INFORMATION AND DATABASE QUALITY

The Kluwer International Series on
ADVANCES IN DATABASE SYSTEMS

Series Editor
Ahmed K. Elmagarmid

Purdue University
West Lafayette, IN 47907

Other books in the Series:

THE FRACTAL STRUCTURE OF DATA REFERENCE: ***Applications to the Memory Hierarchy,*** *Bruce McNutt;* ISBN: 0-7923-7945-4

SEMANTIC MODELS FOR MULTIMEDIA DATABASE SEARCHING AND BROWSING, *Shu-Ching Chen, R.L. Kashyap, and Arif Ghafoor*; ISBN: 0-7923-7888-1

INFORMATION BROKERING ACROSS HETEROGENEOUS DIGITAL DATA: A Metadata-based Approach, *Vipul Kashyap, Amit Sheth*; ISBN: 0-7923-7883-0

DATA DISSEMINATION IN WIRELESS COMPUTING ENVIRONMENTS, *Kian-Lee Tan and Beng Chin Ooi*; ISBN: 0-7923-7866-0

MIDDLEWARE NETWORKS: Concept, Design and Deployment of Internet Infrastructure, *Michah Lerner, George Vanecek, Nino Vidovic, Dad Vrsalovic;* ISBN: 0-7923-7840-7

ADVANCED DATABASE INDEXING, *Yannis Manolopoulos, Yannis Theodoridis, Vassilis J. Tsotras;* ISBN: 0-7923-7716-8

MULTILEVEL SECURE TRANSACTION PROCESSING, *Vijay Atluri, Sushil Jajodia, Binto George* ISBN: 0-7923-7702-8

FUZZY LOGIC IN DATA MODELING, *Guoqing Chen* ISBN: 0-7923-8253-6

INTERCONNECTING HETEROGENEOUS INFORMATION SYSTEMS, *Athman Bouguettaya, Boualem Benatallah, Ahmed Elmagarmid* ISBN: 0-7923-8216-1

FOUNDATIONS OF KNOWLEDGE SYSTEMS: With Applications to Databases and Agents, *Gerd Wagner* ISBN: 0-7923-8212-9

DATABASE RECOVERY, *Vijay Kumar, Sang H. Son* ISBN: 0-7923-8192-0

PARALLEL, OBJECT-ORIENTED, AND ACTIVE KNOWLEDGE BASE SYSTEMS, *Ioannis Vlahavas, Nick Bassiliades* ISBN: 0-7923-8117-3

DATA MANAGEMENT FOR MOBILE COMPUTING, *Evaggelia Pitoura, George Samaras* ISBN: 0-7923-8053-3

MINING VERY LARGE DATABASES WITH PARALLEL PROCESSING, *Alex A. Freitas, Simon H. Lavington* ISBN: 0-7923-8048-7

INDEXING TECHNIQUES FOR ADVANCED DATABASE SYSTEMS, *Elisa Bertino, Beng Chin Ooi, Ron Sacks-Davis, Kian-Lee Tan, Justin Zobel, Boris Shidlovsky, Barbara Catania* ISBN: 0-7923-9985-4

INDEX DATA STRUCTURES IN OBJECT-ORIENTED DATABASES, *Thomas A. Mueck, Martin L. Polaschek* ISBN: 0-7923-9971-4

INFORMATION AND DATABASE QUALITY

edited by

Mario G. Piattini

Coral Calero

Marcela Genero

University of Castilla-La Mancha
Spain

SPRINGER SCIENCE+BUSINESS MEDIA, LLC

Library of Congress Cataloging-in-Publication Data

Information and database quality / edited by Mario G. Piattini, Coral Calero, Marcela Genero.
p. cm. -- (The Kluwer international series on advances in database systems ; 25)
Includes bibliographical references and index.
ISBN 978-1-4613-5260-0 ISBN 978-1-4615-0831-1 (eBook)
DOI 10.1007/978-1-4615-0831-1
1. Database management. 2. Databases--Quality control. I. Piattini, Mario, 1966- II. Calero, Coral, 1968- III. Genero, Marcela, 1966- IV. Series.

QA76.9.D3 I523 2001 2001050340

Printed on acid-free paper.

Contents

LIST OF CONTRIBUTORS (ALPHABETICALLY)

BOBROWSKI, Monica (Chapter 7)
She received the M.S. in Computer Science from Escuela Superior Latinoamericana de Informática (ESLAI) (1991). She currently works as a consultant for Pragma Consultores, specializing on data quality and project management. She is also an assistant professor of computer science in the Department of Computer Science at University of Buenos Aires. She has authored several papers that have appeared in international conferences. Her currently research interests include data quality, software testing, software metrics, and project management. She is active in technology transfer to industry.

BOUZEGHOUB, Mokrane (Chapter 8)
Professor at the University of Versailles in France. He is the Director of the database group in the PRiSM laboratory. His research interests are in database design, data integration, data warehouses, workflows, and software engineering. He is co-editor in chief of the *International Journal in Networking and Information Systems*. He has published different books on databases and object technology. His e-mail address is Mokrane.Bouzeghoub@prism.uvsq.fr

CALERO, Coral (Chapter 4)
MSc and PhD in Computer Science. Assistant Professor at the Escuela Superior de Informática of the Castilla-La Mancha University in Ciudad Real. She is a member of the Alarcos Research Group, in the same University, specialized in Information Systems, Databases and Software Engineering. Her research interests are: advanced databases design, database quality, software metrics, database metrics. She is author of articles and papers in national and international conferences on this subject. She belongs to the ATI association and is member of its Quality Group . Her e-mail is: ccalero@inf-cr.uclm.es

ENGLISH, Larry, P. (Chapter 5)
President and principal of INFORMATION IMPACT International, Inc., is an internationally recognized speaker, teacher, consultant, and author in information quality improvement. He has provided consulting and education in more than 25 countries on five continents. He was featured as one of the "21 Voices for the 21st Century" in the January, 2000 issue of *Quality Progress*. DAMA awarded him the 1998 "Individual Achievement Award" for his contributions to the field of information

resource management. He has organized and chaired 8 Information and Data Quality Conferences in the US and Europe since 1997.
Mr. English's methodology for information quality improvement—Total Quality data Management (TQdM®)—has been implemented in several organizations worldwide. He writes the "Plain English on Data Quality" column in the *DM Review*. Mr. English's widely acclaimed book *Improving Data Warehouse and Business Information Quality*, has been translated into Japanese.

GENERO, Marcela F. (Chapter 2)
Assistant Professor at the Department of Computer Science in the University of Comahue, in Neuquén , Argentine. She received her MS degree in Computer Science from the National University of South, Argentine in 1989. Actually, she is a PhD student at the University of Castilla-La Mancha, in Ciudad Real, Spain. Her research interests are: advanced databases design, software metrics, object oriented metrics, conceptual data models quality, database quality. Her e-mail address is mgenero@inf-cr.uclm.es

KAHN, Beverly K. (Chapter 9)
Associate Professor in the Sawyer School of Management at Suffolk University. She received her Ph.D. from the University of Michigan. Dr. Kahn's research concentrates on information quality, information resource management, database design and data warehousing. Her publications have appeared in leading journals such as MIS Quarterly, Journal of Management Information Systems, Communications of the ACM and Database. Her methodologies have been applied in organizations such as AT&T, Bell Atlantic, Fleet Financial and U.S. Department of Defense. She can be reached at bkahn@acad.suffolk.edu.

KATERATTANAKUL, Pairin (Chapter 3)
He is assistant professor in the Computer Information Systems Program at the Western Michigan University. He received his Ph.D. in Management Information Systems and Master of Arts in Marketing from the University of Nebraska – Lincoln. His research and teaching interests are in electronic business, marketing aspects of electronic commerce, networking, management information systems, information systems discipline, and information systems research.

KEDAD, Zoubida (Chapter 8)
Associate Professor at the University of Versailles in France. She received a PhD. from the Univeristy of Versailles in 1999. Her works mainly concerns database design, specifically schema integration issues and the design of multisource information systems and datawarehouses. Her e-mail address is Zoubida.Kedad@prism.uvsq.fr

LOSHIN, David (Chapter 6)
President of Knowledge Integrity Incorporated (www.knowledge-integrity.com), a consulting and product-development company focusing on knowledge management and information quality. David, who has an M.S. in computer scinece from Cornell University, is the author of three books, the most recent being "Enterprise Knowledge Management - The Data Quality Approach" (Morgan Kaufmann, 2001), and the others focusing on scalable high performance computing. David currently is driving the development of a rule-based data quality and business rule validation system to be used for measuring and managing levels of data quality throughout an interconnected information system.

MARRÉ, Martina (Chapter 7)
She received the M.S. in Computer Science from Escuela Superior Latinoamericana de Informática (ESLAI) (1991), and the Ph.D. in Computer Science from University of Buenos Aires (1997). She is currently an assistant professor of computer science in the Department of Computer Science at University of Buenos Aires. She has authored several papers that have appeared in international conferences and journals. Her currently research interests include data quality, software testing, and software metrics. She is active in technology transfer to industry. Since 1997, she has worked as a consultant, specializing on data quality and software testing.

PAUTKE, Robert, W. (Chapter 1)
Mr. Robert W. Pautke is the executive vice president of Navesink Consulting Group and is based in Cincinnati, Ohio. Bob is an expert at defining and implementing data supplier management programs. Most organizations acquire critical data from external sources, so supplier management is integral to their data quality programs.
Bob has almost two decades of experience in data management, process management and re-engineering. Prior to joining Navesink, Bob led AT&T's first data quality projects, including highly successful efforts with data suppliers. He spent two years at the AT&T Bell Laboratories Data Quality Lab. Bob has consulted with firms in the United States,

Europe and the Pacific Rim helping them gain real business value from improvements in data quality. Bob is published and holds a patent in the field of data quality. He joined Navesink in 1998.

PIATTINI, Mario G. (Chapter 2 and 4)
MSc and PhD in Computer Science by the Politechnical University of Madrid. Certified Information System Auditor by ISACA (Information System Audit and Control Association). Associate Professor at the Escuela Superior de Informática of the Castilla-La Mancha University. Author of several books and papers on databases, software engineering and information systems. He leads the ALARCOS research group of the Department of Computer Science at the University of Castilla-La Mancha, in Ciudad Real, Spain. His research interests are: advanced database design, database quality, software metrics, object oriented metrics, software maintenance. His e-mail address is mpiattin@inf-cr.uclm.es

REDMAN, Thomas C. (Chapter 1)
Dr. Thomas C. Redman is President of Navesink Consulting Group, based in Little Silver, NJ. Known by many as the "guru of data quality," Tom started Navesink in 1996 to help organizations improve their data and information, thereby improving decision-making, increasing customer satisfaction, and lowering cost. Navesink clients include telecommunications, financial services, computer products, dot-coms, and consumer goods companies. Tom's clients find that data are at the heart of everything they do and the need for the highest quality data is paramount. Many clients have reduced expenses by several million dollars per year by improving data quality. Among the first to recognize the need for high-quality data in the Information Age, Tom conceived the Data Quality Lab at AT&T Bell Laboratories in 1987 and led it until 1995. There he created the Applied Research Program that produced the first methods of improving data quality.
Tom holds a Ph.D. in statistics from Florida State University. He is the author of numerous papers, including "Data Quality for Competitive Advantage" (*Sloan Management Review*, Winter 1995) and "Data as a Resource: Properties, Implications, and Prescriptions" (*Sloan Management Review,* Fall 1998). Dr. Redman has written three books, *Data Quality: The Field Guide*, (Butterworth-Heinemann, 2001), *Data Quality for the Information Age* (Artech, 1996) and *Data Quality: Management and Technology* (Bantam, 1992) and was invited to

contribute two chapters to *Juran's Quality Handbook, Fifth Edition* (McGraw Hill, 1999). Tom holds two patents.

SIAU, Keng (Chapter 3)
J.D. Edwards Professor and an Associate Professor of Management Information Systems (MIS) at the University of Nebraska, Lincoln (UNL). He is also the Editor-in-Chief of the Journal of Database Management. He received his Ph.D. degree from the University of British Columbia (UBC) where he majored in Management Information Systems and minored in Cognitive Psychology. His master and bachelor degrees are in Information and Computer Sciences. He has published more than 35 refereed journal articles and these articles have appeared in journals such as Management Information Systems Quarterly, Communications of the ACM, IEEE Computer, Information Systems, ACM's Data Base, Journal of Database Management, Journal of Information Technology, International Journal of Human-Computer Studies, Transactions on Information and Systems, Quarterly Journal of E-Commerce, and many others. In addition, he has published over 55 refereed conference papers in proceedings such as ICIS, ECIS, WITS, and HICSS. He served as the Organizing and Program Chairs for the International Workshop on Evaluation of Modeling Methods in Systems Analysis and Design (EMMSAD) (1996 ? 2001). He has also published two books and 7 book chapters. For more information about him, please refer to his personal website at http://www.ait.unl.edu/siau/

STRONG, Diane M. (Chapter 9)
Associate Professor in the Management Department at Worcester Polytechnic Institute. She received her Ph.D. in Information Systems from Carnegie Mellon University. Dr. Strong's research centers on data and information quality and on MIS application systems, especially ERP systems. Her publications have appeared in leading journals such as Communications of the ACM, ACM Transactions on Information Systems, Journal of Systems and Software, Journal of Management Information Systems, and Information & Management. She can be reached at dstrong@wpi.edu.

YANKELEVICH, Daniel (Chapter 7)
He received the M.S. in Computer Science from ESLAI (1988), and the Ph.D. in Information Technology from Pisa University, Italy (1993). He is co-Founder and Senior Partner of Pragma Consultores, a firm focused on Software Quality and Software Engineering. He is CTO of Dolphin Interventures, an investment firm. He was and is involved in software

development and implementation projects for at least 10 of the 100 Fortune companies in the South Cone region. He is also an Associated Professor of software engineering at the University of Buenos Aires. He has authored several IT articles, published by top specialized journals.

PREFACE

Nowadays, in a global and increasingly competitive market, organisations are driven by information. Data and information are considered their main asset, and CIOs are looking for ways to transform data into true knowledge, which could secure the survival of the organisations. Most organisations have discovered how critical information is to the success of their businesses, however, few of them have effective ways of managing the quality of this information, which is so important to their competitiveness.

In fact, until a few years ago, quality issues were focused on program (ISO 9126, measures for COBOL programs, testing and inspection techniques, etc.) and software process quality (CMM, SPICE, Bootstrap, etc.) but information quality issues were disregarded. During the last decade databases and datawarehouses have become the essential core of information systems, and therefore their quality must be improved as much as possible in order to guarantee successful information systems.

Quality is a relative (the importance of different features varies among stakeholders and over time) and a multidimensional concept, it is therefore important to consider different issues related to information quality (see figure 1).

We can refer to information quality in a wide sense, comprising database/datawarehouse (DB/DW) system quality and data presentation quality. In fact, it is very important that data in the DB/DW reflects correctly the "real world", that is, that data is accurate; but it is also very important that data can be easily and unambiguously understood. DB/DW system quality depends both of the quality of the different processes involved in the construction of the DB/DW: design, loading, collection, transformation, updating, exploitation, etc. and of the quality of the different products of the DB/DW system. Three main products could be identified: the Database Management Systems (DBMSs), the data models (at the conceptual, logical and physical levels) and the data (values) itself.

The main purpose of this book is to provide an overview of some of these issues, covering their organisational and technical aspects. Space limitations prevented us from dealing with each topic in depth or to include others. Readers who want more information about them could consult the references of each chapter.

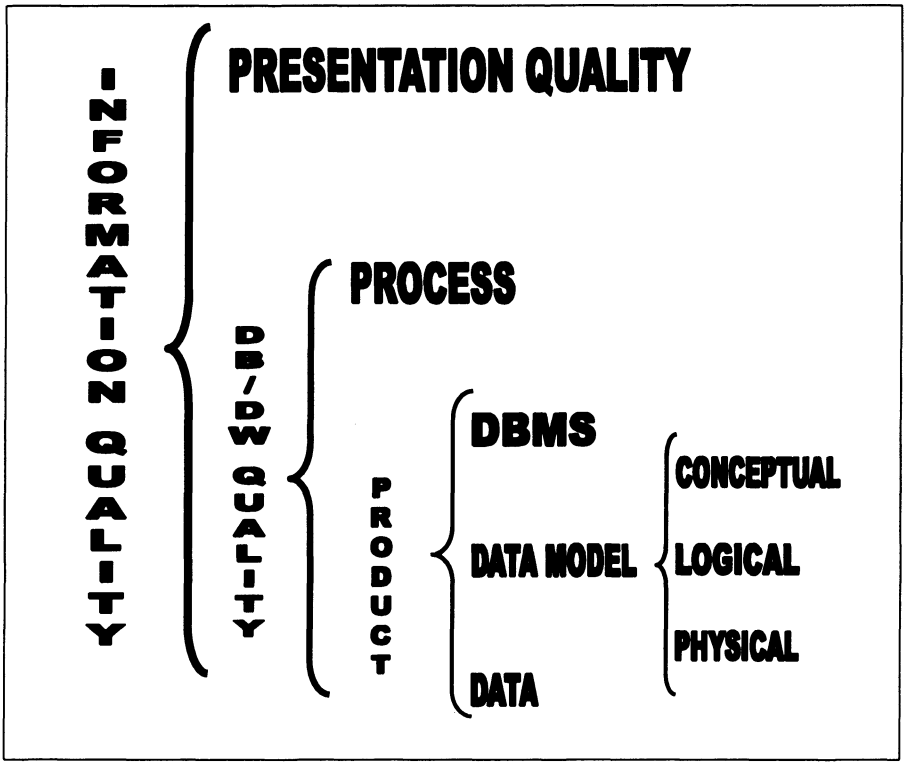

Figure 1. Information Quality Components

The book is divided in 9 chapters. Chapter 1 identifies and discusses the ten most important data issues facing the typical organization at the dawn of the new millennium. There is little debate that data and information (and their more esoteric brethren, knowledge and wisdom) are the critical assets of the Information Age.

Chapter 2 presents the different existing proposals, which deal with the issue of conceptual model quality, looking at the strengths and weaknesses of each one with the aim of providing the reader with a broad insight into the work already done and that which has to be carried out in the field of quality in conceptual modelling. This will help us to get a more comprehensive view of the direction work in this field is taking.

Chapter 3 seeks to develop a framework identifying the key features and facilities of Internet commerce Web sites. Once developed, the framework will enable an assessment of the Web site's design and information quality against a standard set of key characteristics or features.

Chapter 4 gives a series of guidelines which allow us to learn how metrics can be developed, in such a way that they can be used to achieve a specific objective related to the quality database design.

Chapter 5 provides a description of the TQdM® methodology for information quality improvement. It defines what information quality is, why it is essential to the survival of organizations in the Information Age. It describes the processes required to assess and improve information quality in order to achieve business performance excellence. It describes a process for implementing culture change required to achieve a sustainable environment of continuous information quality.

Chapter 6 explores a framework for defining data quality and business rules that qualify data values within their context, as well as the mechanism for using a rule-based system for measuring conformity to these business rules.

In chapter 7 the NEAT methodology is presented. This methodology provides a systematic way of assessing data quality. The methodology is quite simple, and can be applied when data quality should be evaluated and improved. The core part of NEAT is that of deriving metrics to evaluate data quality. The outcome of this work is a suitable set of metrics that establishes a starting point for a systematic analysis of data quality.

Chapter 8 provides a general framework for data warehouse design based on quality.

Chapter 9 examines in detail the mismatch between the information quality skills needed by organizations and the skills taught by universities to future IS professionals and makes recommendations on closing the gap and improving IQ teaching and learning, suggesting improvements to the IS curriculum models.

The book is targeted at senior undergraduates and graduate students, to complement their database courses. Database and datawarehouse professionals, quality managers can also find an interesting overview of these topics and useful hints for their job. The prerequisites for understanding the book is a basic knowledge of databases and software engineering.

Mario Piattini
Coral Calero
Marcela Genero

August 2001

ACKNOWLEDGEMENT

This book compiles works of different authors who have provide their knowledge and experience (both in research and industry) in specific information quality areas. Very special thanks go to all of them for their patience and collaboration.

We also want also to thank Kluwer Academic Publishers and, particularly, Melissa Fearon for her help, her patience and her advice.

Mario Piattini
Coral Calero
Marcela Genero