# Survey of Data Management and Data Quality Texts

CAS Data Management and Information Educational Materials Working Party

**Abstract:**

**Motivation.** Recent focus on corporate governance (e.g., Sarbanes-Oxley) in the United States and the use of predictive modeling techniques in the property/casualty insurance industry have raised the profile of data management and data quality issues in the actuarial profession.

**Method.** Representatives of the Insurance Data Management Association (IDMA) identified seven data management texts they felt would be most helpful for actuaries. Two additional texts were added to fill out the data quality perspective.

**Results.** Actuaries reviewed each of the recommended texts from an actuarial perspective.

**Conclusions.** The working party hopes that this paper will be a resource for actuaries dealing with data management and/or data quality issues. By looking at the summary information in the tables of section 4, readers may be able to narrow down candidate books to those that will best meet their needs and then read the specific reviews in section 3.

**Keywords.** Data Quality; Data Administration, Warehousing and Design; Actuarial Systems; Data Collection and Statistical Reporting; Software Testing.

## 1. INTRODUCTION

Recent focus on corporate governance (e.g., Sarbanes-Oxley) in the United States and the use of predictive modeling techniques in the property/casualty insurance industry have raised the profile of data management and data quality issues in the actuarial profession. Actuaries have a unique role with respect to data quality because they typically understand the process and pitfalls better than management and at the same time they understand the business meaning and impact of errors better than data and systems professionals. For example Francis [1] points out that 80% or more of time spent on large predictive modeling projects is spent on data issues. Also, in December 2004, the Actuarial Standards Board updated their standard of practice on data quality (Actuarial Standard of Practice No. 23) [2].

This paper provides an overview of several resources on information quality by surveying seven non-actuarial data quality and data management textbooks recommended by the Insurance Data Management Association. Two additional texts recommended by a working party member are also reviewed. A discussion of data quality is incomplete without reference to related data management topics such as data structure, data storage, metadata, and software errors. In this paper we will employ the term "information quality" to refer to the broader set of data management topics related to data quality.

Thus, this paper provides resources for actuaries with data management or data quality questions and these resources may provide suggestions to improve data quality. Hopefully

we will motivate our readers to pursue further education on information quality using one or more of the books surveyed. In addition, within our review of the literature, we present an overview of some of the key concepts of information quality.

## 1.1 Research Context

The actuarial literature on data quality and data management is relatively sparse. The Actuarial Standard Board (ASB) Standard of Practice No. 23 on data quality [2] provides a number of guidelines to actuaries when selecting data, relying on data supplied by others, reviewing and using data, and making disclosures about data quality. The guidelines advise actuaries to review data for reasonableness and consistency. The actuary is also advised to obtain a definition of data elements in the data, to identify questionable values and to compare data to the data used in a prior analysis. The actuary is also advised to judge whether the data is adequate for the analysis, requires enhancement or correction, requires subjective adjustment, or is so inadequate that the analysis cannot be performed.

The Casualty Actuarial Society (CAS) Committee on Management Data and Information and the Insurance Data Management Association (IDMA) also produced a white paper on data quality [3]. The white paper states that evaluating the quality of data consists of examining the data for:

- Validity,
- Accuracy, including concepts of absolute accuracy , effective accuracy and relative accuracy,
- Reasonableness, and
- Completeness.

The CAS Committee on Management Data and Information also promotes periodic calls for papers on data management and data quality which are published in the CAS *Forum.* Among the papers on information quality submitted to the program are Francis [1] and Popelyukhin [4]. Francis's focus is mainly on techniques from exploratory data analysis that can be applied by actuaries to detect glitches and other data quality issues in data supplied for an actuarial analysis. Popelyukhin describes data quality issues encountered by actuaries when relying on data supplied by non-actuaries and external data suppliers, such as that supplied by third party administrators and presents the data quality shield as a solution to insurance data quality problems.

The subject of data quality is also of interest internationally. A working party of the U.K. General Insurance Research Organization (GIRO) developed recommendations for improving the quality of reserve estimates. The Reserving (GRIT) working party report [5] recommended more focus on data quality and suggested that U.K. professional guidance notes incorporate standards from U.S. Actuarial Standards of Practice (ASOP) No. 23. Furthermore the GRIT survey found that many respondents expressed concern over data quality.

Note that none of these references specifically addresses data management.

## 1.2 Objective

The objective of this paper is to address gaps in actuaries' knowledge of information quality. The current CAS literature provides a basic introduction to information quality issues. However, there is very little available for those wishing a more advanced knowledge of the subject or for those who have responsibilities involving data management and data validation. Moreover, the current state of actuarial literature on information quality does not equip actuaries to become active advocates for information quality; i.e., to advise management on systems and protocols for improving information quality. This paper attempts to narrow this gap by reviewing several recommended books from an actuarial point of view.

## 1.3 Disclaimer

While this paper is the product of a CAS working party, its findings do not represent the official view of the Casualty Actuarial Society or the employers of the Working Party members. Moreover, while we believe the textbooks reviewed here are good sources of educational material on data management and data quality issues, we do not claim they are the only appropriate ones.

## 1.4 Outline

The remainder of the paper proceeds as follows. Section 2 will discuss the increasing importance of data management and data quality to actuaries, as well as how the reading list was developed. Each subsection of section 3 is a book review of one text. Section 4 summarizes and compares the working party's evaluations of the textbooks on five star rating scales.

## 2. BACKGROUND AND METHODS

### 2.1 Motivation

Information quality issues have come to forefront recently due to several key developments:

- **(Unprecedented) level of detail.** Computerization and cheap data storage along with changes in regulatory requirements have led to extraordinary amounts of data being captured, stored and provided to actuaries. Consequently, enormous amounts of data can amass enormous numbers of errors and inconsistencies.
- **Availability of new tools.** Recent years have seen the proliferation of powerful data analysis packages and technologies: from XML-enhanced data exchange to object-oriented databases to servers enabled with On-Line Analytical Processing.
- **Competition.** Competition encourages pricing techniques to be more and more precise – every percent counts. The precision of estimates is heavily dependent on the quality of the data used in the analyses. In this environment, requirements for quality of data used in pricing algorithms grow immeasurably.
- **Quality of actuaries.** Modern actuaries are more technically prepared for the challenges of dealing with huge amounts of data using contemporary tools and techniques. Prepared with the appropriate information, they should be able to tackle data quality issues with aplomb.

### 2.2 The Reading List

To address these issues, the CAS Committee on Data Management and Information created the Data Management Educational Materials Working Party. A casual search will reveal dozens, if not hundreds, of books on data management. The Insurance Data Management Association (www.idma.org) promotes insurance data management in multiple ways, including accreditation, online courses, information available on their website, seminars, and co-sponsoring forums. Knowing this, the working party asked the IDMA to develop the party's reading list. IDMA representatives narrowed down their syllabus to the texts they felt would be most helpful for actuaries. Louise Francis, four time winner of the CAS Data Management Call Paper program, suggested two additional texts to fill out the data quality perspective.

## 3. THE BOOK REVIEWS

Each of the following sections is the review of one book. The sections are ordered as the underlying texts' focus moves from data quality (3.1, 3.2) to data management (3.3 to 3.7) to special topics (3.8 and 3.9). Some of these reviews have already been published in the *Actuarial Review*. For *Corporate Information Factory* (section 3.5), the text has been altered slightly from that in the *Actuarial Review*. The texts are compared in section 4, so readers may find it helpful to skip to section 4 to determine which text(s) best address their issue(s).

### 3.1 Data Quality: The Accuracy Dimension

*Data Quality: the Accuracy Dimension* [6] by Jack E. Olson (ISBN 1-55860-891-7) focuses on data accuracy, which the author sees as the foundation for the measurement of the quality of data. The author has spent the last 36 years developing commercial software and is an expert in the field of data management systems. This background enables him to address the topic of data quality and accuracy from a practical viewpoint.

There are three parts to this book. The first part defines inaccurate data and shows that many significant business problems arise from inaccurate data. The second part focuses on how a data quality assurance program is constructed using the "inside-out" approach. The last part introduces data-intensive analytical techniques such as data profiling (the use of analytical techniques to discover the true content, structure and quality of data), along with some real world examples of profiling applications.

The author begins the first part, "Understanding Data Accuracy," by introducing real world data quality problems and the concept of data quality assurance technology. The author identifies the essential elements of this technology: experts, educational materials, methodologies, and software tools. In order to define data accuracy in the larger picture of data quality, data is defined as "having quality if it satisfies the requirements of its intended use." Some examples are used to illustrate key aspects of data quality:

- **Accuracy:** An 85% accurate database containing names, address, and phone numbers of physicians in a state would be considered poor quality for notifying physicians of a new law whereas it would be considered high data quality for a new surgical device firm to find potential customers.
- **Timeliness:** A dataset containing monthly sales information which is slow to become complete at the end of each month is poor when it is used to compute

sales bonus in that month whereas it is excellent when it is to be used for historical trend analysis.

- **Relevance:** A dataset without relevant information is of poor data quality for its intended use.
- **Completeness:** A database with 5% of information missing is probably a good quality database for general assessment but is considered to be low quality for evaluation.
- **Understood:** Dataset has to be understood for its intended purpose. *Metadata* is a term used by data management professionals for information about the data such as definitions, a description of permissible values and business relationships that define the data in a database. Comprehensive metadata is a prerequisite for good information quality.
- **Trusted:** Only trusted datasets should be used.

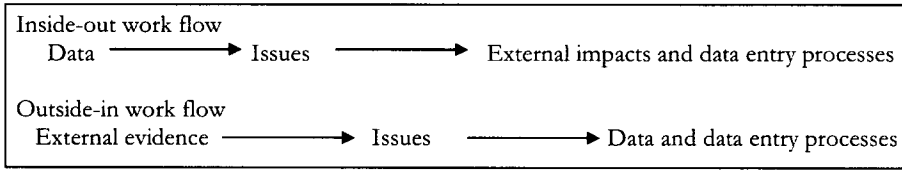
Data accuracy, "the most visible and dramatic dimension of data quality," is then introduced and explained. Data accuracy "refers to whether the data values stored for an object are the correct values." "To be correct, a data value must be the right value and must be represented in a consistent and unambiguous form."

The second part of the book outlines the structure of a data quality program built for identifying inaccurate data and taking actions to improve its accuracy. "A data quality assurance program is an explicit combination of organization, methodologies, and activities that exists for the purpose of reaching and maintaining high levels of data quality." An **inside-out** methodology is believed to be the best way to address accuracy. This method works from a complete and correct set of rules that define data accuracy for a particular dataset. The author defines "inaccurate data evidence" as a collection of facts which are aggregated into issues. The facts might include tabulations of the number of invalid values for variables in the data, totals of the number of missing values, etc. This evidence is produced by the data profiling process defined above. The issues are then analyzed to determine the external impact.

The second approach, the **outside-in** method, looks for issues in the business rather than looking at data. "It identifies facts that suggest that data quality problems are having an impact on the business." The facts are then examined to determine the degree of culpability attributable to defects in the data and if the data has inaccuracies that contribute to the

problem.

Summarizing the two approaches to data quality programs (page 72, *fig.* 4.3):

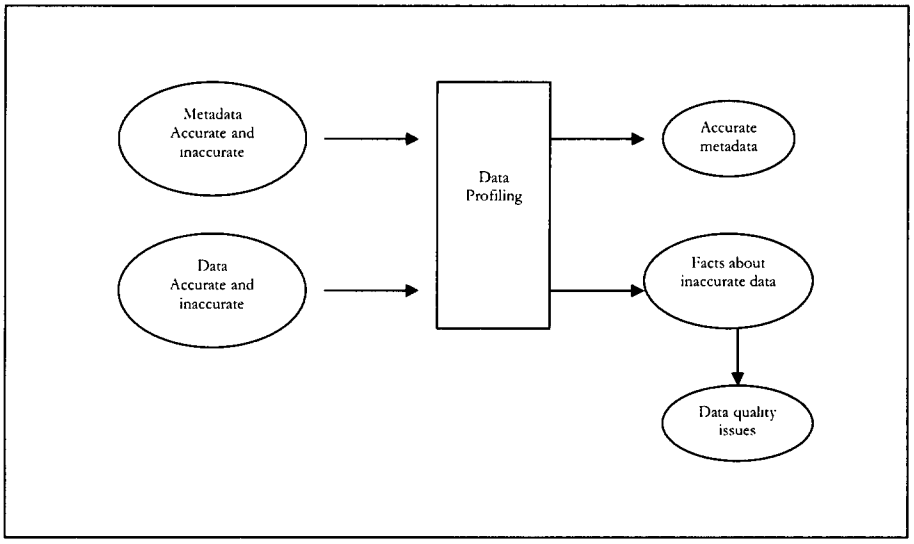



The data quality assurance program also requires an assurance team to decide how it will engage the corporation to bring about improvements and return value for their effort. The author advocates that team members should only be assigned to the data quality assurance team, i.e., this is their full time job -- not a project.

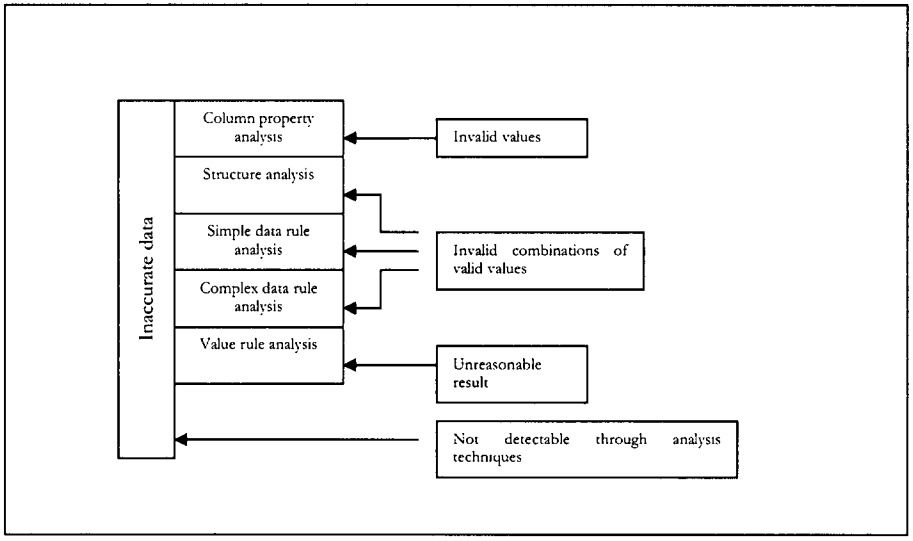
Some of the key technologies used to create and maintain an effective data quality assurance program are:

- **Metadata repositories:** metadata should define what constitutes accurate data. It is essential for determining inaccuracies in data profiling.
- **Data cleaning:** identifying and cleaning up data after data problems have been discovered. It is valuable to clean up data before moving to the next step of data profiling to avoid distortions in the discovery processes of later steps.
- **Data profiling:** the use of analytical techniques to discover the true structure, content, and quality of a collection of data.
- Data filtering
- **Data monitoring:** looking at individual transactions before they cause database changes or looking at the entire database periodically to find issues.

"Data profiling is a new technology that has emerged in the last few years." It uses any known metadata and the data itself to discover the presence of inaccuracies within a database. The general model of a data profiling process can be shown as follows (page 123, *fig.* 7.1):




Data profiling uses a bottom-up approach. It starts at the most basic level of the data and then goes to progressively higher levels of structure. The following diagram (page 131, *fig.* 7.2) illustrates how the major steps of data profiling (in the middle column) can address data issues (in the right hand column):




Within each data profiling step there can be processes for discovery, assertion testing, or value inspection. The outputs of these processes are used to make decisions. The author

discusses each step in a separate chapter with real world examples of the rules and the types of investigative thought required to be effective. The author believes data profiling is probably the single most effective technology for improving the accuracy of data in corporate databases.

Overall, the book provides a thorough introduction to data accuracy and the data profiling technology that could significantly improve data quality. A reader could probably develop a data quality assurance program including data profiling after reading the text, although there is not much on statistical methodologies commonly used to detect data problems. However it does serve as a good reference for data quality structures and concepts.

## 3.2 Exploratory Data Mining and Data Cleaning

The primary topic of the book *Exploratory Data Mining and Data Cleaning* [7] by Tamraprni Dasu and Theodore Johnson is data quality. In data mining circles this book is the reference of choice on data quality and its authors are invited to speak on the topic at many conferences. It combines a review of the most common methods used for screening data for quality with some novel approaches developed by the authors. It also provides a review of key data quality concepts along with some data management concepts relevant to data quality.

An overview chapter summarizes the topics covered in the rest of the book and presents the authors' philosophy towards data quality. The authors lay out the methods of exploratory data mining they will be using: These include parametric summaries (measures of central tendency, dispersion and skewness), as well as non-parametric summaries such as quantiles, histograms and OLAP cubes. The authors believe in "end-to-end-data-quality," that is there are many stages in the data assembly process where data quality needs to be monitored and improved, such as during data gathering, data storage, data analysis and data integration. Their equation:

DATA + ANALYSIS = RESULTS

reflects in equation form the well known adage "garbage in – garbage out." The authors are also proponents of measuring data quality in order to promote data quality improvement.

The book has a chapter on "Exploratory Data Mining" that presents graphical and statistical techniques largely from the exploratory data analysis literature. The methods of

exploratory data analysis were pioneered and the practice given its name by John Tukey (see *exploratory data analysis* at www.wikipedia.org). Its methods are widely accepted in the statistical community as a key activity within any statistical project and its methods are widely implemented in statistical software. Exploratory data mining is an application of exploratory data analysis to large databases that can be used to understand the structure of a database and to detect outliers (data glitches are often found by examining outliers). In this chapter, the authors introduce the novel concept of data depth. Data depth provides a measure of how far a record is from the center of the data or from typical data values. In order to construct such a measure, one needs a way to quantify the notion of "center" and the notion of "distance" from the center. The authors provide the Mahalanobis depth as one way to measure data depth.

In the chapter "Partitions and Piecewise Models" the authors discuss data cubes as a mechanism for exploring data. Data cubes are single or multidimensional tabular summaries of data. Statisticians have long used cross-tabulations, or slicing and dicing of data to develop a high level understanding of the structure of databases. Among practicing actuaries, pivot tables are a common example of data cubes. In this chapter, the authors introduce the concept of data pyramids for comparing two databases for changes. Unfortunately, this concept was a little difficult to follow, even after a couple of readings of the material. The authors also introduce two data mining methods which can be used to model nonlinearities and other data complexities in this chapter: piecewise regression and naïve Bayes.

In their chapter on Data Quality, the authors detail all the mishaps affecting data that create quality problems. Some of the sources of data quality problems are: unreported changes in layout, unreported changes in measurement, temporary reversion to defaults, missing and default values and gaps in time series. Being mindful of the sources of data errors, one can detect, remediate and most importantly, prevent them.

In the Data Quality chapter the authors are strong proponents of implementing data quality measures. The authors believe that in order to motivate improvements in data quality, it is imperative that data quality be measured, even when the measures are somewhat subjective. In developing their measurement approach, both static and dynamic constraints are described. Some of the metrics quantify traditional data quality components such as accuracy, consistency, uniqueness, timeliness and completeness. Others capture other features of data quality such as extent of automation (sample some transactions, follow them through the database creation processes and tabulate the number of manual interventions),

successful completion of end-to-end processes (count the number of instances in a sample that, when followed through the entire process have the desired outcome), and glitches in analysis (measure the number of times and severity in a sample that data quality errors cause errors in analyses). The different metrics are weighted together into an overall data quality index using business considerations and the analysts' goals to develop weights.

The book provides a wrap-up chapter that applies the authors' quantitative techniques to the detection, correction and prevention of data quality problems. In the chapter, methods for detecting and correcting glitches are illustrated. For instance, to address the missing value problem, the authors present techniques (including data imputation) that can be used to create values that substitute for the missing data. The chapter presents an introduction to techniques for joining different data sets, including approximate joining techniques when exact matches are not found between the key fields of two databases. Finally and most importantly, the authors also stress the crucial role of metadata, the information describing the data, and discuss ways of creating good metadata.

Overall, the book provides a thorough introduction to data quality at a level that can be understood by the practicing actuary (with the exception of the material mentioned above on data pyramids).

### 3.3 Improving Data Warehouse and Business Information Quality

*Improving Data Warehouse and Business Information Quality* [8] (ISBN: 0-47125-383-9), by Larry P. English, is a complete detailed treatment of information quality for any type of business. The main theme in this book is that data is a material for informational product and (like in manufacturing) the quality of the product is determined by customer satisfaction. According to the book, *everyone* in the organization has a role in establishing and maintaining information quality to deliver a quality product to the customer. Thus actuaries, as consumers and producers of information, should establish data quality standards and communicate their data quality requirements to the stewards of all their data sources.

The book is multifaceted; it is "a concept book, a textbook, a reference book, and a practitioner's guide." It is generic enough to cover a lot of ground (scenarios, situations, setups) while detailed enough to serve as a step-by-step guide full of relevant examples. Throughout the book the author consistently uses a 4-part template for every proposed step (Input, Output, Techniques & Tools and Process Description) which makes the text immensely useful.

The book is divided into three sections: "Principles of Information Quality Improvement," "Processes for Improving Information Quality" and "Establishing the Information Quality Environment."

In section one, "Principles of Information Quality," the author lays the ground work by defining what data is, what quality is and is not and why we should be interested in information quality in the first place. He then builds upon this foundation work with detailed discussions about the high cost of low data quality and how to measure data quality with detailed examples. He continues with a discussion of quality principles applied to information as a product and each stakeholder's role in producing, planning, controlling, leading, funding, and continuously improving information.

Section two uses many flow diagrams to demonstrate the various process steps for improving information quality. For example, there are diagrams to show the steps in measuring non-quality information costs, establishing the information quality environment, establishing data quality definitions, and assessing data quality. The chapter on data definition and information architecture quality is particularly detailed as the author provides instructions on how to construct data names, build metadata repositories, and provide guidelines for quality business rules. The chapter on information quality assessment shows how to determine sample size and also includes numerous quality assessment templates to show different ways quality measurements and customer satisfaction can be presented. The author places great emphasis on data defect prevention through the process of continuous improvement as "the cost to react to quality problems can be 5 to 10 times as much as the cost of prevention."

Section three shows how "Deming's 14 points of quality" can be applied to the information product. It describes the roles and accountabilities of everyone in the organization, from information producer to executive management, as stewards of information quality. The author points out that management commitment is essential to having a quality improvement environment. He then describes how to start implementing it step by step, including: "creating a vision and objectives, identifying critical success factors, managing change, conducting an information customer survey, selecting a small manageable pilot project, defining the business problem, and assessing the systemic barriers." You clearly get the idea that this is not just about data but about managing processes and people.

With time the book has acquired the flavor of a cautionary tale about obsolete systems. If in 1999 the book was considered to be mostly about cleansing legacy systems and converting

them into new shiny-bright data warehouses, nowadays it can be read as a powerful reminder of how to keep systems current and relevant in a constantly changing environment in order to avoid their transformation into "legacy" systems. According to the book, maintaining data definitions and business rules will make long strides into keeping information from becoming legacy data in need of remediation.

The book's content translates directly to the actuarial situation: actuaries rely on many pieces of data (loss runs, premiums bordereaux, claims classification, etc.), which may be quite imperfect. The caveat is that actuaries rarely (if at all) have control over their data, while the book implicitly assumes that the reader can perform the suggested data cleansing and transformation procedures. Nevertheless, the book is very useful: actuaries would definitely benefit from knowing which data defects may cause problems and of what size. Actuaries should determine the types of potential data errors with the largest impact and presumably should be able to estimate the effects they may have on their data. Ideally, actuaries would use data quality assessment reports to calculate the level of data accuracy.

The book is an extremely valuable source of information for anyone potentially affected by data quality. It can be read as a textbook, as a practitioner's guide, as a cautionary tale, or as an inspirational book. Indeed, learning about data quality problems at source level may even inspire actuaries to incorporate an estimate of data uncertainty into their methods. In summary, even though this is a very long book it does contain a wealth of ideas and techniques that can be used by everyone in the information value chain in carrying out their information quality stewardship responsibilities.

### 3.4 Enterprise Knowledge Management

The purpose of *Enterprise Knowledge Management* [9] (ISBN: 0-12455-840-2) by David Loshin is to provide an enterprise-wide framework for data quality. The author likens the flow of data within an organization to the assembly process in a manufacturing plant, often referring to an organization's data production as "the information factory." The author uses many quality control ideas from the world of manufacturing and applies them to the process of manufacturing information in an enterprise.

The book is divided into chapters each of which outlines one building block of an enterprise data quality program. The book is at once both technically detailed and conceptually rich.

Technical data quality concepts are illustrated by a number of real world data examples. The data examples are not insurance specific, but rather generic, typically using universal business elements such as name, address, location, and phone number. Nevertheless the concepts are universal and especially applicable in an industry like insurance, where data drives the business. The actuary will recognize many of these concepts, described generically in the text, as applicable to the actuarial applications of ratemaking, reserving, or modeling.

While containing some technical details, the text is curiously abstract, relying mostly on high-level conceptual material. It resembles an Actuarial Standard of Practice in that for each topic a list of conceptual considerations and best practices are given, but with few concrete recommendations as to which are most important. That determination is left up to the practitioner's judgment. The text is oriented towards professionals who oversee information flow within an organization: the CIO, the systems manager, or the actuary who oversees information infrastructure.

The author begins with a section on how to build support for data quality management within an organization. The first step is to get senior management buy-in for the program. Start with a small but visible data quality issue. In choosing an initial task, the author invokes the Pareto or "80-20" rule, which states that 80% of the impact is usually generated by 20% of the cases. Quantify both the soft and hard costs of allowing the issue to linger. The author recommends using a process known as COLDQ (cost of low data quality) that maps the information chain, and then builds a Data Quality Scorecard to identify potential problem nodes in the information manufacturing chain.

For instance, if the issue is faulty customer addresses, the associated costs might include hard impacts like the cost to repair data and increased customer service expense; but also soft impacts like increased customer attrition or delay in analysis and initiative implementation dependent on the data. In an insurance setting, these "soft" costs might be manifest in the inability to analyze catastrophe data or to reorganize rating territories, for example. Next, to gain buy-in, demonstrate to management the operational benefit and rate of return associated with fixing the issue. Once the issue is addressed, celebrate the solution and thereby build support and enthusiasm to address further data quality issues. A key component of the solution is to establish a data ownership policy. The author gives many different paradigms for "who should own the information" in various settings, but it should always be formalized and agreed upon.

The author discusses various dimensions of data quality, e.g., completeness, flexibility,

robustness, essentialness, granularity and precision, among others, as they relate to data models, data values, information domains, information presentation, and even the corporate information policy itself. One or two indices are given as guidelines for how to compute each measure of data quality. Foe instance, a complete database is one that contains all of the data required for an analysis while the analyst may request additional data be added to an incomplete database. To measure completeness one might chart the number of requests to add new data fields over time.

Once data quality measures and thresholds have been established, they can be measured either statically or dynamically. Static measurement involves collecting and analyzing past data, usually after the end of a time cycle, and is useful for identifying chronic data quality issues. Dynamic measurement involves inserting data probes into the information chain and measuring output in real time. This is useful for identifying acute data quality issues. Data quality measurement is often implemented via a rules-engine containing data and business rules, and acceptable tolerance thresholds for each. The author spends a fair amount of time in listing considerations when evaluating different rules-based systems and products. Often the choice of a particular rules engine will depend upon whether measurements are primarily static or dynamic.

The author then devotes several chapters to data cleansing. Data cleansing is the act of "fixing" data, i.e., appending, supplementing, or overwriting data whose quality has tested low. Often data quality problems arise when merging data from two different data sources. The author describes techniques used to determine if two different data fields' members come from the same domain. The concepts of overlap, agreement, and disagreement are discussed and a formula given for computing the degree of each between two data sets.

If a data domain is unknown (this usually occurs in string fields housed in legacy mainframe data systems), a number of domain discovery techniques are given; among them agglomerative, divisive, hierarchical, and K-means clustering. Each of these clustering-based methods relies on a notion of distance between data points. Distance rules are typically Euclidean (d = ((x1–x2)^2 + (y1– y2)^2)^0.5), city block (d = |x1-x2| + |y1-y2|), or Exact Match. "Exact match" distance rules are used to compare the distance between strings and are extremely helpful in data clustering, data cleansing, spelling, and address checking routines.

One distance rule used to compare strings is "edit distance" or the minimum number of

basic operations (i.e. insert, delete, transpose) needed to transform a candidate string to a target string. For example the edit distance between "intermural" and "intramural" is 3.

The author also gives a number of approximate matching techniques to match like strings using the notion of distance in combination with various word and phonetic coding schemes such as: Soundex, New York State Identification and Intelligence System (NYSIIS), Metaphone, and N-grammimg. Each of these methods attempts to simplify the phonetic representation of a word (by omitting vowels, coding like sounds, etc.) and then uses the above notions of distance on the coded entries to identify approximate string matches.

As an aid to these matching and clustering techniques, the author enumerates a number of common error paradigms and their causal conditions. For example a data format that is too strict, e.g., insisting on a middle initial for every name entry, will tend to generate erroneous, "placeholder" data entries. These are redundant records added to a database as a result of an erroneous match, or more appropriately, not finding the correct match due to inconsistencies in the fields used to join the data from two datasets.

In a specific data cleansing case study, the author describes a technique for standardizing residential and business addresses based on data rules established by the US Post Office. The author then proceeds to describe a number of general data cleansing and enhancement tools including: date/time, contextual, geographic, demographic, psychographic, and inferential data enhancement. An example of an inferential enhancement might be to assign a "primary decision maker" field to a household database based on the most frequent credit card user within the household.

Finally, the text summarizes each of the chapters as building blocks needed to build data quality practices for an enterprise. This book is a good primer on data quality concepts. It lists, in a systematic and formal way, many of the things that an actuary knows to look for intuitively in their work, but may not know how to articulate formally. While it is a long book, it is not an especially difficult read. It could be put to good use in constructing a checklist of data quality best practices that one would run through when building or implementing a new database or system architecture.

### 3.5 Corporate Information Factory

*Corporate Information Factory* [10] (ISBN 0-471-39961-2) provides an overview of information technology architecture for modern corporations. Its authors, Inmon (described by many in the industry as the father of the data warehouse), Imhoff and Sousa,

describe a way of thinking about various technologies available today to give the reader a structure to incorporate them in their company's systems. The authors feel their proposed approach is "the best way to meet the long-term goals of the information processing company." Two clear strengths of their approach are that it can be implemented incrementally and it is designed to be flexible to adapt to changing business needs.

The book is divided into four parts. The first two chapters summarize the evolution of the "corporate information factory." Chapters 3 through 14 review each element of the architecture and how they are combined. Chapters 15 to 17 discuss constructing and managing the corporate information factory. Finally, the appendix contains guidelines for examining and assessing a particular corporate information factory.

The authors write: "Three fundamental business pressures are fueling the evolution of the information ecosystem: growing consumer demand, increased competition and complexity, and continued demands for improvements in operating efficiencies..." The corporate information factory can help corporations respond to these pressures by aiding them in:

- **Business operations:** running the day-to-day business,
- **Business intelligence:** helping companies understand what drives their business and the likely impact of decisions, and
- **Business management:** "If business intelligence helps companies understand what makes the wheels of the corporation turn, business management helps direct the wheels as the business landscape changes."

The authors see the big picture as follows. "The alpha and omega of the corporate information factory is the external world in which business is transacted." Information flows from the external world to the data acquisition applications of the corporate information factory. From there it can be condensed into operational reports or transformed and integrated with other data before being forwarded to primary storage management. Primary storage management includes the operational data store (ODS) and the data warehouse including historical data. The final phase, data delivery, can include data marts, decision support services and an exploration warehouse or a data mining warehouse. Managing metadata (information about the data) embraces and integrates across all three phases of the corporate information factory: data acquisition, primary storage, and data delivery.

The authors look at each of the dozen components from several points of view:

- What is the purpose or function of the component?
- What is its structure?
- How does information flow?
- What types of data does it work with?
- What types of users use it?
- What is the level of centralization versus decentralization in processing?
- How does this component interface with others?

The concepts of ecology and evolution are used frequently in the book. Corporate information systems are like an ecosystem where raw energy in the form of data is transformed by organisms, i.e., the various component information systems, into "food" or output which is then recycled into other "organisms" or information systems. These systems are never static, but evolve over time as circumstances and requirements of the users change.

Corporate information resides in a number of different data stores. These include data warehouses, data marts and operational data stores, each with their own role in meeting corporate information needs. The data warehouse is a big data repository of much of the company's data that is needed to run business intelligence systems. A data mart is a subset data repository, used for specific functions and applications containing smaller data subsets and aggregations of data. It is needed for efficiently running applications. Both play an important role in managing corporate information needs. Finally, an ODS "is a collection of detailed data that satisfies the collective, integrated operational needs of the corporation." The focus of the ODS is on information for operations, so it only contains current detailed information, not a data warehouse's multiple snapshots and summaries.

Each of the various corporate data stores has a development life cycle involving requirement gathering, analysis, design, programming, testing, and implementation. The book discusses a general database management strategy, as well as the strengths and weaknesses of various software and hardware solutions.

The different kinds of data storage may have different management requirements. For instance, the data warehouse needs are "characterized by volumes of data and unpredictable

workload" (p. 251). The authors discuss the various needs and how they are addressed, how the various data stores are integrated as well as the security needs associated with the different databases.

A further consideration of corporate information systems is archival of stale data. The authors discuss how long management should wait before archiving data and what the best mechanism is for archiving it

The authors also include a discussion of multiple data warehouses and the integration of data from multiple systems. Integration of data from separate systems can be necessitated by corporate mergers and acquisitions or by the need to do more advanced analysis. Such integrations contain their own challenges, such as who owns the data, who creates and manages the new database, what types of data the database contains, and the nature of the sharing that will occur.

The book is something like the "Cliff Notes" of information management. It is a concise summary of current theory and practice with respect to developing and maintaining information systems, but it is not weighted down with a lot of technical detail. As such, the book provides an easy-to-read introduction for those not working in the area but might be too simplistic for those with deep experience in data management and information systems. The text is clearly written, but because of the multi-faceted approach sometimes it is difficult to tell where the authors are going with a discussion. Acronyms often appear in the book and their frequency sometimes becomes annoying. Also, a lot of the diagrams are trivial: they don't really illustrate their point any better than the text. Finally, despite all the points of view, there does not seem to be a lot of actionable information: this is a good text for learning about concepts, but not for implementing them.

*Corporate Information Factory* provides a good introduction to the broad world of information technology. This book can help actuaries better understand IT structure, concepts, issues, and goals to better frame their interactions with IT. If you are interested in a quick introduction to the topic that covers the key concepts and techniques, this book will meet your goals. If you need a more substantial introduction to information management, reference another book, perhaps one of the many books referenced by the authors.

### 3.6 Data Quality, the Field Guide

The focus of the book *Data Quality, the Field Guide* [11] (ISBN: 1-55558-251-6), by

Thomas C. Redman, Ph.D., is on data quality programs and efforts inside organizations. This book provides many constructive approaches to establishing or improving the data quality programs in businesses. It is a "how to" manual for those new to data management and a great refresher to those who have been in the field for a while. As the quality of the work product that an actuary produces depends so much on the quality of the input, with data being one of the key ingredients, the topic of this book should be of high interest to actuaries.

The book first reinforces that all disciplines and levels in an organization have a stake in quality data. The author presents the viewpoints of various stakeholders from the CEO to the customers of the organization.

For an actuary who has had the responsibility for data management and/or data quality in their organization, this book brings no surprises. But the nice feature for the experienced data manager – actuary or non-actuary – is the well-organized presentation of the issues with many charts and logical pictorial diagrams of data quality concepts and data quality processes to illustrate the author's points.

For those actuaries who regularly encounter quality issues in data supplied from internal or external systems and who are starting out in the area of data management and data quality, this book should be required reading. It quickly presents many concepts that a new data manager needs to know and the author presents them succinctly on a high level. Again, the illustrations and diagrams will help solidify the concepts quickly and can be adapted by readers to their own situations. Your adaptations of his charts and diagrams to a business case plan for data quality improvements will lend an authoritative flavor to your plans. It is a book worth reading for actuaries who have interactions with those responsible for data in their organizations. The knowledge and insights gained by the reader will help put them on equal footing with those who are responsible for the data.

One very important point that the author makes is that clean-ups of a database do not scale; you need to fix the source or cause of the data quality issue. "Organizations must recognize that finding and fixing data errors is time consuming, expensive, non-value-added work." Otherwise resources will be forever dedicated to cleaning a database and the problem will never go away. As the author says, "any form of clean-up without prevention is wrong-headed." In Section E of the book, the author describes the elements of a successful data quality system and how those tasks are accomplished. While you may not want to follow his solution or methods exactly, the book does provide a lot of ideas to consider as you work to

improve data quality in your organization. The author makes the point that you need to consider all costs of errors – the immediate cost and the cost to those downstream; and you need to know where errors occur.

A "statistical control process" is described in Chapter 23. The process as presented is ever vigilant, focusing on continuous improvement and bottom line impact.

Another point made by the author is that I. (information or data) is not IT (information technology). It is clear from the book that one should not use IT to automate a poorly designed information chain. An information chain, as defined by the author, is an "end-to-end process that starts with original data sources creates 'information products' and continues through to the use of data in operations, decision-making, and planning." First improve the information chain and then automate using IT to reduce cost and to free up people for other tasks; IT plays a subordinate role.

The author also presents other concepts in the book that may be more applicable to the data manager such as a business case plan for data quality; the competitive advantage derived from quality data; techniques for cleaning a database; and the common elements of a successful data quality program. Reading through these sections should help practicing actuaries improve their communications with the data managers in their organizations.

Throughout the book, the author presents seventy-one "tips." For ease of reference sixteen of the most important tips are repeated at the end of the book and reorganized according to several subjects. A glossary of terms is also provided at the end of the book.

Overall the book is a quick read and presents many concepts in an easy to understand fashion for the practicing actuary.

### 3.7 Data Management: Databases and Organization

*Data Management: Databases and Organization, fifth edition* [12] (ISBN 0-47171-536-0) by Richard T. Watson is an introductory data management text. It focuses on the core skill of data modeling using SQL (structured query language) to implement the data models. It also covers such topics as the managerial perspective of data management, database architecture, emerging technologies, and data integrity.

Overall, this text is very well written. The topics are self contained, although the concepts of data modeling and SQL run throughout, so those sections should not be skipped. For actuaries, it is probably best to use the text as a reference book on particular

topics because of the length (approximately 600 pages). Watson divides his book into five sections, and a brief synopsis of each follows.

Section 1, "The Managerial Perspective," defines the concept of organizational memory which includes not only computers, but also people, paper files, manuals, reports, etc. He also draws distinctions between data, information, and knowledge. According to Watson, "data are raw, unsummarized, and unanalyzed facts," while "information is data that have been processed into a meaningful form." Finally he states that "knowledge is the capacity to use information." Watson makes the interesting point that the preceding perspectives on data and information are relative. One person's information is another person's data.

In section 2, "Data Modeling and SQL," Watson considers data modeling and SQL skills as fundamental to data management. As such he devotes approximately half of the book to this topic. The style of this section is very straightforward and should be accessible to any actuary with some exposure to relational databases, such as Microsoft Access, SQL Server, Oracle, etc. He goes through in detail the basic building blocks of data modeling: modeling a single entity, one-to-many relationships, many-to-many relationships, one-to-one relationships, and recursive relationships.

The author repeatedly uses the same approach to explain new concepts, thus making the text easy to follow. First, he builds his examples using a standard data modeling diagramming syntax., Second, as each new modeling concept is introduced, a model is developed and then implemented in SQL. This is an effective technique for both data modeling and SQL since the concepts reinforce each other.

Watson also uses examples from standard relational databases such as Access and Oracle. While the book is not an Access reference and many advanced SQL features are not supported in Access, the text does give a good indication of the theoretical underpinnings about how a relational database product such as Access should be used. The text is filled with numerous exercises on both data modeling and SQL. It is a good primer for those actuaries that are interested in moving beyond Access or doing advanced database work using the macro programming capabilities of Access.

The author thoroughly illustrates the concept of normalization as a method for increasing the quality of a database design. He goes through the development of six normal forms and describes the issues that these normal forms resolve. This is perhaps a little advanced for most actuaries, but it is interesting reading if one is willing to devote the effort.

Finally, Watson provides an "SQL playbook" that contains 61 sample queries that should handle most of the data manipulation tasks that an actuary may encounter.

Section 3, "Database Architectures and Implementations," deals with more of the technical aspects of data management such as data structures and storage. It also provides an introductory background on data processing architectures such as client/server technology. If nothing else, this section and Section 4 define much of the terminology that is used in many IT shops today. This is of great use to actuaries that need to understand the key concepts of various technologies to liaise with their IT departments.

Watson devotes a chapter in this section to object-oriented (OO) data management. He does a good job of describing the object-oriented paradigm and then contrasting it with the relational-paradigm. Since the relational model is primarily used in data management, and the OO model is used primarily in software engineering, Watson posits that it is important to be able to translate between the two. Among the differences that he cites between the two paradigms is that the OO paradigm has its basis in the software engineering principles of coupling, cohesion, and encapsulation, while the relational paradigm is based on the mathematical concepts of set theory.

Section 4, "Organizational Memory Technologies," covers a potpourri of technologies. Watson devotes a chapter in this section that touches on data warehousing, data mining, and the multi-dimensional database (MDDB) or cube environment. Given that MDDB is (arguably) the best storage arrangement for actuarial triangles, this section should be of great interest to actuaries. Unfortunately, it barely scratches the surface on data warehousing and data mining. He also devotes two chapters to the Web and provides some extensive examples on how to use SQL within Java. Finally, he closes the section with a good treatment of XML (extensible markup language) and its emerging use as a data management standard.

The final section "Managing Organizational Memory" covers two topics that most actuaries should find of interest: data integrity and data administration. In this time when actuaries are being asked to become advocates for data quality, it is important for them to understand what data quality really means. Watson states that maintaining data integrity involves three goals:

1. Protecting the existence of the data so it is available whenever it is needed;
2. Maintaining the quality of the data so that it is accurate, complete, and current; and

3. Ensuring confidentiality of data so that only those authorized can access it.

He then describes many techniques to achieve these goals.

The author also covers what he calls the 18 dimensions of data quality. As an example, let's look at three of the dimensions—Accuracy, Timeliness, and Accessibility—and what conditions Watson sets for high quality (see Table 1).

Table 1

| **Dimension** | **Conditions for high quality data** |
|---|---|
| Accuracy | Data values agree with known correct values. |
| Timeliness | A value's recentness matches the needs of the most time critical application requiring it. |
| Accessibility | Authorized users can readily access data values through a variety of devices from a variety of locations. |

These three dimensions, as well as the other 15 dimensions outlined in the book, are an ongoing pursuit and not a destination. It is worthwhile for actuaries to look at all 18 dimensions and see how each of their organization's data stacks up against them.

Overall, I would highly recommend Data Management: Databases and Organization to those actuaries that are interested in learning more about the principles and challenges of data management.

## 3.8 Software Testing in the Real World

Edward Kit's main goal in this book [13] (ISBN 0-20187-756-2) is to prove to software companies that they need dedicated testing departments at least as big as their development departments. Given that actuaries are not in the business of making shrink-wrapped software packages for numerous outside customers, the "real world" in the title practically never intersects with the actuarial universe.

Some of the main thoughts of the book, however, will be of interest to actuaries. Considering that actuaries implement their models in software, this activity could be conceivably called "software development." Thus some notions of testing should not be fully foreign to actuaries; they just have to be adapted to the actuarial situation.

Testing according to the book should start from the "specifications" and end with the "final product" evaluation, and should be performed by an "outsider." Testing techniques range from verification to validation, i.e., from checking the "code" to examining "final

product" outcomes. In the actuarial paradigm, the final product could be an Excel spreadsheet, Mathematica notebook, or Oracle stored procedure. Correspondingly, specifications could be a reserve test or pricing method, and the "code" would be formulae in cells, VBA subroutines, or SQL statements. Evidently, checking everything from methods and assumptions to auditing spreadsheet formulae and query results makes perfect sense.

The content of Kit's book is broken into 4 parts. Part I includes chapters 1 through 3. The material in these chapters is somewhat esoteric. There is a lot of discussion about what is needed to get started on the testing of software and the history of software testing. These chapters would not be applicable to the actuarial science field.

Chapters 4 through 6, which form Part II of the book, establish a framework for conducting tests on software. This section establishes some decent terminology that one could use to test a student's familiarity with testing procedures. The question we need to ask is will everyone in the industry adhere to the same terminology? For example, in the 4th chapter, there are several terms used to establish a general failure in the software code. Such terms include: "mistake, fault, failure, [or] error." Could we get some of these terms generally accepted in the actuarial industry? There are several examples of these types of definitions of principles within this section. There is one principal in particular that could prove to be useful in the actuarial science field: "the purpose of testing is to discover errors." It is a nice short and sweet principal. Chapter 5 seems to be getting to some substance. One question that it attempts to answer is when a tester should be giving special attention to the testing process. Discussions about verification (checking the code) and validation (testing the program) are also discussed in Chapter 5. Chapter 6 is not very helpful to actuaries. This chapter seems to be regurgitating different top-down methods on how to approach testing. This is probably more useful to software engineers than to actuaries. At this stage of the book, some examples would have proven to be helpful. Several lists of questions are developed for testing methods but none of the questions are ever answered. It is unclear if we are supposed to be learning how to ask or how to develop questions. More testing standards are talked about in a theoretical sense but no lists of standard questions are given. The section on "Testware" (a collection of software tools for testing) is somewhat useful. This section discusses what is actually used to test software and calls for maintaining the best testware tools beyond the testing of a single product.

Part III, which includes chapters 7 through 12, provides several different testing methods. Some of the material can be applied to what we do in actuarial science. For example, the

methods used for verification could become a basis for technical reviews of an actuary's work. Still, the text lacks examples and exercises for the reader to follow. There do not appear to be definitive methods to apply to specific circumstances. The recommendations at the end of the chapters contain many phrases such as: "usually it is better to do..." or "there's a real trade-off when you do..." A decisive recommendation on a method to use in a particular situation would have been more helpful. There is a relevant exercise given on p. 67 of the book. It refers to documents in Appendices B and C. The exercise shows how verification testing can produce gains on developing software for a minimal amount of effort. Also, the section on how a tester should report an author's mistakes (in Chapter 7) is useful.

Part IV includes topics on structural designs for testing software, practices used by software engineers in testing, and getting gains from software testing. This section would not be applicable to the field of actuarial science.

The appendices follow these 4 parts, and are clearly the most useful part of the book to actuaries. There appears to be much more order and less theory in this section of the book. Appendix A gives lists of Software Testing Standards. This section may be very useful when a tester has to present results to a management team or to a group of people within the industry. For testing actuarial work, one could refer to similar standards much like we do for reserving and valuation methods. Appendix B gives good verification checklists. It is ironic that there is a functional design checklist which has a requirement to look out for designs "without examples or examples that are too few." The author could have taken this requirement and applied it to the earlier chapters in the book. Appendix B has a good deal of sample checklists which would be useful. Appendices C and D contain verification and validation exercises (respectively) and solutions that seem very useful, however extensions would be needed to translate the exercises into practical advice for Excel "developers." Appendix E contains a bibliography which is a good reference for guides on software testing. Appendix F gives source information on conferences, journals, and newsletters which may be useful for someone desiring more information on software testing. Appendix G gives a list of software technology used to check software. Appendix H contains a list of improvements in the area of terminology, product requirements, tools used for testing, and documentation which should be considered. The text should have referred to the lists and information in the appendices much more frequently.

In conclusion, actuarial practitioners who are heavily involved in spreadsheet design may

occasionally find some useful tidbits in this book. However there simply are not enough examples or case studies to make any of the testing methods easy to implement. Therefore actuaries not heavily involved in systems development should probably pass on this text and wait for a more directly applicable book or article on the subject. Please note that do not wish to minimize the importance of software errors to actuaries. However this book may not be the appropriate reference for the kinds of software development projects that actuaries encounter.

### 3.9 Insurance Data Collection and Reporting, eighth edition

This book [14] (ISBN 1-877796-27-1), edited by Rose Castro, is the first in a series of eight books published by the Insurance Data Management Association (IDMA) designed to educate data managers. As a textbook, it is well written and quite easy to follow. There are 10 chapters in total.

The first three chapters introduce underwriting and actuarial ratemaking, highlighting the necessity of high quality insurance data that underlie these functions. As the author rightly points out in the first chapter, both line underwriters and staff underwriters need data to perform their daily jobs. Moreover, actuaries rely heavily on data to analyze loss reserves and conduct rate level experience reviews. Chapter 2 discusses general ratemaking procedures widely used by property/casualty actuaries. These procedures include pure premium method, loss ratio method and distribution of an overall indication to territories/classes. Workers Compensation ratemaking, a different animal as usual, is elaborated in the third chapter. NCCI has three types of systems to perform ratemaking functions: the administered pricing system, the advisory rate system, and the loss cost system.

Chapters 4 to 9 focus on various types of statistical agents such as ISO, NAII, and NCCI. Chapter 4 gives a general background of insurance regulation and statistical reporting. Two important court decisions (Paul v. Virginia and South-Eastern Underwriters Association) and two laws (McCarren-Ferguson Act and All-Industry Rating Bills) are cited. These help readers understand the historical context in which insurance regulation has evolved. Chapter 4 also gives a high-level review of statistical agents. Chapter 5 summarizes various statistical agent reports and three basic report designs (annual statistic compilations, Fast Track Monitoring System, and accelerated reports). Chapter 6 gives a detailed description of ISO. Besides highlighting ISO's statistical plans, the author also touches upon the process that ISO goes through after receiving data. In chapter 7 and 8, the NAII and NCCI statistical

plans are described in detail. Chapter 9 identifies organizations specializing in data collection which do not fall into the above categories: mostly involuntary pools.

Chapter 10 focuses on state insurance departments including the history of insurance regulation regarding insurance data and state data needs.

Overall, this book provides excellent study material for data managers to get a good understanding of insurance data collection/reporting. Actuaries have learned most of the contents of this book through CAS exams. For them, this book not only gives a good review but also helps to piece together an understanding of data management to the insurance enterprise.

## 4. CONCLUSIONS

There is an actuarial standard of practice with respect to data quality and some actuaries have data management responsibilities, but there is almost nothing in actuaries' formal training to prepare them for these tasks. Furthermore, current CAS literature is comparatively cursory in its coverage of information quality topics. To fill this gap, these nine texts have been recommended to actuaries seeking more information on data quality or data management.

To help identify the best text for a specific situation, the texts are compared below in three ways. The first table (Table 2) describes the subjects covered in each book and should be helpful in determining which books are most appropriate for particular data quality and data management goals. In this table, five solid circles mean the particular topic is excellently covered in a way readily accessible to actuaries. Conversely, five empty circles mean the subject is either barely covered or addressed from a point of view that is of limited use to actuaries. Finally, a blank rating means the particular subject is not covered at all in the particular text.

Table 2: Coverage of Topics

| Author | Section | Data Quality | Principles of Data Quality | Metadata | Exploratory Data Analysis | Data Audits |
|---|---|---|---|---|---|---|
| Olsen | 3.1 | ●●●●○ | ●●●●○ | ●●●○○ | ●●○○○ | ●●●○○ |
| Dasu | 3.2 | ●●●●○ | ●●●○○ | ●●●○○ | ●●●●○ | ●○○○○ |
| English | 3.3 | ●●○○○ | ●●●◖○ | ●●●◖○ | | |
| Loshin | 3.4 | ●●●○○ | ●●●○○ | ●●●○○ | ●●●○○ | ●●○○○ |
| Inmon | 3.5 | ●○○○○ | ●●○○○ | ●●●○○ | | |
| Redman | 3.6 | ●●●○○ | ●●○○○ | | ●●●○○ | |
| Watson | 3.7 | ●●●●○ | ●●●●○ | ○○○○○ | | |
| Kit | 3.8 | ○○○○○ | | | ○○○○○ | ○○○○○ |
| IDMA | 3.9 | ○○○○○ | ○○○○○ | | ○○○○○ | ○○○○○ |

| Author | Section | Processing Quality | Presentation Quality | Measuring Data Quality | Data Quality Improvement Strategies | Data Management | Statistical Plans |
|---|---|---|---|---|---|---|---|
| Olsen | 3.1 | ●●●●○ | | ●●●●○ | ●●●●○ | ●●○○○ | |
| Dasu | 3.2 | ●●○○○ | | ●●●○○ | ●●●○○ | ●●○○○ | |
| English | 3.3 | | ○○○○○ | ●●○○○ | ●●○○○ | | |
| Loshin | 3.4 | ●●○○○ | ●○○○○ | ●●●○○ | ●●○○○ | ●●●●○ | ○○○○○ |
| Inmon | 3.5 | | | | ●○○○○ | ●●○○○ | |
| Redman | 3.6 | | ●●○○○ | ●●○○○ | ●●●○○ | ●●●○○ | |
| Watson | 3.7 | | | ●○○○○ | ●○○○○ | ●●●●○ | |
| Kit | 3.8 | ●●○○○ | | ○○○○○ | ○○○○○ | | |
| IDMA | 3.9 | ○○○○○ | | ○○○○○ | ○○○○○ | ●●○○○ | ●●●●○ |

Table 3: Definitions of Topics

| Topic | Definition / Description |
|---|---|
| Data Quality | What is it? Why does it matter? How to achieve it? |
| Principles of DQ | Key attributes of "quality data" |
| Metadata | Information about data, e.g. business rules |
| EDA | Statistical and graphical tests to identify suspicious values in a data set |
| Data Audits | Reconcile the data intended for use to its original source(s) |
| Processing Quality | Ensuring quality in models and software through design, implementation and testing |

| | |
|---|---|
| Presentation Quality | Clear, correct, consistent presentation of results |
| Measuring DQ | Statistics to track key attributes of quality |
| DQ Improvement Strategies | What should an organization do to determine the level of quality required and how to achieve it? |
| Data Management | The bridge between those who are responsible for the collection and repository of data and those who will use the data in analyses |
| Statistical Plans | Examples, motivation and uses of mandated data (not detailed instructions on specific statistical plans) |

The next table contains brief summaries of the dominant characteristics of each book.

Table 4: Synopsis

| Author | Section | Comment |
|---|---|---|
| Olsen | 3.1 | Well written, easy to follow data quality program for companies |
| Dasu | 3.2 | Good introduction to use of exploratory data analysis in data quality |
| English | 3.3 | Complete data quality guide aimed at IT and management rather than at actuaries |
| Loshin | 3.4 | Good generic data management text. Not specific to actuarial science, but covering many thorny data issues which actuaries may encounter. |
| Inmon | 3.5 | An easy to read introduction to concepts and systems architecture |
| Redman | 3.6 | Easy to follow book with concepts an actuary can easily pick up on |
| Watson | 3.7 | Very good text on data modeling and SQL |
| Kit | 3.8 | This book should only be used by actuaries who are involved with designing software. Even so, adaptation of any of the material will be needed prior to use. The appendices and the last few chapters are the most applicable to actuaries. |
| IDMA | 3.9 | Good introduction to data collection and various agencies |

The final table contains summary ratings by text. The ratings assigned provide an assessment of how suitable each book is at covering topics and what audience it is best suited for. For instance, each book is rated on whether it is geared towards beginners in information quality or at a more advanced audience that is already familiar with some of the literature. As another example, since insurance applications were not a focus of any of the books, each book is rated on its relevance to actuaries. Whereas Table 2 focuses on the topics covered in the book (such as data quality or metadata) Table 5 focuses primarily on qualities of the book as a whole that determine what audience it is best suited for (such as beginner/advanced, those wanting a more theoretical as opposed to practical knowledge, etc.). The book reviews in section 3 also contain information on the technical level of the

book and the audience it is written for. Note: although many of the books were written for an audience that is actively involved in data management or data quality, they also provide a good introduction to the topic for a general audience that interacts regularly with data supplied by others.

Table 5: General Characteristics

| Author | Section | (a) Actuarial Relevance | (b) Beginner / Advanced | (c) Practical / Theoretical | (d) Micro or Macro Focus | (e) Overall |
|---|---|---|---|---|---|---|
| Olsen | 3.1 | ●●●○○ | ●●○○○ | ●●●○○ | ●●●○○ | ●●●○○ |
| Dasu | 3.2 | ●●●○○ | ●●●●○ | ●●●●○ | ●●○○○ | ●●●○○ |
| English | 3.3 | ●○○○○ | ●●●○○ | ●●○○○ | ●●○○○ | ●◖○○○ |
| Loshin | 3.4 | ●○○○○ | ○○○○○ | ●●○○○ | ●●○○○ | ●●◖○○ |
| Inmon | 3.5 | ●○○○○ | ●●●●○ | ○○○○○ | ●○○○○ | ●●○○○ |
| Redman | 3.6 | ●●○○○ | ●●●○○ | ●●○○○ | ●○○○○ | ●●●○○ |
| Watson | 3.7 | ●●◖○○ | ●◖○○○ | ●●○○○ | ●●○○○ | ●●●○○ |
| Kit | 3.8 | ○○○○○ | ◖○○○○ | ●◖○○○ | ●●○○○ | ●○○○○ |
| IDMA | 3.9 | ●○○○○ | ●●●○○ | ●○○○○ | ●●○○○ | ●●○○○ |

Notes: Generally, five solid circles is most relevant to an actuarial analyst

(a) 5 solid circles = text is written for actuaries. 5 empty circles = need to modify or extend ideas in the text before an actuary could use them.

(b) 5 solid circles = beginner: no prior IT knowledge required. 5 empty circles = advanced, e.g. reader should have worked in the field.

(c) 5 solid circles = purely practical, e.g. a tip sheet with no reasoning behind the tips. 5 empty circles = purely theoretical.

(d) 5 solid circles = hands-on analyst advice such as a book of C programs. 5 empty circles = only high-level advice, e.g. strictly executive issues.

(e) 5 solid circles = a "must-read" for all actuaries. 4 solid circles = a "must-read" for actuaries with data management responsibilities. 5 empty circles = the information is not worth the time it takes to read it.

By reviewing the summary information in these tables, the reader may be able to identify candidate books that will best meet his or her needs.

The working party hopes that this paper will be a resource for actuaries dealing with data management and/or data quality issues. More information on these issues can be found at the idma.org web site.

**Acknowledgment**

The working party thanks IDMA for narrowing the field of data quality and data management texts to those they felt would be most relevant to actuaries. We also thank our IDMA liaisons for their feedback and support throughout this project.

The working party also thanks the Insurance Services Office, Inc., for the use of excerpts from their homeowners module of the ISO personal lines statistical plan (other than auto).

## 5. REFERENCES

[1] Francis, Louise A. "Dancing with Dirty Data: Methods for Exploring and Cleaning Data." CAS Forum Winter 2005: 198-254.

[2] Actuarial Standards Board of the American Academy of Actuaries. *Actuarial Standard of Practice No. 23: Data Quality,* revised edition. Schaumburg, Illinois: American Academy of Actuaries, 2004.

[3] CAS Committee on Management Data and Information. "White Paper on Data Quality." CAS Forum Winter 1997: 145-168.

[4] Popelyukhin, Aleksey. "Watch Your TPA: A Practical Introduction to Actuarial Data Quality Management". CAS Forum Winter 1999: 239-254.

[5] Copeman, P., Gibson, L, Jones, T., Line, N, Lowe, J, Martin, P., Mathews, P., Powell, D., "A Change Agenda for Reserving: A Report of the General Insurance Reserving Issues Task Force" 2006, www.actuaries.org.uk

[6] Olson, Jack E. *Data Quality: the Accuracy Dimension.* Morgan Kaufman, 2003.

[7] Dasu, Tamraprni and Theodore Johnson. *Exploratory Data Mining and Data Cleaning.* Wiley, 2003.

[8] English, Larry P. *Improving Data Warehouse and Business Information Quality.* New York: Wiley, 1999.

[9] Loshin, David. *Enterprise Knowledge Management.* Morgan Kaufman, 2001.

[10] Inmon, William and Claudia Imhoff and Ryan Sousa. *Corporate Information Factory, second edition.* New Jersey: Wiley, 2000.

[11] Redman, Thomas C. *Data Quality, the Field Guide.* Boston: Digital Press, 2001.

[12] Watson, Richard T. *Data Management: Databases and Organization, fifth edition.* New Jersey: Wiley, 2005.

[13] Kit, Edward. *Software Testing in the Real World.* New York: Addison-Wesley, 1995.

[14] Castro, Rose. *Insurance Data Collection and Reporting, eighth edition.* New Jersey: IDMA, 2005.

**Abbreviations and notations**

Collect here in alphabetical order all abbreviations and notations used in the paper

| | |
|---|---|
| ASB, Actuarial Standard Board | MDDB, Multi-dimensional Database |
| ASOP, Actuarial Standard of Practice | NAII, National Association of Independent Insurers |
| CAS, Casualty Actuarial Society | NCCI, National Council on Compensation Insurance |
| CIO, Chief Information Officer | NYSIIS, New York State Identification and Intelligence System |
| COLDQ, Cost of Low Data Quality | ODS, Operational Data Store |
| GIRO, General Insurance Research Organization | OLAP, On-Line Analytical Processing |
| GRIT, General insurance Reserving Issues Taskforce | OO, Object Oriented |
| IDMA, Insurance Data Management Association | SQL, Structured Query Language |
| ISO, Insurance Service Organization? | VBA, Microsoft Visual Basic Application |
| IT, Information Technology | XML, Extensible Markup language |

## Biographies of Working Party Contributors

**Robert Campbell** is Director, Commercial Lines Actuarial at Lombard Canada in Toronto, Canada. He has a Bachelor of Mathematics in Business Administration from the University of Waterloo. He is a Fellow of the CAS and a Fellow of the Canadian Institute of Actuaries. He is chair of the Data Management Educational Materials working party, participates on the CAS Committee on Data Management and Information, and was a participant on the 2006 GIRO Data Quality working party.

**Lijuan Zhang** is senior actuarial analyst at Insurance Service Office in New Jersey. She is responsible for ZIP Code based territory revising analysis and pricing. She has a degree in Economics from Youngstown State University. She is a Fellow of the CAS and a Member of the American Academy of Actuaries.

**Louise Francis** is a Consulting Principal at Francis Analytics and Actuarial Data Mining, Inc. She is involved in data mining projects as well as conventional actuarial analyses. She has a BA degree from William Smith College and an MS in Health Sciences from SUNY at Stony Brook. She is a Fellow of the CAS and a Member of the American Academy of Actuaries. She is serves on several CAS committees /working parties and is a frequent presenter at actuarial and industry symposia. She is a four-time winner of the Data Quality, Management and Technology call paper prize including one for "Dancing with Dirty Data: Methods for Exploring and Cleaning Data (2005)."

**Rudy Palenik** is the Commercial Actuary at Westfield Insurance Group in Westfield Center, Ohio. He is responsible for the development of rates for all the commercial lines of business. He has a degree in Math from Marquette University in Milwaukee, Wisconsin and is a Fellow of the Casualty Actuarial Society and a member of the American Academy of Actuaries. Rudy participates on a number of CAS committees including: Data Management and Information, Actuarial Education and Research Foundation, Research Paper Classifier and University Liaison.

**Aleksey Popelyukhin** is a Vice-President of Information Systems with the 2 Wings Risk Services and a Head of Quantitative Analytics Group with the Wall Street North Consulting in Stamford, Connecticut. He holds a Ph.D. in Mathematics and Mathematical Statistics from Moscow University (1989). Aleksey actively participates in CAS research and is frequent presenter on CAS conferences. CAS recognized Aleksey's contributions by awarding him the very first prize in "Data Management" papers competition and inviting him to the very first Working Party (on presentation of DFA/DRM results). In addition to numerous publications Aleksey helps to advance actuarial science by building convenient software tools for actuaries such as Triangle Maker®, Affinity and Actuarial Toolchest™. For those actuaries having troubles explaining statistics to the management Aleksey built a DRM presentation template available from CAS website. And for those who have troubles fitting clean models to dirty data Aleksey developed advanced data quality service called Data Quality ShieldSM. Aleksey is currently developing an integrated pricing/reserving/DRM computer system for reinsurance called "SimActuary" and also an action/adventure computer game tentatively called "Actuarial Judgement."

**Gregory Scruton** is Senior Vice President, Actuarial and Planning at Middlesex Mutual Assurance Company in Middletown, CT. He is responsible for pricing, reserving, planning and management information. He has a degree in Mathematics from Renssalaer Polytechnic Institute in Troy, NY. He is a Fellow of the CAS and a Member of the American Academy of Actuaries. He has participated on the CAS examination committee, and currently participates on the CAS Committee on Data Management and Information.

**Virginia R. Prevosto** is Principal, Consulting at Insurance Services Office, Inc. Ms. Prevosto is a Phi Beta Kappa graduate of the State University at Albany with a Bachelor of Science degree in Mathematics, *summa cum laude*. She is a Fellow of the CAS and a Member of the American Academy of Actuaries. She serves as General Officer of the CAS Examination Committee and as liaison to various other CAS admission committees. She also serves on the CAS Committee on Management Data and Information. In the past Ms. Prevosto also served on the Data Quality Task Force of the Specialty Committee of the Actuarial Standards Board that wrote the first data quality standard of practice for actuaries. Virginia has been a speaker at the Casualty Loss Reserve

Seminar on the data quality standard and to various insurance departments on data management and data quality issues. Ms. Prevosto authored the paper "Statistical Plans for Property/Casualty Insurer" and "Study Note: ISO Statistical Plans" and co-authored "For Want of a Nail the Kingdom was Lost – Mother Goose was right: Profit by Best (Data Quality) Practices" for the IAIDQ.

**Dave Hudson** is an Actuary for St. Paul Travelers in Hartford, CT. He has a MS degree in Mathematics from Washington State University in Pullman, WA. He is a Fellow of the CAS and a Member of the American Academy of Actuaries. He is also a member of the CAS Committee on Data Management and Information.

**Keith Allen** is the associate actuary for United Educators and is responsible for underwriting duties within the public school sector and general corporate actuarial issues. Allen has 13 years of experience in the insurance industry as an underwriter, claims adjuster, and actuary. Keith previously worked for Tillinghast-Towers Perrin as an actuarial specialist where he did reserving, pricing, and forecasting for various public and private entities. Prior to that, Allen worked as a claims adjuster and underwriter for State Farm Insurance where he helped develop the "Reinspection Program" used to assess coastal risks. Before joining the insurance industry, Allen was a teacher at Bellaire High School in Houston, TX. Allen holds a bachelor's degree in mathematics from the University of Texas and is an Associate of the Casualty Actuarial Society.

**Shiwen Jiang** is assistant vice president/actuary at Arch Insurance Group in New York. He is responsible for technical pricing. Prior to this, he has held various positions in Arch Insurance Group, The Hartford and W.R. Berkely Corp. He has a M.S. degree in Statistics from the Pennsylvania State University. He is a Fellow of the CAS and a Member of the American Academy of Actuaries. He participates on the CAS Committee on Data Management and Information and Asian Regional Committee.