturing companies at various stages of maturity in the quality management process. From least to most mature, Crosby calls these categories: uncertainty, awakening, enlightenment, wisdom, and certainty. For a more thorough explanation of these five stages of quality management refer to Appendix C. The cost of quality estimates for each stage appears below.

[5] Garvin, D.A. 1983. "Quality on the Line." *Harvard Business Review.* Sept.-Oct. pp. 65-75.

[6] Crosby, P.B. 1979. *Quality is Free.* New York: McGraw-Hill.

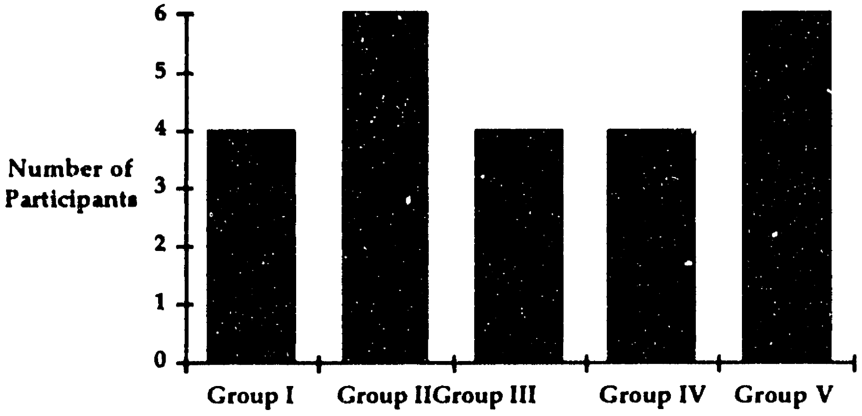
Crosby

Total Cost of Quality as a Percentage of Sales				
Uncertainty	Awakening	Enlightenment	Wisdom	Certainty
20%	18%	12%	8%	2.5%

Total Cost of Data Quality Estimates

Like Garvin and Crosby, we have subdivided our respondents into five categories of quality maturity based on the ratings which they provided in the surveys. To accomplish this, we averaged each company's rankings on the six survey questions estimating accuracy, interpretability, availability, timeliness, departmental data, and auditability. The results were scaled so that each company was given an overall quality rating ranging from 0 to 10. If a company rated itself perfectly in all six categories, it would achieve a score of 0. Likewise, rating itself as poorly as possible in all six categories would achieve a score of 10. Based on clustering of the ratings, we then defined five groups based on a score, s , in the following ranges: Group I $s < 1.5$; Group II $1.5 < s < 2.5$; Group III $2.5 < s < 3.5$; Group IV $3.5 < s < 4.5$; Group V $s > 4.5$. The distribution of scores is given below.

Distribution of Survey Participants Across Groups



As indicated above, our assumption, in using the Garvin and Crosby estimates, is that the room for improvement in total costs from one group to the next is the same for data quality as it is for manufacturing quality. For example, Garvin estimates that Japanese Plants enjoy a cost of quality which is three times lower than that of Average US Plants. In applying the Garvin estimates to our survey respondents, we assume that the companies in Group I enjoy a cost of quality which is three times lower than the companies in Group IV.

Extrapolation Based on the Garvin and Crosby Data

If the Garvin numbers are applied to the five groups categorizing our survey respondents, then the average amount by which a company in our survey could lower its total cost of data quality is 60%. If the Crosby numbers are applied to the five groups categorizing our survey respondents, then the average improvement factor which companies in our survey could realize becomes 80%. Clearly these extrapolations have not been sufficiently grounded in objective research. Nonetheless, they indicate that a substantial opportunity exists for lowering costs through data quality improvements.

Conclusions

Although the results presented in this section need to be placed on a more rigorous empirical footing, it is reasonable to assume that the calculations given above may serve as ballpark estimates for the savings which many IS organizations could achieve through better data quality management. Hence, long term goals for a data quality improvement program should include the significant costs savings which may be achievable from productivity improvements. Specifically, these numbers indicate that large organizations, such as Fortune 500 companies, could use data quality improvement programs to free up hundreds of millions of dollars for vital backlogged projects.

Part III

Achieving Zero Defect Data

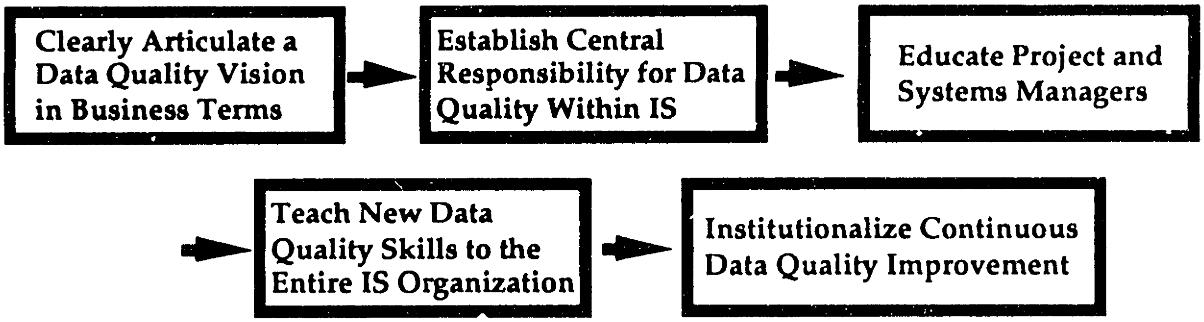
Chapter 6

Implementing a Data Quality Improvement Plan

In this chapter, attention is focussed on describing how organizations can implement a program to manage data quality. Section 6.1 applies some of the lessons learned from manufacturing quality management in the context of data quality. Specific recommendations are presented for implementing organizational and technological solutions to data quality problems. Section 6.2 goes on to establish operational goals, in the form of critical success factors, for the successful management of data quality within organizations.

6.1 Five Categories of Change

The art and science of data quality management remains in its infancy. As pointed out in Section 5.2, because most organizations have little or no formal systems for data quality management, the opportunities for improvement and resulting economic gain in this area are tremendous. Such improvement, however, cannot be achieved without significant organizational changes. In this section, we address the changes which need to be made. For convenience of analysis, we have grouped the changes into five categories which loosely parallel those of Tribus and Tsuda presented in Section 1.6. Our categories are based on impressions and data gathered from conversations with the companies listed in Appendix A. The resulting progression toward superior data quality management appears below.


Clearly Articulate a Data Quality Vision in Business Terms
Establish Central Responsibility for Data Quality Within IS
Educate Project and Systems Managers
Teach New Data Quality Skills to the Entire IS Organization
Institutionalize Continuous Data Quality Improvement

Clearly Articulate a Data Quality Vision in Business Terms

As discussed in Chapter 1, quality can be defined as conformance to standards. Hence, in order to improve quality, one must first set standards. At the highest levels, standards are set by users: the external and internal customers for the data produced by information systems. Such standards are expressed in business terms. In this manner, the first step toward implementing a data quality improvement plan is for top IS management to clearly articulate a data quality vision in business terms. The following example from the Bank X's 1990 Data Administration Task Force report illustrates this principle very well.

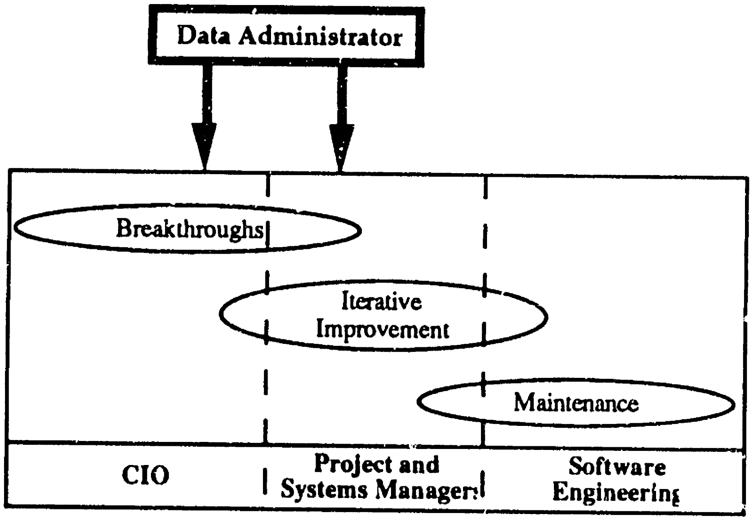
> *Customer service and decision making at the Bank X will be unconstrained by the availability, accessibility, or accuracy of data held in automated form on any strategic platform.*

Since leadership is crucial in the early stages of any quality improvement program, the data quality vision must be clearly identified with the top level management (e.g., the CIO) within information systems. At this stage, top management's goal is to begin organizational awareness of data quality problems and start everybody moving in the same direction. Toward this end, the CIO must make it clear to the entire organization that data quality has become a top priority.

Establish Central Responsibility for Data Quality Within IS

Once a vision has been articulated, the organization needs to establish central responsibility for data quality. Ultimately, this responsibility rests with the chief information officer, but another person, reporting directly to the CIO, needs to be given day to day responsibility for data quality. Many organizations are tempted to proclaim that data quality is "everybody's responsibility", but in practice this approach leads to confusion and inaction. Implementing a data quality improvement program requires significant organizational change as well as the adoption of new management techniques and technologies. For these reasons, we recommend that a position of data administrator[1] be established with primary responsibility for ensuring data quality.

Data administration is a managerial, rather than technical, function distinct from data base administration. The data administrator is responsible for making ;ure that data resources are managed to meet business needs. In this manner, data quality falls naturally within this sphere of responsibility. The data administrator should head up the Data Administration staff which serves as a center of expertise on the application of quality management within the information systems organization.



Allocation of Responsibility for Data Quality Improvements

The diagram abov(indicates that the data administrator has responsibilities spread equally across the two highest levels of quality management discussed in Chapter 1: breakthroughs and iterative improvements. In the area of breakthroughs, the data administrator coordinates work with the CIO and

[1] For more information regarding the role of the data administrator see Date, C.J. 1990. *An Introduction to Database Systems,* Addison-Wesley

senior level management to identify systems redesign projects and new technologies which could have tremendous impact on the organization's management of data quality. An example of such a breakthrough is the data warehousing project being developed and implemented at Bank X. This data administration initiative[2] will revolutionize management of corporate data resources in order to improve data quality.

In terms of iterative improvements, the data administrator serves as a central source of information and guidance which project and systems managers can access regarding data quality matters. In this manner, Data Administration should understand how to apply the principles of statistical quality control to information systems. Furthermore, the data administrator should be responsible for evaluating new development methodologies such as CASE and Data Modeling with respect to data quality on a project by project basis.[3] Finally, Data Administration should also be given responsibility for tracking the critical success factors of data quality management discussed in Section 6.2. In this instance, the data administrator would coordinate with the data base administrator to ensure that technologies and systems are being developed to ensure that the data quality goals established with respect to the critical success factors are met.

Educate Project and Systems Managers

[2] See Section 7.6 for more information on data warehouses.

[3] See Chapter 7 for more information on SQC, CASE, Data Modeling and other technologies which can be used to improve data quality.

Once central responsibility for data quality management has been established, the stage is set to begin educating the key people within the organization who will take charge of iterative improvements in data quality. Within IS, these people are the project and systems managers. These managers must learn the relationship between quality and productivity so that they will invest the time and resources necessary to improve data quality. Beyond this, they must learn the specific methods of data quality improvement that are relevant to their projects or systems. For project development managers, this means learning to view data quality as a fundamental design goal. For systems managers, it means learning to apply the principles of statistical quality control to monitor systems.

Of critical importance to these managers will be the job of identifying and implementing the appropriate quality control methods for their projects or systems. Toward this end, we have developed the following grid which segments information systems into four groups by type of application and type of data reported: factual management information systems, aggregate management information systems, factual data processing systems, and aggregate data processing systems.

		Type of Data Reported	
		Factual	Aggregate
Type of Application	Management Information	Hospital Patient Records	Marketing Analysis
	Data Processing	Airline Reservations	General Ledger

Factual Management Information Systems:

These systems provide individual facts which form the basis of management decisions. A good example is a hospital's patient information system which doctors and nurses use to manage patient care. In such systems the accuracy and timeliness of individual data elements is critical.

Aggregate Management Information Systems:

Data is analyzed in the aggregate by these systems in order to drive models or statistical analysis supporting management decisions. Marketing analysis models which consumer products companies run off their scanner data bases constitute one example of this type of system. Systematic accuracy errors or data availability problems are likely to comprise the most important data quality problems with respect to these systems.

Factual Data Processing Systems:

These systems are characterized by transaction processing of individual data elements. A good example is an airline reservation system. In such applications the control of systemic data accuracy errors introduced through software enhancements are likely to be the most problematic. Interpretability problems in the presentation of data to endusers may also pose problems (e.g., bank statements mailed to customers).

Aggregate Data Processing Systems:

Data is processed in the aggregate by these systems in order to produce reports or other products, such as address labels, which can be run in batch. Examples include the general ledger system or monthly sales reporting systems in a retail environment. Interpretability problems are fundamental to this category. The temptation with automated reports is to always produce too much information and frequently systems are not designed to provide data in a readily useable form. Systematic accuracy errors are also a primary concern here. For these systems, a moderate level of random data accuracy errors may be acceptable. For example, a direct mail marketing company would not be very concerned if 3% of its addresses were incorrect. In contrast, 3% accuracy errors is totally unacceptable for an airline reservation system.

Teach New Data Quality Skills to the Entire IS Organization

Management holds most of the responsibility for ensuring data quality. The reason for this is that most data quality problems result from poor systems design and administration: areas which are management's responsibility. However, responsibility for the successful implementation and maintenance of data quality programs belongs to the entire organization. Hence, the entire IS organization must learn the skills required to put data quality improvement programs into place.

The skills required by an individual will vary according to his or her responsibilities. In general, data quality responsibilities will fall into one or more of the following three categories: inspection and data entry, process control, and systems design. Knowledge of statistical quality control is essential for work in all three areas and therefore SQC techniques must be universally understood throughout the IS organization. SQC in the context of information systems is discussed in greater depth in Section 7.2. Below we discuss the three categories of data quality responsibility and the relevant skills required for each.

Inspection and Data Entry:

Inspection and data entry involves responsibility for the accuracy of data as it is entered into a system or is processed by a system. Current practice for the inspection of data remains mostly manual. However, since manual inspection of data is totally inadequate for superior data quality management,[4] in the future the use of advanced expert systems based data quality measurement tools may become mandatory. This will necessitate the training of workers in the proper use of such tools. In terms of technology that is here today, modern interactive and forms based user interfaces require some training in order to fulfill their potential for minimizing data entry accuracy problems.

Process Control:

Process control involves maintaining and monitoring the performance of existing systems with respect to data quality management. In addition to SQC, the training required here involves the use of auditability tools for tracking down the source of data quality problems. In our survey, over 50% of respondents expressed difficulty in tracking down the sources of data quality problems. Proper training in the tools and methods required to isolate problems could have a significant impact on data quality management. Finally, our conversations with IS organizations indicate that people with process control responsibilities frequently need training in the proper procedures for the uploading and downloading of data. The transferring of

[4] See Section 7.1.

data from one database to another is a significant source of quality problems, especially when one of the databases resides on a PC.

Systems Design:

Finally, systems design involves building new systems or upgrading existing applications with data quality management as a primary design goal. In this area there are a host of tools and techniques which professional IS developers need to learn in order to design systems which are compatible with data quality goals. These are discussed in Chapter 7 and include the use of: CASE tools, data modeling, intelligent user interface design, data warehouses, and auditability tools.

Institutionalize Continuous Data Quality Improvement

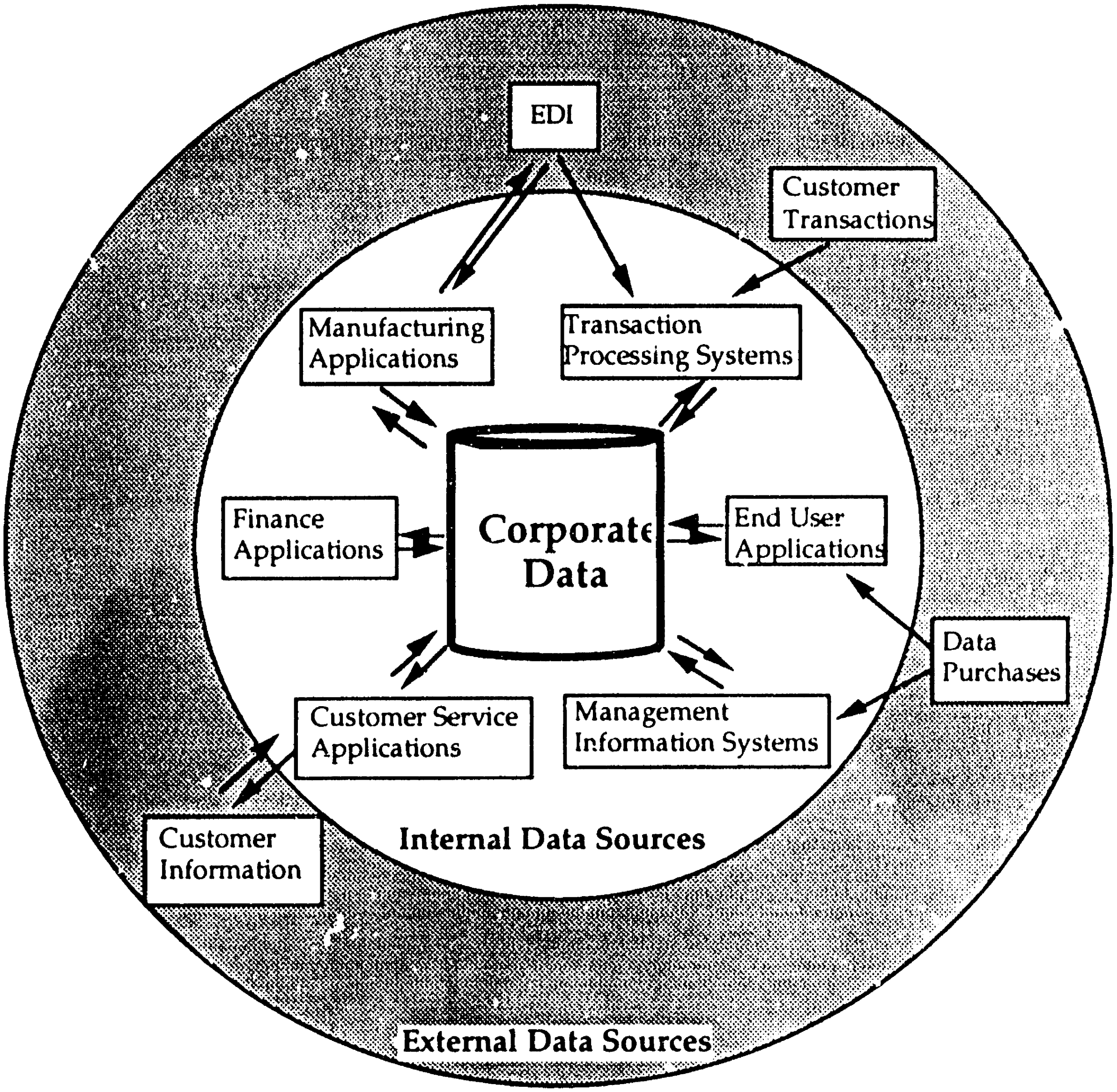
Once the entire organization has received the necessary training, and data quality improvement plans have been put into action, it is necessary for top management to ensure that the data quality improvement process becomes institutionalized. This requires leadership from the CIO and other top management in the form of visible continuous interest in data quality activities. For example, regular meetings, presentations, and reporting structures should be established to track the organization's progress in meeting data quality goals. Additionally, data quality improvement projects need to become a regular part of the budgetary process. Data quality programs cannot be put in one quarter when money is available, and taken out the next when finances get tight.

6.2 Critical Success Factors for Achieving Zero Defect Data

In the preceding section, broad categories of organizational change were discu3sed. In this section more specific operational goals for achieving zero defect data are presented. These goals are expressed as *critical success factors* : operational objectives which are absolutely critical for the successful management of data quality. Based on interviews and surveys, five critical success factors have been identified.

- Certify Existing Corporate Data
- Standardize Data Definitions
- Certify External Sources of Data
- Control Internal Generation of Data
- Provide Data Auditability

Below is a diagram which can help in understanding which systems and data sources the five critical success factors impact. In this example, the goal of the information systems organization is to ensure that the corporate level data exhibits superior quality across all four parameters: accuracy, availability, interpretability, and timeliness.



Certifying the existing data implies providing a guarantee that the corporate data, pictured in the center, is 100% accurate. Standardizing data definitions ensures that all data flows, indicated by the arrows, among internal data sources can be implemented in a straightforward manner. The result is a high degree of availability for corporate data. Certifying external sources of data involves ensuring that none of the sources pictured in the outer ring are contributing accuracy errors to the corporate data. Likewise, controlling internal data generation implies certifying all of the applications pictured in the inner circle, and their interfaces with the corporate data. Finally, providing data auditability implies that when data quaiity problems are detected in the corporate data, they can be traced to the source, whether it be internal or external.

Certify Existing Corporate Data

Most organizations do not have a clear understanding of the quality of their existing data. Frequently, this data is worse than they think, because much of it entered their systems long before any serious attempts were being made at data quality control. For example, one insurance company related a story of a mailing which they conducted recently to everyone on their central customer database. Over fifty percent of the letters were returned because addresses were incorrect. To a large extent, this is a problem of timeliness. Much of the address data was just too old. However, for many organizations, insurance companies particularly, the customer database is a strategic resource which should not be allowed to deteriorate to such a poor state of quality.

Clearly, in order to begin reaping the benefits of high quality data, in terms of decision making, customer service, and operations, organizations must start by certifying their existing data. However, this is easier said than done. Below, a brief sketch of how organizations are approaching this problem, is given.

Current Practice:

Most certification today is done manually. A good example of this appears above in the discussion of the data rework activities conducted at Hospital E. Typically, a small number of people are assign' ! half time responsibility to sort through files and records correcting mistakes. Sometimes, more sophisticated and automated means are applied. As data is used, it may be passed through some simple rule based checking. When problems are detected, the data may be fed back to the source for verification or correction.

Most organizations make some attempt to certify data after any migration process: when data moves from one system to another and changes format. The most basic certification consists of checking to see that the migration process did not degrade the quality of the data. This process is frequently given an official sounding name like "acceptance testing", but usually just involves running all of the old reports on the new systems and checking that they appear the same as before and don't crash. More motivated organizations may use the data migration time as an opportunity to try to upgrade the quality of their data. Again, however, this will mostly be approached by human verification. Various new reports will be run, and the output will be fed back to users or data entry people to check for errors.

Example:
A few, leading edge, organizations use expert systems technology to certify the quality of their existing data. In particular, Information Resources, Inc. (IRI) has invested millions of dollars in software for automated checking of data quality. IRI is in the business of buying scanner data from grocery stores, processing it, and reselling it to consumer marketing firms. Their data quality systems use the latest artificial intelligence and expert systems techniques to verify and automatically correct problems with incoming scanner data. Since IRI is in the business of selling data, their customers let them know quickly when any problems occur. In this case, there is a very direct link between data

quality and the company's bottom line, so management does not hesitate to invest money in solutions.

Standardize Data Definitions

The concept of standardizing data definitions really encompasses two sets of issues. One set is technological and refers to data incompatibilities resulting from incompatible technology standards (e.g., IMS and DB2). The other set of issues is administrative and refers to data incomparabilities resulting from inconsistent definitions of business terms (e.g., Corning's operating margins). Both of these sets of issues are difficult to address, although the first set is becoming easier as technological solutions based around data modeling tools are beginning to bear fruit in this area.

Current Practice:

Most organizations today do not have any data definition standards in place. Almost every organization participating in this research indicated that the have plans for standardizing their data definitions. Usually, these plans involve adopting a specific data modeling tool which will be used by data administrators throughout the organization in a manner consistent with standards set forth by the central IS people. Plans for moving the applications development cycle to a standardized, repository based systems, are usually mentioned along with the goal of data standardization.

Currently, however, the data in most organizations consists of a hodge podge of incompatible standards stored within various incompatible database systems. It is unlikely that this situation will change at any great rate in the near future because most organizations have concluded that the costs of migrating data from old (IMS or IDMS) database systems to new (DB2) technology, are not worth the benefits. Certainly, all new applications are being developed according to the new standards, but the old applications and data will probably be left alone for a long time to come.

Example:

Take Health Maintenance Organization G as an example of an company which is wrestling with the problems of data definition standards. In this case, the problem is with the precise definition of business terms rather than with incompatible technology standards. For business efficiency purposes, the organization has developed along decentralized lines. Historically, HMO G existed as separate business entities in several regions (e.g., Braintree, Methuen, Peabody) which have now consolidated into 8 health centers. Each center has developed and maintained its own rules, policies, procedures and data standards. Consider the definitior of "encounter" (visit to health center) and the procedures involved:

In Methuen, if a family of 5 (mom, dad, 3 kids) came in for checkups, they would be charged $3 each = $15 and reimbursed for the rest. They fill out 5 encounter forms. In Peabody, the same family also fills out 5 encounter forms, but they are charged for the kids as a group: total = $9. In Braintree, the same family only fills out 1 form, and is charged $3 for "the sick one". As a result of the different definition of "encounter", corporate level utilization reports showed that Methuen had 3 times as many encounters as Braintree, even though Methuen has 10,000 members and Braintree has 20,000. Since part of the reimbursements are made according to encounter forms, this inconsistency clearly affects the member, the health center, and Blue Cross.

Certify External Sources of Data

As data collection technology has improved, a market has developed for the resale of data by third parties. This market is the best developed within the financial services and consumer marketing industries. Firms such as Dun & Bradstreet, Information Resources, and Nielsen collect scanner data from retail stores, process the data, and resell it to large consumer marketing firms such as Proctor & Gamble or General Mills. In the financial markets, firms such as Reuters, Dow Jones, and Compuserve perform a similar economic function. As the practice of purchasing third party data has become commonplace, organizations have been forced to consider methods for insuring the quality of this data.

Another external source of data which is becoming increasingly important is electronic data interchange data (EDI). This is an electronic network by which customers can communicate orders to suppliers and suppliers can send invoices to customers. Anecdotal evidence suggests that both suppliers and customers must be careful to check the quality of data received via EDI.

Current Practice:

Most organizations take the same approach to insuring the quality of third party data: they assume it is the suppliers responsibility. Typically, this means proceeding with operations under the assumption that the data is accurate. When something goes wrong, IS departments complain to the supplier and force him to fix the problem. Many firms expressed the goal of being able to work with only one data supplier. In this manner, they hope to build a better relationship with that supplier which would give them greater leverage over the quality of his data. This strategy is very similar to the the single sourcing philosophy which is now very popular in the manufacturing world. The assumption is that higher quality levels can be achieved by building strong working relationships between buyers and suppliers.

Example:

Manufacturer H has developed extensive relationships with customers and suppliers that involve the transmission of data using electronic data interchange (EDI). They find that the vast majority of the data they get in has significant quality problems. So significant that they find it impossible to integrate this data into their manufacturing process without extensive human involvement. Manufacturer H believes strongly that the responsibility for ensuring the quality of this data lies with the source, and is working with their customers to clean up their EDI data.

Control Internal Generation of Data

Traditionally, most corporate data has been generated internally. This includes sales data, customer data, financial records, etc. All of this data must have been originally entered into a machine by an employee through some particular user interface. This suggests the importance of well designed data entry procedures and user interfaces to insure data quality.

It is now well understood that the most economic place to solve data quality problems is at the source. Within the manufacturing world, practical rules of thumb suggest that a quality problem which costs $1 to fix at the source, costs $10 to fix in production, and $100 to fix in the field. Within the information systems world, it is probably reasonable to assume that a data quality problem which would cost $1 to fix at data entry, would cost $10 to fix during beta test, and $100 to fix once the full scale production system is running.

Current Practice:

Although the point of data entry is the most convenient and economic place to begin a data quality program, surprisingly few organizations have given much thought to how their systems interfaces impact the quality of their data. Forms-based interfaces with automated edits of fields go a long way toward improving the quality of data. Automated edit checks are able to apply a few simple rules to fields which ensure the avoidance of egregious errors (e.g., age: 165). Context-based help has also had an impact. This technology allows users to interact with the interface to determine whether or not they are entering data correctly. An old stand-by technique from the days of batch systems, double entry, is still used effectively to cut down on data entry errors. A few organizations have even implemented rule-base systems which go a step beyond automated edit checks to apply complicated rules to data in more that one field (e.g., if sex=male then not condition=pregnant). However, as people are infinitely creative in their efforts to enter bad data into a machine, these technologies by no means guarantee data quality. The only real solution to this problem may be educating people to the importance of accurate data and pushing down responsibility for that quality within the organization.

Example:

Many organizations are currently rewriting their user interfaces in order to make them more conducive to data quality. Bank I is currently developing a sophisticated user interface for their traders in an effort to cut down on the number of incorrectly executed trades. In this regard, they have many of the same problems discussed previously in terms of Bank D's error account. This will be a forms-based interface with automated edit checking on the front end, and a rule-based system on the back end.

However elaborate the user interface, there will still be mechanisms to defeat it, as these examples show. Hospital E related a story of a registration clerk who entered her own zip code, instead of the patient's, whenever the patient was unable to provide it or it was blank on the form. The automated edit checker never caught on. Doctors are notorious for defeating user interfaces. The same hospital reports that if their doctors can't remember a diagnosis code, they will make one up, or enter another that they know. The solution to this second problem, of course, is to make the diagnosis code transparent to the user, and prompt doctors for diagnoses by name.

Provide Data Auditability

Among the critical success factors for zero defect data, the issue of data auditability has received the least attention in the field. When asked a direct question, however, over fifty percent of the survey participants indicated that data quality problems were at least "somewhat difficult" to track down and correct. In large part this is probably the case because none of them have any technology or formal systems in place to audit their data sources. Many IS managers also note that data quality problems are increasingly difficult to trace, the further downstream from the source they show up.

Current Practice:

Current practice for tracing data quality problems is only automated in one area: systems processing induced quality problems. If data accuracy is being corrupted by a malfunctioning system, there are debugging tools and systems documentation which can aid the IS staff in tracking down the source of the errors. However, this assumes that the problem is detected close enough to the source, that it is obvious which system is corrupting the data. Another tracking method which a few IS managers mentioned involves using the audit trail generated by a transaction processing system. After some careful analysis, contaminated data can frequently be matched with the transactions which caused the problem.

Data quality problems which are not caused by processing errors can be harder and more costly to trace. Most participants in this research indicated that tracing such problems involved manually working back through the systems to the data source, like a detective, by making lots of phone calls and studying standard IS departmental reports.

Example:

When HMOs or insurance companies discover data quality problems with patient information, they call hospitals to check their data against hospital records. This is a time consuming process which may end up correcting the data quality problem without discovering the source of the error and preventing it from happening again.

6.3 Obstacles to Achieving Zero Defect Data

Return on Investment Analysis

Data quality improvement programs can often be difficult to justify because the benefits are not yet well understood. Even in manufacturing, where the principle that quality improvements imply productivity improvement is well understood, investments in quality are often hard to justify on an ROI basis. In the data center, where it is extremely difficult to estimate the true cost of data quality, ROI justifications are even more difficult. In this manner, research needs to be undertaken which would establish a direct and measurable link between data quality in productivity in the information systems department.

Management Blaming Everyone but Themselves

Data quality is primarily the responsibility of IS management, because management designs the systems which capture and process data. Many managers, however, do not accept this responsibility, and prefer to blame data entry clerks, programmers, the quality assurance department, or anybody except themselves for data quality problems. Such behavior defeats the primary goal of a quality oriented organization: to iteratively improve IS processes through the cooperation of all parties in the identification and correction of sources of data quality problems. Behavior which blames one group or another for quality problems can only be counter productive.

One example of demeaning management behavioɪ comes from a manufacturing company participating in this research. Manufacturer Y produces a variety of products which are physically indistinguishable, but have different operating characteristics. Products are automatically coded before shipment to make sure that mixups do not occur. At one point, however, a routine data quality error caused a shipment to be miscoded. The mistake cost the company over $500,000 to fix. As a result, a division manager at Y placed the miscoded product on display in a company lobby to remind the workers how important data quality can be. To our knowledge, no effort was made to determine the source of the error, or even to determine if the workers could have prevented the problem from occurring. In a quality oriented company, management takes responsibility for designing systems which improve data quality and cooperates with the workers to implement such systems. Prominently displayed mistakes which implicitly blame careless workers do little to achieve the goals of a quality oriented company.

Direct Measures of Data Quality are Difficult

Another obstacle to data quality improvement involves the difficulty of employing direct measures of data quality. In a manufacturing setting, defects are frequently easy to define and inspect. In a database, the variety of accuracy errors which can occur is, for all practical purposes, infinite. In this manner, data quality management involves determining which variety of data quality errors are most likely and then implementing means of inspecting for these. Such data quality checking is further complicated by the fact that human inspection is totally inadequate for the job. The databases employed today are simply too large to be manually inspected. Unfortunately, human judgement and creativity are often the ingredients necessary to discover and correct data accuracy errors. To address this problem, new technology needs to be developed which can automate the inspection and defect resolution process. Expert systems will probably play a role in automating these tasks. In fact, some leading organizations are already developing expert systems based inspection systems[5] .

Lack of Awareness of the Problem

A final obstacle to data quality improvement involves the simple lack of awareness of the severity of the problem. Most IS managers take a fire fighting approach to data quality. If the problem becomes so bad that customers or user begin complaining about it, then try to resolve it. Widespread human inspection of data is the hallmark of an organization employing this approach to data quality. When a data quality problem breaks out, the fire fighting organization will assign people to inspect the data coming out a system and ask them to correct the problems manually. Needless to say, this is a terribly expensive and inefficient approach to improving data quality. Even more importantly, widespread inspection contributes nothing to productivity through better systems design. Organizations which are aware that data quality is an important problem should take a proactive role in employing the management techniques described earlier in this chapter.

An excellent example of this lack of awareness is offered by Company Z which participated in the early phases of this research. A number of managers

[5] See Section 7.7.

interviewed at Z insisted that the company has no significant data quality problems. This assertions came to us in spite of well documented newspaper and magazine reports describing widespread examples of incorrect customer billings.

Chapter 7

Technologies for Improving Data Quality

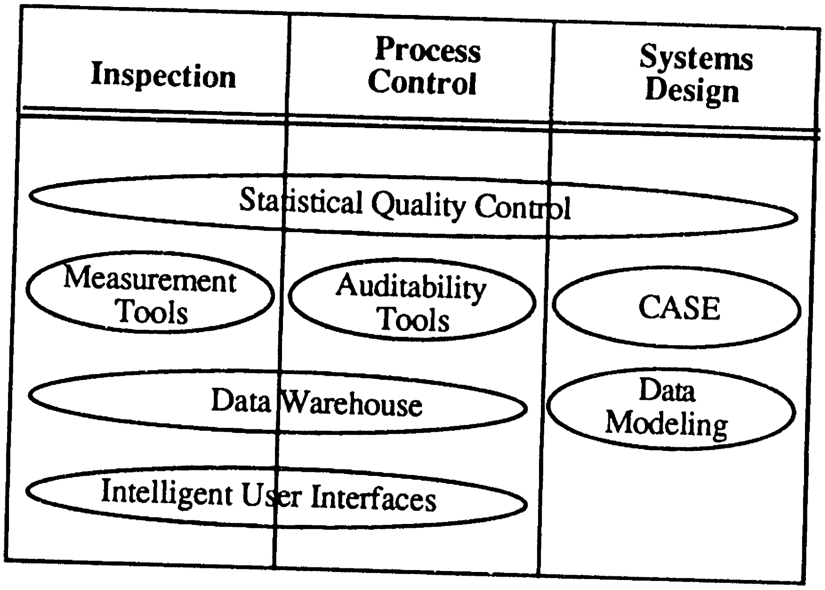
As we have seen, managing data quality constitutes a complex organizational and operational challenge. Organizationally, we have identified five categories where progressive changes are needed in order to achieve superior data quality management. Operationally, five critical success factors have been established for achieving zero defect data. In this chapter, recent technological advances are examined which could help organizations manage data quality along the lines described in Chapter 6. These technologies fall into the following seven groups: statistical quality control, CASE tools, data modeling and repositories, intelligent user interfaces, data warehouses, expert systems based measurement tools, and auditability tools. This chapter offers the reader a very brief introduction to these seven technology groups and their potential impact on the organizational and operational aspects of data quality management[6].

7.1 The Role of Technology in Achieving Zero Defect Data

The role of technology is an area in which data quality management differs significantly from manufacturing quality management. In manufacturing technology, and robotics in particular, have been discredited as an effective means of improving quality management. In the manufacturing arena, quality and productivity have been linked primarily to improvements in product and process design. In the world of information systems, however, improvements in product and process design are intimately linked to technology. CASE tools, data modeling and repositories, intelligent user interfaces, data warehouses, expert systems based measurement tools, and auditability tools all represent technological tools which can be critically important to improving information systems design. Finally, data quality management also suffers from the problem that defects are not always as easy to measure as they are in the manufacturing world. An insurance company's 500 million row claims database cannot be manually inspected for quality problems. In this manner, if inspection and statistical quality control are to play a significant role in data quality management, new technologies are going to be needed to facilitate the quality measurement process.

[6] The potential applications of each of these seven technologies represent an important area where further research needs to be undertaken.

To set the context for our discussion of technology's role, however, we first offer a taxonomy of the seven technologies presented in this chapter along three categories of data quality management techniques. These categories are inspection, process control, and systems design. Inspection involves measuring the data for accuracy errors. Process control involves monitoring and managing software processes to prevent data quality errors from being introduced and to trace quality problems to their source. Systems design requires building and maintaining systems to ensure data quality along all four parameters. Below appears a table which indicates which of the technologies impact which of the three categories. This is a useful diagram to keep in mind while reading the following seven sections.



Impact of Various Technologies on the Three Categories of Data Quality Management Techniques

7.2 Statistical Quality Control

As discussed in Section 1.4, statistical quality control provides manufacturing organizations with a common language for the measurement and interpretation of quality parameters. SQC constitutes the primary means for evaluating manufacturing processes, identifying quality problems, and identifying solutions. Surprisingly, the principles of statistical quality control have not been adopted by information systems organizations to perform similar functions. In this section, we indicate how SQC could be employed in the data center to realize significant improvements in data quality management. As in a manufacturing environment, SQC should be applied to data quality management in five principles areas: situational analysis, process parameterization, process control, cause and effect analysis, and acceptance or rejection. For each area, we describe the goals, implementation, and deliverables related to statistical quality control.

Situational Analysis

The goal of situational analysis is to discover the principle data quality problems which a particular information system experiences. It provides a means of determining which problems are most important so that they can be addressed first. To implement situational analysis with respect to a given system, the following steps should be taken:

- Canvass the users and IS staff to determine which types of data quality problems they have noticed. Group these into categories for further analysis.

- Measure the frequency of errors for each category.

- Create a pareto chart (see Appendix D) of the error categories to determine which data quality problems are most pervasive.

The deliverable, at the end of situational analysis, consists of a pareto classification of the most common data quality problems for a given system. For example, a hospital may determine that incorrect zip codes or patient diagnoses codes are the most common errors encountered by users of a patient data base. An airline may learn that the most common problem experienced by users of a reservation system is that it misrepresents the number of available seats in one particular fare class.

Process Parameterization

The goal of process parameterization is to characterize the normal operating properties of each system and develop measurement techniques to determine whether it is performing normally. Implementation involves these steps:

- Target the most important quality parameters discovered during situational analysis. For each, statistically characterize and measure the normal operating conditions for the system. For example, with respect to diagnosis codes in the hospital patient database, determine the expected distribution of diagnoses which enter the system during a given hour, day , or week. This can be represented as a histogram (see Appendix D).

- Devise statistical tests for unlikely events. For example, an unusually large number of the same diagnoses in one day, or a particular diagnosis code occurring only at specific times of the day. These statistical tests should be designed to reveal the occurrence of the quality problems isolated during situational analysis as being the most important.

For process parameterization, the deliverable is a statistical profile (histogram or distribution chart) describing normal data along the critical parameters identified during situational analysis. Such a statistical profile acts as a surrogate for direct verification of data accuracy and allows the information systems organization to measure data quality without having to manually validate data.

Process Control

The goal of process control is to use the statistical profiles developed during process parameterization to continuously monitor systems for data quality problems. Implementation involves the following steps:

- Measure systems with respect to the statistical profiles (e.g., zip code or diagnosis distributions) developed during process parameterization. Wherever possible, automate the statistical tests which were developed so that systems parameters can be checked continuously.

- Have the automated statistical testing software flag IS staff when violations occur signifying unlikely systems events. For example, if a hospital patient database receives 4 times the normal number of a particular diagnosis codes within a half hour interval a flag would warn systems staff of probable quality problems occurring with the diagnosis code data.

The deliverable in this case is an automated data quality checking system. This system continually monitors applications with respect to key process parameters and warns IS staff whenever problems may be occurring. For example, if American Airlines[7] had such an automated data quality checking system in place to monitor seat availability during the summer of 1988, they could have avoided costly underbooking of flights. Such a data quality system could monitor the level of seat availability and compare with the typical distribution for specific routes and times of day. In this manner, after the faulty software enhancement, the data quality checking system would have immediately picked up on the drop in seat availability and flagged the IS department to look into the problem.

Cause and Effect Analysis

Cause and effect analysis addresses the goal of determining the cause of data quality problems uncovered as a result of process control. Here, we are interested in interpreting the statistically unlikely events which are flagged to the IS staff. Implementation involves two steps:

- Develop process flow diagrams describing all applications software systems which are being monitored.

- Construct Ishikawa diagrams (see Appendix D) which illustrate the causes of the most common data quality problems.

[7] See the Introduction for a description of the American Airlines problem.

In this case, the deliverable of cause and effect analysis consists of a set of diagrams which can be used by IS staff to track down data quality problems. For example, a hospital patient database system may flag IS that an unusually large number of a particular diagnosis occurred between 9:00 and 10:00 on June 31. Process flow diagrams would indicate that data entry points for the particular system are doctors working in a particular set of departments. An Ishikawa diagram may indicate that one of the causes of unusually large numbers of the same diagnosis involves doctors entering diagnosis codes incorrectly. Putting these pieces of information together, IS should be able to track down the doctor who is entering inaccurate data and work with him to improve the situation.

Acceptance or Rejection

The goal of acceptance of rejection analysis involves maintaining a certified set of data. For example, if IS wants to guarantee that a particular database meets certain accuracy standards, then acceptance tests need to be developed and applied to data before it is incorporated into this database. Implementation involves two steps:

- As in process parameterization, statistical tests should be devised which can be used to evaluate incoming data. This involves developing a statistical profile describing what normal data looks like and then defining the statistically unlikely events which indicated data quality problems.

- All channels from which data might enter the certified database must be screened using the statistical test. If data fails the tests, it should not be permitted to enter the database.

For acceptance and rejection testing, the deliverable is a plan to implement statistical tests at all points of data entry for a specific system or database. Such testing will become increasingly important as IS organizations continue to purchase more and more data from third party sources. This is particularly relevant with respect to EDI, where electronic orders can set purchasing, shipping, and accounting systems into action with little or no human intervention.

7.3 CASE Tools: Standardizing Applications Development

CASE Tools address data quality problems primarily by reducing the likelihood of applications design flaws. Applications design flaws impair data quality along two of the four defining parameters: accuracy and interpretability. Poor data accuracy can result when a flawed application is run on existing high quality data and systematic errors are introduced. Data interpretability is impaired when applications output (e.g., reports or on-line information) is poorly designed and users cannot easily extract or understand the information in that output which is relevant to them.

Information technology organizations which have large, complicated, and old applications are particularly susceptible to data accuracy problems introduced by systems design flaws. Problems typically occur when an existing application is altered or partially redesigned to address new business conditions. The changes introduced frequently have unanticipated side effects which can cause data contamination. Such side effects are difficult to prevent because many older applications were developed without regard for maintainability, and as a result do not have proper documentation or modular design. CASE (computer aided software engineering) tools comprise a technological means of addressing some of these problems.

CASE Tools and AD/Cycle

CASE tools allow information technology organizations to standardize and automate the various phases of the software design and delivery cycle. This cycle can be conceived of as roughly containing five phases: requirements definition, analysis/design, production, build/test, and maintenance. Independent CASE tools vendors have delivered products which automate various portions of this cycle. Some, such as Texas Instruments, offer integrated CASE tools which span the entire cycle. AD/Cycle[8] is IBM's proposed integration of its own CASE tools with those of some business partners to standardize and automate the entire cycle.

As illustrated in the following diagram, CASE tools address both the logical and physical design of both data and processes. This section considers primarily the process design. Data design is discussed in the following section.

[8] IBM Corp. 1989. *Getting started–planning for IBM's AD/Cycle,* marketing brochure.

	Logical	Physical
Data	Data Modeling	Database Design & Admin.
Process	Process Modeling	Systems Design

source: Gartner Group

The primary benefits, in term of data quality, to be derived from CASE based process design are:

- Process modeling improves applications maintenance by revealing the side effects of systems changes.

- Automated code generation from the process model into the systems design phase greatly reduces the number of programmer introduced errors which can contaminate data.

Process Modeling

A process model is a blueprint which represents the functionality of a program. It graphically displays the interrelationships between the various modules and sub-modules within an application. In an integrated CASE approach, the process model is directly linked to the applications code and serves as a pictorial form of documentation. Hence, when the applications needs to be updated, the side effects of any changes can be anticipated by examining the process model. This greatly reduces the likelihood of data being contaminated as a side effect of an applications update.

Automated Code Generation

Integrated CASE tools sometimes provide for automated code generation (or pseudo-code generation) directly from the process model. This greatly reduces the likelihood of applications design flaws being introduced during the code building phase. As a result, data is less likely to become contaminated because of poorly designed applications.

From a data quality perspective, perhaps the most valuable contribution of CASE is that the methodology provides IS organizations with a detailed blueprint of their systems architecture. In the world of construction, no builder would dare to make structural modifications to an existing building without an accurate and detailed blueprint. The consequences for the structural integrity of the building could be disastrous. Likewise, no software engineer should undertake significant modifications of existing systems without and accurate CASE diagram to serve as a blueprint to guide his efforts.

Example: American Airlines

During the summer of 1988, a software enhancement in American Airlines' Sabre reservation system resulted in inaccurate data concerning the number of discounted seats available on American flights.[9] As a result, the inaccurate data presented to travel agents caused underbooking of American flights for

[9] Computerworld, Sept. 19, 1988, pp. 2, "Airline hurt by faulty fare estimations."

an estimated $50 million in lost revenue. Large, complicated, and mission critical systems such as Sabre represent the applications where CASE offers the most potential benefit. In this ca᷉e, an accurate process model of the Sabre module which was being enhanced could have prevented the IS staff from introducing the flaw which corrupted the reservation data.

7.4 Data Modeling[10] and Repositories[11]

Perhaps the most universally acknowledged data quality problem among the organizations participating in this study concerns the lack of data availability stemming from incompatible data standards across systems. This is partially a result of the fact that most information systems organizations focussed on applications design standards long before data design standards. In fact, until recently, with CASE tools and relational database technology forcing the issue, most organizations hadn't given much thought to data design at all.

Data design is defined to be the logical structure of the mechanisms which are used to store the data. (e.g., tables within a relational system, data structures within a flat-file system) Data design flaws can cause data quality problems along two of the four defining parameters: availability and interpretability. Data availability is impaired when an organization uses incompatible data definitions across different systems. When it becomes necessary to build an application which can access data from both systems, the incompatible data

[10] For an introduction to data modeling see Chen, P.P.S. 1976. *The Entity-Relationship Approach to Information Modeling and Analysis*. North-Holland. Amsterdam.

[11] For an introduction to repositories see IBM Corp. 1989. *AD/Cycle: Blueprint for a more productive future,* marketing brochure.

definitions make it very difficult to design and deliver such a system. Data interpretability is hindered when poor data design prevents the extraction of information from the database in a convenient and usable form. For example, an employee database may have been designed without any provision for querying for the manager relation. Perhaps the manager relation is only indirectly stored in the database. In order to decide who Joe Smith's manager is, you might have to first query for Joe's department and then search all of the employees in that department to determine which one has the title "department manager". In this case, it is very difficult to interpret the data in a useful manner because of poor data design.

In this manner, the advent of data modeling tools, and their extension, through CASE, to data repositories impacts data quality management in two important ways.

- Data modeling tools facilitate thoughtful data design with respect to business needs.

- Repositories enable data definition standards to be enforced throughout an organization.

Most organizations which participated in this research were beginning to use data modelling tools and planned to implement a repository at some point in the next five years.

Data Modeling

Data modeling tools are concerned with logical database design. A logical database design (e.g., an entity relationship diagram) formalizes business relationships for representation in a database. For any given business system there are innumerable ways to draw up the logical database design. One important consequence of data modeling tools is that they force database designers to work within the restrictions of a particular methodology. By reducing the number of alternatives for logical design, data modeling tools naturally lead to organizational standards for representing data.

More importantly, these tools provide a medium in which such data standards can be represented. Prior to data modeling, designers undertook physical database design directly, without much consideration for logical design. As a result, standards had to be expressed in English, or pseudo-code, were difficult to implement, and remained largely ignored. Now that these tools provide a means for expressing logical database designs, software engineers can more easily design for standards.

Repositories

A repository is the centerpiece of any integrated CASE software development strategy. (e.g., IBM's AD/Cycle) The repository is a database management system which stores the deliverables for each phase of the logical and physical design of data and processes. The repository stores process-flow diagrams, entity-relationship diagrams, database designs, compiled code, etc. Repositories will be important to data quality management because they will provide a central focal point for the database design process throughout the organization. In short, a repository will be able to store and enforce the data definition standards which an organization wishes to impose on its software developers. Once all corporate and departmental applications conform to the data definition standards set forth in the repository, integration between database and applications will be facilitated and data availability will improve dramatically.

Example: Bank X

As mentioned in Section 3.2, The Bank X offers a classic example of an IS organization suffering from a lack of data standards. In order for IS to build a real estate portfolio system which can monitor the financial status of the bank's loans, data needs to be accessed from the commercial loan system (when lending information resides) and the real estate appraisal system (where current asset valuations reside). Because these two systems use incompatible data definitions it has been very difficult and time consuming to build an application which can access data from both. In the future, the bank hopes to avoid such problems through the use of data modeling tools from vendors such as Bachman. If the commercial loan system database and the real estate appraisal system database both had accurate data models, then IS developers would realize two tremendous technical benefits. First, loans and valuations would be stored in standardized, compatible formats. Second, applications developers would have accurate blueprints of how all relevant business data is stored in both systems. Both of these benefits would serve to increase applications development productivity tremendously.

7.5 Intelligent User Interfaces

The user interface is the point of entry for the vast majority of data accuracy errors. A surprising number of these errors could be eliminated simply by employing better user interface designs. In manufacturing, the term "design for manufacture" refers to the process of designing products in a manner that will render them particularly suitable to high quality production. Likewise, in the data center, user interfaces should be "designed for data quality". The following example will serve to illustrate how interfaces can be designed to encourage quality data entry. At one Boston area hospital, doctors are required to enter patient information into a central database. The user interface requires the entry of an obscure "diagnosis code" for each patient. Since the doctors often cannot remember this code, they make one up, or enter an incorrect one which they know. To avoid this problem, the user interface should provide some facility, such as a help screen or choice list, which would allow the doctor to select the diagnosis from a predefined list. Simple design principle like the one illustrated by this example could go a long way toward improving the quality of many organizations' databases.

Moving beyond more user-friendly designs, there are several technologies which can be employed in order to make an interface more intelligent and therefore less likely to pass bad data. One of these technologies, which is becoming widely accepted, involves automated edit checks. In this case, an interface will automatically apply some straightforward range checks to data before it accepts the entry. Examples involve checking that dates are valid, checking that data falls into a specific range, ensuring that zipcodes match the city and state, etc. More sophisticated edit checking involves what is known as "rule-based edit checking." Rule-based checking is an application of expert systems technology to create intelligent user interfaces.

Intelligent user interfaces are capable of applying more complex rules than ordinary edit checking, to screen a data entry. For example, an interface to a patient database may employ a number of rules to an entire patient record before accepting the fields that comprise it. Some of the rules employed may involve: checking that a female diagnosis is only applied to patients with sex = F; ensuring that a patient's surgeon is not classified as an internal medicine physician, etc. These rules need to be designed to catch the most common classes of data entry errors.

An important example of a class of errors which can be prevented with rule-based checking involves the proliferation of names for a particular item. For example, the name of a hospital may be stored in an insurance database as: "Mt. Auburn Hospital", "Mount Auburn Hospital", "Mount Auburn", "Mt Auburn Medical Center", etc. Such proliferation can make it almost impossible for the insurance company to list all of the claims paid for patients at Mount Auburn. An intelligent user interface could potentially use rule-based methods to screen hospital entries and ensure that they are represented in a consistent manner.

7.6 Data Warehouses

A data warehouse[12] consists of a set of technologies and administrative procedures which an information systems department puts into place in order to offer the organization a guaranteed accurate source of data. Typically, the emphasis is placed on keeping a particular set of corporate databases clean. Furthermore, the concept of a data warehouse extends beyond simply maintaining an accurate source of data. It also includes organizational rules specifying when and for which purposes the clean data must be used. In this manner, data warehouses are designed to address two primary sources of data accuracy errors throughout the organization: inaccurate data input and end-user corruption.

Inaccurate Data Input

Data can enter a corporate system from a number of internal and external sources (see diagram in Section 6.2). It can be keyed in through an internal application, uploaded from a departmental source, purchased from an external source, or received electronically from a customer or vendor (EDI). A data warehouse establishes screens which check the quality of entering the system from all of these sources. Strict rules are also established which determine what types of data from which sources can be uploaded into the warehouse.

End-User Corruption

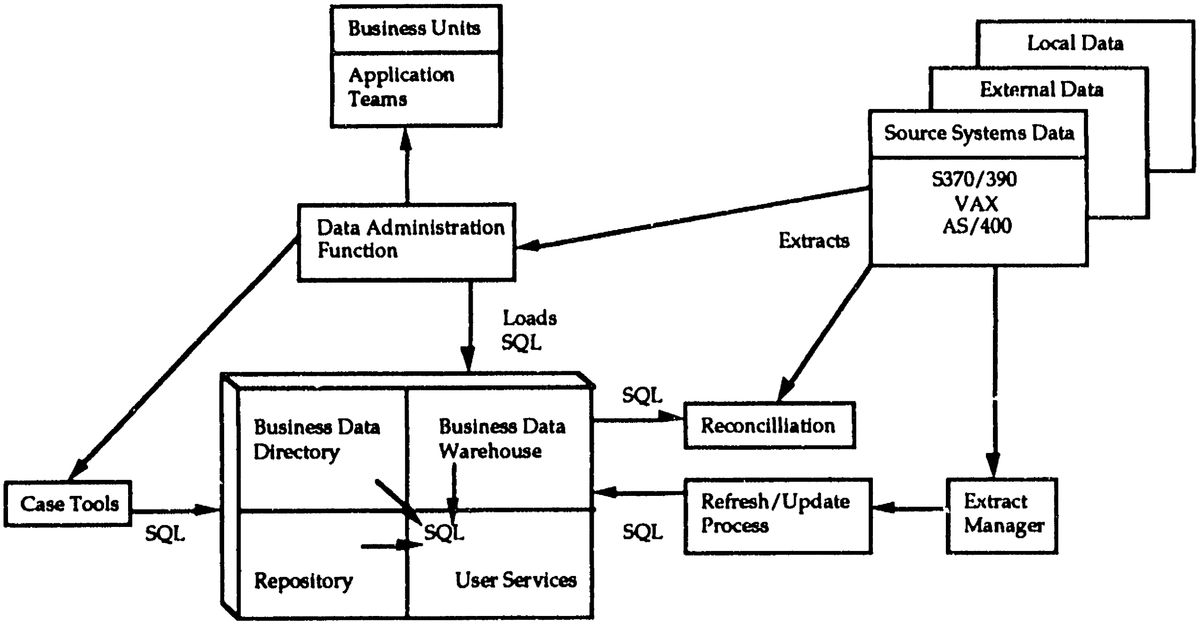
For many organizations, data warehouses are intended to address the rampant problem of, data accuracy being impaired by end-user corruption.

[12] In the course of this research, many organizations presented plans for implementing secure and accurate sources of data. These were most commonly referred to as "data warehouses".

These data quality problems have been one of the negative consequences of the PC revolution. Once data is downloaded into a PC, corporate MIS has no control over what the end user will do with it. As a result, many management reports that are generated on departmental desktops contain errors. These errors result from problems with data accuracy and timeliness. Frequently, data which has been downloaded onto PCs is not kept up to date by the enduser and as a result "old data" is often used to generated important reports. Potentially even more serious problems occur, however, when contaminated data gets uploaded from PCs back into corporate systems. Organizations are often interested in distributing data which resides in PC, but because of quality problems they are reluctant to do so.

Example

One large bank which we spoke with is implementing a data warehousing concept which they refer to as a data delivery utility. Bank X's envisions this utility as a combination of standard technologies and procedures which will enable them to deliver data quality along all four parameters to meet their business needs. The diagram below shows a process view of the data delivery utility.



Bank X's Data Delivery Utility Architecture

At the core is a business data warehouse which stores all of the corporate business data and is guaranteed to be accurate. The business data warehouse is proposed to be an SQL based relational database with strictly controlled access paths. Data uploads from source systems data, external data, or local data, must either be cleared by the data administration function or else pass through an extract managcr which verifies quality. Downloading of data is managed through user services. One of the functions of user services is to establish rules defining which corporate purposes require data from the business data warehouse. The intention is to prevent end-user corruption of data which is used for strategic purposes.

A business data directory is integrated into the data delivery architecture to provide standard data definitions and a dictionary to the business data warehouse. Furthermore, a repository will be implemented for storing the standardized data models and managing the CASE applications development life cycle.

7.7 Expert Systems Based Measurement Tools

An expert systems based measurement tool is a sophisticated software application which can be used to identify data accuracy problems. A crude example of this idea has been incorporated into most commercially available relational databases in the form of range restrictions on data elements. In this case, the database is smart enough to reject data outside of a specified domain. The objective of expert systems based measurement tools is to carry this idea further, to the point where systems can recognize many of the data accuracy problems which a human inspector would be able to identify.

Such systems could best be used to check for statistically unlikely events and business rule violations. Software and hardware technology has advanced to the point where a significant body of business and statistics knowledge can be practically incorporated into an application and applied to a large database. In this manner, these types of expert systems are beginning to be installed by a group of lead users for whom data accuracy is particularly important: third party data providers. For a detailed example of expert systems based measurement tools, we consider the work which has been done at Information Resources, Inc (IRI).

IRI's Expert Systems Tools

IRI buys scanner data from supermarkets across the country and then resells it to consumer marketing firms such as Proctor & Gamble and General Mills. In order to check the quality of the data which they purchase from supermarkets, IRI has developed a number of sophisticated measurement tools. These tools are preventative in nature: they ensure that the data which enters IRI's database is accurate.

One system checks all incoming UPC codes against an existing UPC code dictionary and also checks to ensure that the price paid for any item is reasonable with respect to that UPC code. Another system is responsible for quality control on each individual tape that comes in from the stores. Typically, one week of data comes in at a time. This system performs standard, rule-based tests for statistically unlikely events on all the data in each tape. It tracks 30 measures (e.g. total number of UPC codes, total store volume, UPCs with the same volume from week to week, etc.) and uses a number of rules to compare these measures. Comparisons are made to previous weeks' data and intercorrelations among the 30 measures are performed. At this point, if problems are detected, another rule base is employed to do preliminary diagnosis of the statistically revealed data quality problems.

Furthermore, IRI claims that a final system employs neural network software technology to look through every single UPC record each week and compare it against historical data. The organization processes 35 million records per week in this manner. The applications developers assert that the neural net is trained on a certain set of data quality cases based on actual data. These cases are instances of particular data quality problems (e.g. incorrect pricing, incorrect volume, unexpected increase in sales, etc.). The neural net software flags suspicious individual events and makes a first guess at a diagnosis. For a routine diagnosis, the software fixes the data without human intervention. For a more complicated, or rare, diagnosis, the software will flag a human operator. It is claimed that this system allows IRI to go beyond the very crude algorithmic code of traditional rule-based systems, to incorporate fuzzy rules and human intuition into their data quality analysis.

The IRI systems point out the tremendous unexploited potential that exists in todays hardware and software technology. Most IS organizations are probably not aware that it is technologically feasible to process 35 million records per week through a sophisticated rule based system at reasonable cost. The implications for data quality management here are considerable since this technology offers the possibility of implementing data inspection at a reasonable cost.

Examples: Manufacturer H; Hospital E

Two examples from Section 6.2 point out some potential applications of expert systems measurement tools. First, Manufacturer H has indicated that the vast majority of the EDI data which they receive has significant quality problems. As a result, extensive human involvement is required to integrate this data into their manufacturing process. Systems such as those implemented at IRI could offer considerable benefit in reducing the cost of EDI data inspection. Second, Hospital E related the story of a registration clerk who continually entered her own zip code, instead of the patient's, into an admissions database. Rule based checks for statistically unlikely events would pick up on the preponderance of patients from one zip code area and alert the IS staff to such data entry problems.

7.8 Auditability Tools

More than 50% of the participants in the data quality survey expressed difficulty in tracking down data quality problems. As has been indicated previously, because of the survey bias, we can assume that data auditability represents a considerable problem for most organizations. The fundamental problem, as expressed by many companies, is that frequently data quality problems are detected so far downstream from the data source, that it is difficult to determine their cause. Recent work at MIT's Sloan School of Management points to some interesting technology which could be developed to address this problem.

Source Tagging

This technology is known as data source tagging and addresses the following two fundamental problems. First, source tagging provides a means of determining where data comes from. Secondly, it indicates which intermediate sources were used to arrive at the data. Wang and Madnick [WM90] have developed a sophisticated mathematical model, the polygen model, for resolving these data source problems in the context of relational databases. As expressed by these authors, the fundamental idea of data source tagging is that each data value stored in a relational table has a tag indicating its original source, and a tag indicating its intermediate source. The polygen algebra provides a mechanism for updating these tags with respect to the five orthogonal operators defining the relational algebra: projection, cartesian product, restriction, union, and difference. In this manner, the polygen algebra represents a straightforward extension of the relational algebra which can be used to incorporate source tagging.

Putting source tagging into practice in most IS organizations would require the support of a commercial vendor. This could happen if an existing relational database vendor (e.g., IBM, Oracle, Sybase) extended their relational database engines to incorporate the polygen algebra. Source tagging also represents an opportunity for a new relational database vendor to offer a product, based on the polygen model, which targets organizations where data auditability is critical. As we discuss below, source tagging may be particularly attractive to organizations which are implementing data warehouses.

Incorporating Source Tagging into the Data Warehouse Concept

As discussed in Section 7.6, data warehouses are designed to provide a guaranteed accurate source of corporate data. However, as Wang and Madnick [WM90] have pointed out, for many managers source knowledge is a critical aspect of accuracy. Most managers feel more comfortable with data if they can apply their own judgement to the credibility of the source. In fact, many indicate that without source knowledge data can be total useless to them. This indicates that IS organizations need to provide some mechanism for providing source information along with data stored in a data warehouse.

In addition, however, source tagging provides a powerful mechanism to aid in the implementation of data warehouses. One of the primary concerns expressed by companies during the course of this research related to the fact that data is frequently contaminated at the departmental or enduser level and then incorporated into corporate databases. Source tagging could go a long way toward preventing this sort of contamination. In an environment with source tagging, data would reflect any intermediate processing which occurred at the departmental level, or on an enduser's desktop. In this case the data warehouse could be designed to refuse uploaded data which had experienced such intermediate processing. On the other hand, if the data warehouse did not have such sophisticated screening tools, at least when IS finally detects the corrupted data, they will be able to trace the problem back to the department or desktop and therefore determine the extent of the problem and design a solution.

Directions for Further Research

In this thesis the reader has been presented with a broad introduction to data quality management. This is an important and ripe field for research and this work represents a first step toward understanding the business implications of data quality problems and proposing some solutions. In the author's opinion, considerable benefit could be derived from additional focused research efforts in this area. As a result, these final paragraphs offer three suggestions to guide researchers interested in pursuing more focused studies of data quality.

Focused Field Research

In order to define and explore many parameters of data quality, the field research conducted to support this thesis was necessarily very broad in scope. At this point, a specific industry or type of information system (e.g. reservation, order entry, securities processing) should be examined in detail. Careful and detailed observations of quality levels and management techniques should be made and compared across a small group of organizations. This would allow a detailed evaluation of the effectiveness of varying data quality management techniques. It would also offer the opportunity to identify best practice and quantify the economic benefits of superior data quality management.

Technology Study

Technology also represents an important area which needs to be explored in depth. One or two of the technologies discussed in Chapter 7 could be studied across a number of organizations and assessed as a tool for managing data quality. In addition, there is a need for further technical research to develop and test specific applications of data quality management technology.

International Research

Finally, significant insights could conceivably be realized by studying data quality management on a global basis. Just as comparative research involving Japan and the United States[1] has shed light on the subject of manufacturing quality, so a survey of international approaches to data quality management could be enormously revealing.

In conclusion, data quality represents an area where there is the potential for academic research to make a significant contribution to solving vital business problems. As American business moves into the 1990s, the value which management places on information as a strategic asset will most likely continue to increase dramatically. However, as more and more organizations are going to discover, the value of information as a corporate asset depends critically on the quality of the underlying data. As a result, corporations can be expected to increasingly require innovative new technologies and techniques for managing the quality of their data.

[1] Garvin, D.A. 1983. "Quality on the Line." *Harvard Business Review.* Sept.-Oct.

Appendix A

Organizations Participating in the Data Quality Research

The following organizations provided input at various points in this research either through interviews, surveys, or personal experience.

Insurance
Aetna Life and Casualty
Blue Cross/Blue Shield
Hanover Insurance
John Hancock
New England Life
State Mutual Assurance Company

Health Care
Baystate Health Care
Brigham and Womens Hospital
Children's Hospital, Boston
Dana Farber Cancer Institute
Massachussetts Department of Public Health
Medical East
New England Medical Center

Financial
Bank of Boston
Federal Reserve Bank of Boston
Fidelity Investments
First Boston
Goldman Sachs & Co., Inc.
John Hancock Financial Services
Merrill Lynch
Morgan Stanley

General
Ames Department Stores
Computer Horizons Corporation
Corning Inc.
DMR Group
First Church of Christ, Scientist
GTE Government Systems
Hewlett Packard

Information Resources Inc.
Instron Corporation
MCI
Massachussetts Department of Vital Records and Statistics
New England Power Service Company
Nynex Information Resources
Ocean Spray
Proctor & Gamble
University of Massachusetts at Amherst

Appendix B

MIT Data Quality Survey

Name: ______________________ Company: ______________________
Title: ______________________ Phone: ______________________

Consider the *information products and services* such as reports, decision support systems, accounting records, customer files, customer service, mailing lists, which are produced from corporate data. How accurate are these products?

□.....100% Information products and services never contain errors.
□.....99% Information products and services rarely contain errors.
□.....95% Information products and services occasionally contain errors.
□.....90% Information products and services frequently contain errors.
□.....below 90% Information products and services are plagued with errors.

Now consider the data underlying these information products and services. The quality of this data can be defined in terms of four parameters: accuracy, interpretability, availability, and timeliness. Accuracy measures the correctness of the information stored in data. Interpretability measures how easy it is to extract understandable information from the data. Availability measures how quickly information stored in corporate data can be gathered by the people who need it. Timeliness measures how up to date the information stored in the data is.

Please estimate the quality of your corporate data along these four parameters.

Accuracy:	□ Excellent.	□ Good.	□ Fair.	□ Poor.	□ Terrible.
Interpretability:	□ Excellent.	□ Good.	□ Fair.	□ Poor.	□ Terrible.
Availability:	□ Excellent.	□ Good.	□ Fair.	□ Poor.	□ Terrible.
Timeliness:	□ Excellent.	□ Good.	□ Fair.	□ Poor.	□ Terrible.

List the three major challenges your IS organization faces in maintaining the quality of corporate data.

a. ____
____
b. ____
____
c. ____
____

Does your company have any decision support systems? □ Yes □ No
If so, what techniques do you use to ensure that the data which drives them is clean?
____
____

When you migrate data, how do you verify its quality? ______________________
____
____

How long does it take?______________________________________

Consider the data stored and maintained within your company's various departments. How comfortable do you feel about using this data?

Appendix B

MIT Data Quality Survey

Name: ______________________ Company: ______________________
Title: ______________________ Phone: ______________________

Consider the *information products and services* such as reports, decision support systems, accounting records, customer files, customer service, mailing lists, which are produced from corporate data. How accurate are these products?

□.....100% Information products and services never contain errors.
□.....99% Information products and services rarely contain errors.
□.....95% Information products and services occasionally contain errors.
□.....90% Information products and services frequently contain errors.
□.....below 90% Information products and services are plagued with errors.

Now consider the data underlying these information products and services. The quality of this data can be defined in terms of four parameters: accuracy, interpretability, availability, and timeliness. Accuracy measures the correctness of the information stored in data. Interpretability measures how easy it is to extract understandable information from the data. Availability measures how quickly information stored in corporate data can be gathered by the people who need it. Timeliness measures how up to date the information stored in the data is.

Please estimate the quality of your corporate data along these four parameters.

Accuracy:	□ Excellent.	□ Good.	□ Fair.	□ Poor.	□ Terrible.
Interpretability:	□ Excellent.	□ Good.	□ Fair.	□ Poor.	□ Terrible.
Availability:	□ Excellent.	□ Good.	□ Fair.	□ Poor.	□ Terrible.
Timeliness:	□ Excellent.	□ Good.	□ Fair.	□ Poor.	□ Terrible.

List the three major challenges your IS organization faces in maintaining the quality of corporate data.

a. ____
____
b. ____
____
c. ____
____

Does your company have any decision support systems? □ Yes □ No
If so, what techniques do you use to ensure that the data which drives them is clean?
____
____

When you migrate data, how do you verify its quality? ______________________
____
____

How long does it take?______________________________________

Consider the data stored and maintained within your company's various departments. How comfortable do you feel about using this data?

□.....Absolutely comfortable. Important business decisions are based on this data.
□.....Moderately comfortable. Suitable for informal analysis and internal use.
□.....Slightly uncomfortable. Check twice before using it for anything important.
□.....Very Uncomfortable. The departmental data is almost unusable.

Many IS organizations would like to be able to use the information stored in their departmental databases for corporate or inter-departmental purposes. (e.g., building inter-departmental applications or decision support systems) How easy is it for your IS organization to use data stored in departmental databases?
□ Very easy. □ Easy. □ Somewhat difficult. □ Difficult. □ Impossible.

List three obstacles your IS organization faces when trying to use departmental data.
a. ______________________________
____
b. ______________________________
____
c. ______________________________
____

A large portion of the data in corporate databases is generated internally through user interfaces to applications software. Which of the following technologies does you company use to ensure data quality at the user interface?
□ Double entry. □ Choice lists. □ Context-based help.
□ Automated edit checking. □ Rule-based checks.
□ Other ______________________________

Is there a certification program in place for end user data? □ Yes. □ No.
If so, please describe it. ______________________________
____

Most companies today buy data from outside vendors such as Reuters, Information Resources, Dun & Bradstreet, etc. Estimate the quality of the data which you buy from outside vendors.
□ Excellent. □ Good. □ Fair. □ Poor. □ Terrible.

How do you recognize quality problems in data you purchase?______________________________
____
How do you solve these problems?______________________________
____

When data quality problems are discovered, how easy is it to track down the source?
□ Very easy. □ Easy. □ Somewhat difficult. □ Difficult. □ Impossible.

List two methods which you use to track down the sources of data quality problems.
a. ______________________________
____
b. ______________________________
____

Which of the following technologies does your company use to ensure data quality?
□ Expert Systems □ Statistical Sampling □ Human verification
□ Other ______________________________

Appendix C

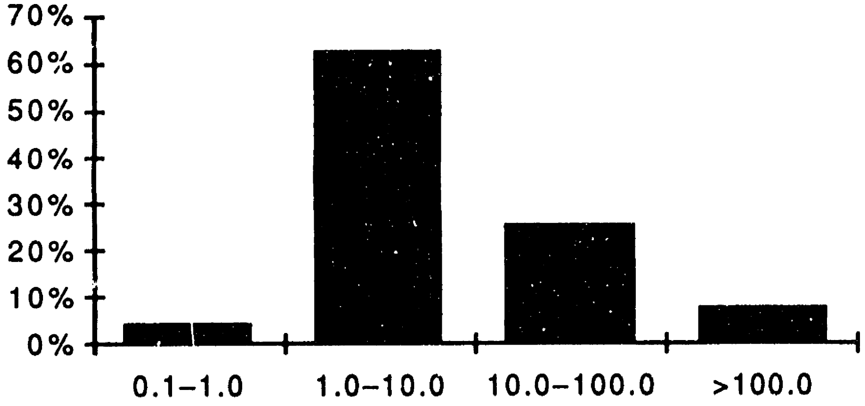
Quality Management Maturity Grid					
Measurement Categories	Stage 1: Uncertainty	Stage 2: Awakening	Stage 3: Enlightenment	Stage 4: Wisdom	Stage 5: Certainty
Management understanding and attitude	No comprehension of quality as a management tool. Tend to blame quality department for "quality problems."	Recognizing that quality management may be of value but not willing to provide money or time to make it all happen.	While going through quality improvement program learn more about quality management; becoming supportive and helpful.	Participating. Understand absolutes of quality management. Recognize their personal role in continuining emphasis.	Consider quality management an essential part of company system.
Quality organization status	Quality is hidden in manufacturing or engineering departments. Inspection probably not part of organization. Empahsis on appraisal and sorting.	A stronger quality leader is appointed but main emphasis is still on appraisal and moving the product. Still part of manufacturing or other.	Quality department reports to top management, all appraisal is incorporated and manager has role in management of company.	Quality manager is an officer of company; effective status reporting and preventitive action. Involved with consumer affairs and special assignments.	Quality manager on board of directors. Prevention is main concern. Quality is a thought leader.
Problem handling	Problems are fought as they occur; no resolution; inadequate definition; lots of yelling and aqusations.	Teams are set up to attack major problems. Long-range solutions are not solicited.	Corrective action communication established. Problems are faced openly and resolved in an orderly way.	Problems are identified early in their development. All functions are open to suggestion and improvement.	Except in the most unusual cases, problems are prevented.
Cost of quality as % of sales	Reported: unknown Actual: 20%	Reported: 3% Actual: 18%	Reported: 8% Actual: 12%	Reported: 6.5% Actual: 8%	Reported: 2.5% Actual: 2.5%
Quality improvement actions	No organized activities. No understanding of such activities.	Trying obvious "motivational" short-range efforts.	Implementation of the 14-step program with thorough understanding and establishment of each step.	Continuing the 14-step program and starting Make Certain.	Quality improvment is a normal and continued activity.
Summation of company quality posture	"We don't know why we have problems with quality."	"Is it absolutely necessary to always have problems with quality?"	"Through management commitment and quality improvement we are identifying and resolving our problems."	"Defect prevention is a routine part of our operation."	"We know why we do not have problems with quality."

Source: Crosby, P.B. 1979. Quality is Free. New York: McGraw-Hill.

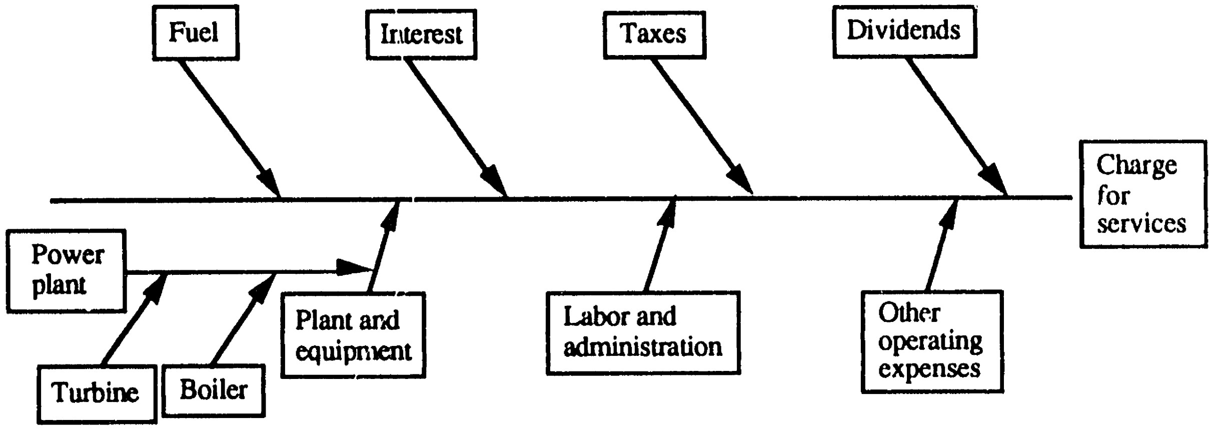
Appendix D

Pareto Analysis:
Pareto analysis is simply the classification of defect types into categories. Such analysis allows management to isolate the most common quality problems and focus attention on solving them.

Defects Caused by Dust Particles in Particular Size Ranges



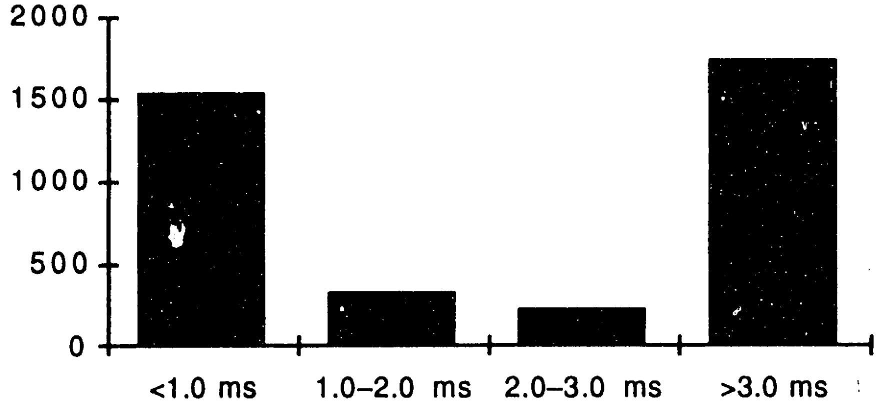
Ishikawa Diagrams:
Ishikawa diagrams are used to illustrate cause and effect relationships. Management uses them to trace observed quality problems back to their source.



Components of all costs for a typical electric company.
Source: Deming, *Out of the Crisis*

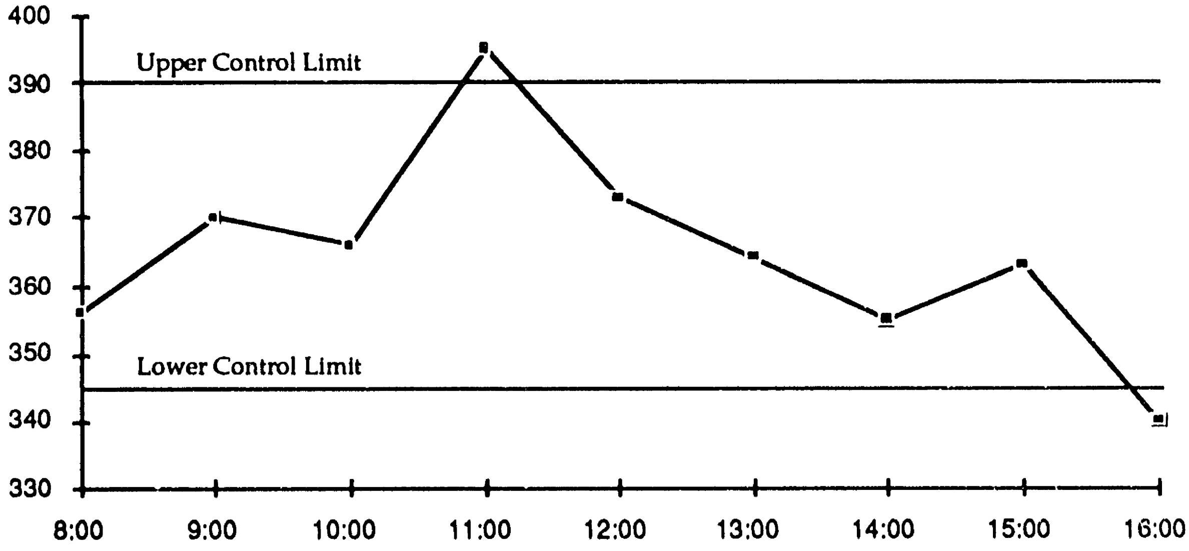
Histograms:
Histograms illustrate the distribution of measurements around a specification. These charts graphically represent the degree of variation in a process.

Distribution of Read Response Times for Memory Chip XK142



Control Charts:
Control charts are used to monitor the effective operation of manufacturing processes. Upper and lower control limits are set for parameters which are measured at regular intervals. As long as the measurements remain within the control limits, the process is assumed to be operating effectively and efficiently. When a process moves outside of the control limits, management can take immediate corrective action.

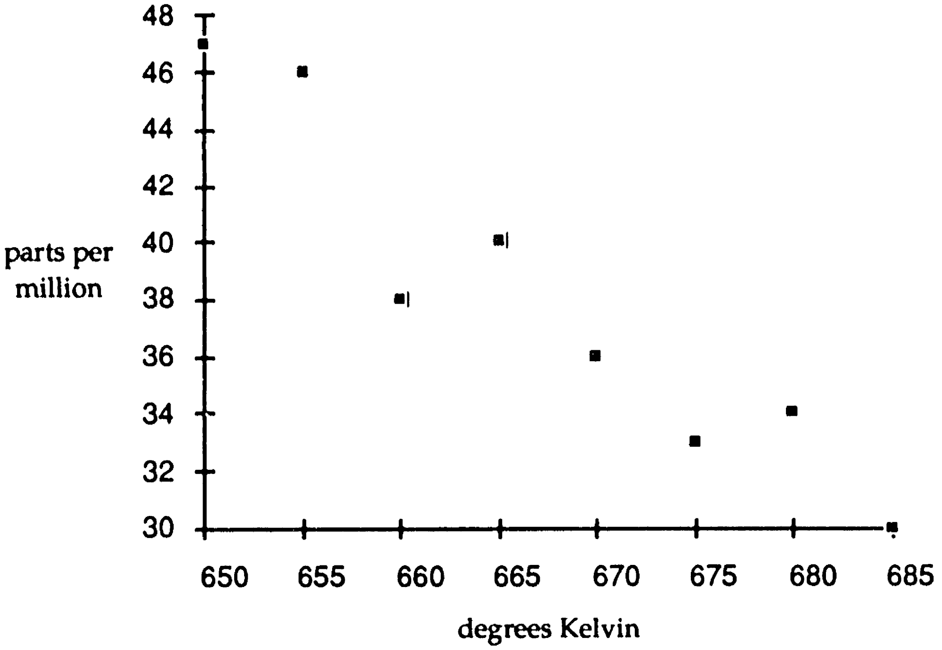
Hourly Temperature Measurements for Reactor 11A



Scatter Plots:

Scatter plots are used to discover correlations between variables. For example, a quality manager may have a hunch that furnace temperature and chemical purity are related in a manufacturing process. In this situation, he would develop the following scatter plot.

Impurities vs. Temperature



Bibliography

[B87] Bock, R.N. 1987. *Scan Data Quality: The Nielsen Approach,* The Nielson Researcher, Vol. 1, Issue I, Spring, pp. 2-9.

[BW39] Brigham & Women's Hospital. 1989. *The Informant,* Spring.

[BW90] Brigham & Women's Hospital. 1990. *The Informant,* Fall.

[C90] Corning Inc. 1990. *Information Services: World Class Quality Plan,* Sept.

[C83] Chen, P.P.S. 1976. *The Entity-Relationship Approach to Information Modeling and Analysis.* North-Holland. Amsterdam.

[C79] Crosby, P.B. 1979. *Quality is Free.* New York: McGraw-Hill.

[D90] Date, C.J. 1990. *An Introduction to Database Systems,* Addison-Wesley

[D86] Deming, W.E. *Out of the Crisis,* Massachusetts Institute of Technology, Center for Advanced Engineering Study, 1986, pp. 3.

[D85] Durrell, W. 1985. *The Politics of Data.* Computerworld. Sept. 9. pg. ID25-36

[E81] Edelman, F. 1981. *The Management of Information Resources: A Challenge for American Business.* MIS Quarterly. March. pp. 17-27.

[F83] Feigenbaum, A.V. 1983. *Total Quality Control.* 3rd ed. New York: McGraw-Hill.

[F90] Fine, C. 1990. *Total Quality Management,* MIT Sloan School of Management, Aug.

[FB87] Fine, C.H. and Bridge, D.H. 1987. *Quest for Quality: Managing the Total System.* Industrial Engineering and Management Press

[G83] Garvin, D.A. 1983. "Quality on the Line." *Harvard Business Review.* Sept.-Oct.

[GW90] Glasser, J.P. and Williams-Ashman, A. 1990. *There's Something Wrong with the Data,* Brigham and Women's Hospital, draft, Oct.

[GQR88] Goodhue, D.L., Quillard, J.A., Rockart, J.F., 1988. *Managing The Data Resource: A Contingency Perspective.* MIS Quarterly, Vol. 12 No. 3. Sept.

[HC88] Hauser, J.R. and Clausing, D. 1988. *The House of Quality.* Harvard Business Review. May-June. pp. 63–73

[IBM89a] IBM Corp. 1989. *AD/Cycle: Blueprint for a more productive future,* marketing brochure.

[IBM89b] IBM Corp. 1989. *Getting started–planning for IBM's AD/Cycle,* marketing brochure.

[IBM88] IBM Corp. 1988. *Systems Application Architecture: An Overview,* Nov.

[I91] Information Week. *MIS and the Pursuit of Quality.* Jan. 7 1991 pg. 36.

[I76] Ishikawa, K. 1976. *Guide to Quality Control.* Tokyo: Asian Productivity Organizations.

[J64] Juran, J.M. 1964. *Managerial Breakthrough.* New York: McGraw-Hill.

[J3] Juran, J.M. (ed.). Quality Control Handbook. 3rd ed. New York: McGraw-Hill

[LB88] Lin, E. and Blanton, R. 1988. *Selecting Hospital Information Systems,* Journal of Systems Management, May, pp. 24-27

[ML90] Merrill Lynch & Co. 1990. *Measuring the Cost of Quality, ,* Inc., personal communication, Oct.

[OA88] Oman, R. and Ayers T. 1988. *Improving Data Quality,* Journal of Systems Management, May, pp. 31-35

[PR90] Ragozzino, P.P. 1990. *IS Quality – What Is It?,* Journal of Systems Management, Nov. pp. 15-16

[R90] Rin, N.A. 1990. *Case: Concepts, Benefits, Challenges and 5-year Outlook,* N. Adam Rin, Gartner Group, Inc., presentation to MIT Sloan School of Management, Nov. 1990.

[SK88] Sass, J. and Keefe, T. 1988. *MIS For Strategic Planning and a Competitive Edge,* Journal of Systems Management, June. pp. 14-17.

[S82] Schonberger, R.I. 1982. *Japanese Manufacturing Techniques.* New York: The Free Press

[S31] Shewhart, W.A. 1931. *Economic Control of Quality of Manufactured Product*. New York: McGraw-Hill

[S90] Software Magazine. 1990. *Carving Out Systemview: Research at IBM,* Nov. pp. 22-52

[ST89] Sullivan-Trainor, M. 1989. *The Push for Proof of Information Systems Payoff,* Computerworld, April 3, pp. 55-57

[TT84] Tribus, M. and Tsuda, Y. 1984. "Creating the Quality Company." Working Paper. Center for Advanced Engineering Study. M.I.T.

[WM90] Wang, Y.R. and Madnick, S.E. 1990.*Where Does the Data Come From: Managing Data Integration with Source Tagging Capabilities,* MIT Sloan School of Management, personal communication, Oct. 1990.