DATA 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

DATA QUALITY

by

Richard Y. Wang

Mostapha Ziad

Yang W. Lee

KLUWER ACADEMIC PUBLISHERS
New York / Boston / Dordrecht / London / Moscow

eBook ISBN: 0-306-46987-1
Print ISBN: 0-792-37215-8

©2002 Kluwer Academic Publishers
New York, Boston, Dordrecht, London, Moscow

All rights reserved

No part of this eBook may be reproduced or transmitted in any form or by any means, electronic, mechanical, recording, or otherwise, without written consent from the Publisher

Created in the United States of America

Visit Kluwer Online at: http: www.kluweronline.com
and Kluwer's eBookstore at: http: www.ebooks.kluweronline.com

To our families ...

Table of Contents

Preface xiii

Chapter 1 1

Introduction 1

- *Fundamental Concepts* 2
  - Data vs. Information 2
  - Product vs. Infomation manufacturing 3
  - The Information Manufacturing System 4
  - Information Quality is a Multi-Dimensional Concept 4
  - The TDQM Cycle 5
- *A Frameworkfor TDQM* 6
  - Define IP 7
  - Measure IP 12
  - Analyze IP 13
  - Improve IP 14
- *Summary* 14
- *Book Organization* 14

Chapter 2 19

Extending the Relational Model to Capture Data Quality Attributes 19

- *The Polygen Model* 20
  - Architecture 21
  - Polygen Data Structure 22
  - Polygen Algebra 23
- *The Attribute-based Model* 25

Data structure 26
Data manipulation 27
QI-Compatibility and QIV-Equal *28*
Quality Indicator Algebra *29*
Data Integrity 31
Conclusion *32*

Chapter 3 **37**

Extending the ER Model to Represent Data Quality Requirements **37**

Motivating Example *38*
Quality Requirements Identification *39*
Requirements Modeling *40*
Modeling Data Quality Requirements *41*
Data Quality Dimension Entity 42
Data Quality Measure Entity 42
Attribute Gerund Representation 43
Conceptual Design Example 44
Concluding Remarks *47*

Chapter 4 **49**

Automating Data Quality Judgment **49**

Introduction *50*
Quality Indicators and Quality Parameters *51*
Overview *51*
Related Work *52*
Data Quality Reasoner *52*
Representation of Local Dominance Relationships 53
Reasoning Component of DQR 54
First-Order Data Quality Reasoner *57*
The Q-Reduction Algorithm 58
Q-Merge Algorithm 60
Conclusion *61*

Chapter 5 **63**

Developing a Data Quality Algebra **63**

Assumptions 64
Notation 65
Definitions 65
A Data Quality Algebra *69*
Accuracy Estimation of Data Derived by Selection 69
Worst case when error distribution is non-uniform 70
Best case when error distribution is non-uniform 71
Accuracy Estimation of Data derived by Projection 71
Worst case when error distribution is non-uniform 72
Best case when error distribution is non-uniform 73
Illustrative Example *73*
Conclusion *76*

Chapter 6 **79**

The MIT Context Interchange Project **79**

Integrating Heterogeneous Sources and Uses *80*
The Information Integration Challenge *81*
Information Extraction and Dissemination Challenges 82
Information Interpretation Challenges 82
Overview of The Context Interchange Approach *84*
Context Interchange Architecture 84
Wrapping 85
Mediation 86
Conclusion *91*

Chapter 7 **93**

The European Union Data Warehouse Quality Project **93**

Introduction *93*
The D WQ Project *95*
DWQ Project Structure 95

DWQ Objectives 96
Data Warehouse Components 97
The Linkage to Data Quality 98
D WQ Architectural Framework *100*
The Conceptual Perspective 101
The Logical Perspective 101
The Physical Perspective 102
Relationships between the Perspectives 103
Research Issues *103*
Rich Data Warehouse Architecture Modeling Languages 104
Data Extraction and Reconciliation 105
Data Aggregation and Customization 105
Query Optimization 106
Update Propagation 107
Schema and Instance Evolution 108
Quantitative Design Optimization 108
Overview of the DWQ demonstration *108*
Source Integration (Steps 1,2) 109
Aggregation and OLAP Query Generation (Steps 3,4) 113
Design Optimization and Data Reconciliation (Steps 5,6) 113
Summary and Conclusions *114*

Chapter 8 **119**

The Purdue University Data Quality Project **119**

Introduction *120*
Background *121*
Related Work *122*
Searching Process 122
Matching Process 123
Methodology *124*
Data Pre-processing 125
Sorting, Sampling & the Sorted Neighborhood Approach 126
Clustering and Classification 127
Decision Tree Transformation and Simplification 128
Experiments and Observations *130*
Conclusions *135*

Chapter 9 **139**

Conclusion **139**

Follow-up Research *140*
Other Research *143*
Future Directions *143*

Bibliography **149**

Index **163**

PREFACE

> *If you would not be forgotten,*
> *As soon as you are dead and rotten,*
> *Either write things worth reading,*
> *Or do things worth the writing.*
>
> Benjamin Franklin

This book provides an exposé of research and practice in data quality for technically oriented readers. It is based on the research conducted at the MIT Total Data Quality Management (TDQM) program and the work of other leading research institutions. It is intended for researchers, practitioners, educators and graduate students in the fields of Computer Science, Information Technology, and other inter-disciplinary fields. This book describes some of the pioneering research results that form a theoretical foundation for dealing with advanced issues related to data quality. In writing this book, our goal was to provide an overview of the cumulated research results from the MIT TDQM research perspective as it relates to database research. As such, it can be used by Ph.D. candidates who wish to further pursue their research in the data quality area and by IT professionals who wish to gain an insight into theoretical results and apply them in practice. This book also complements the authors' other well-received book *Quality Information and Knowledge* that deals with more managerially- and practice-oriented topics (Prentice Hall, 1999) and the book *Journey to Data Quality: A Roadmap for Higher Productivity* (in preparation).

The book is organized as follows. We first introduce fundamental concepts and a framework for total data quality management (TDQM) that encompasses data quality definition, measurement, analysis and improvement. We then present the database-oriented work from the TDQM program, work that is focused on the data quality area in the database field. Next, we profile research projects from leading research institutions. These projects broaden the reader's perspective and offer the reader a snapshot of some of the cutting-edge research projects. Finally, concluding remarks and future directions are presented.

We thank Cambridge Research Group for sponsoring this project and providing a productive work environment. We are indebted to Professor Stuart Madnick at MIT for his continuous, unreserved support and encouragement. The support of Dean Louis Lataif, Senior Associate Dean Michael Lawson and Professor John Henderson at Boston University's School of Management is appreciated. Also appreciated is the help and support from Dean John Brennan, Professors Warren

Briggs, Jonathan Frank, and Mohamed Zatet at Suffolk University. In addition, we appreciate the support from Dean Ira Weiss and Senior Associate Dean Jim Molloy at Northeastern University's College of Business Administration.

We thank John F. Maglio and Tom J. Maglio for their assistance in fulfilling many ofbook requirements such as the index and format. Thanks are also due to our colleagues and researchers whose work is included in this book, including Don Ballou, Ahmed Elmagarmid, Chris Firth, Jesse Jacobson, Yeona Jang, Henry Kon, Matthias Jarke, Harry Pazer, Stuart Madnick, Leo Pipino, M. P. Reddy, Giri Kumar Tayi, Diane Strong, Veda Storey, Yannis Vassiliou, Vassilios Verykios, and Yair Wand.

We must also thank the many practitioners who have made significant contributions to our intellectual life. Although it is not possible to list all ofthem, we would like to express our gratitude to Jim Funk at S.C. Johnson, Kuan-Tsae Huang at TASKCo.com, Ward Page at DARPA's Command Post of the Future program, Sang Park at Hyundai Electronics, Brent Jansen at IBM, Jim McGill at Morgan Stanley Dean Witter, Mel Bergstein, Alan Matsumura, and Jim McGee at Diamond Technology Partners, Jim Reardon at the Under-Secretary's Office of the Defense Department for Human Affairs, Gary Ropp and David Dickason at U.S. Navy Personnel Research, Studies, and Technology, Terry Lofton and Frank Millard at Georgia State's Public Health Division, Bruce Davidson and Alein Chun at Cedars Sinai Medical Center, Frank Davis and Joe Zurawski at Firstlogic, Madhavan Nayar at Unitech Systems Inc., and Les Sowa and Beam Suffolk at Dow Chemical.

In addition, we thank D.J. Chang at Dong-Nam Corporation, Kyung-Sup Lee at Korean Airlines, Dwight Scott and Karen O'Lessker at Blue Cross Blue Shield, Kary Burin at Hewlett Packard, Don Hinman and John Talburt at Axiom Corporation, Peter Kaomea at Citicorp, Lu-Yue Huang at Edgewave, Col. Lane Ongstad of the U.S. Air Force Surgeon General's Office, Lt. Col. Dave Corey of the U.S. Army Medical Command, Steve Hemsley, Dane Iverson and Raissa Katz-Haas at United Health Group, Sue Bies at the First Tennessee National Corporation, and Rita Kovac at Concept Shopping.

We are grateful for the contributing authors and publishers for granting us permission to reprint their materials in this book. Specifically, Chapter 1 is based, in part, on the paper, *A Product Perspective on Total Data Quality Management,* authored by Richard Wang and published the Communications of ACM in 1998. Chapter 2 is based, in part, on the paper, *A Polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective,* co-authored by Richard Wang and Stuart Madnick and published in the Proceedings of the International Conference on Very Large Database Conference in 1990. It is also based on the paper, reprinted from *Decision Support Systems,* 13, Richard Wang, M. P. Reddy, and Henry Kon, *Toward quality data: An attribute-based approach,* 349-372, 1995, with permission from Elsevier Science.

Chapter 3 has been contributed by Veda Storey and Richard Wang based on the paper, *An Analysis of Quality Requirements in Database Design,* that they co-authored and published in the Proceedings of the *1998 Conference on Information*

Quality. Chapter 4 is based, in part, on the M.I.T. CISL working paper, *A Knowledge-Based Approach to Assisting In Data Quality Judgment,* co-authored by Yeona Jang, Henry Kon, and Richard Wang. Chapter 5 is based, in part, on the M.I.T. TDQM working paper, *A Data Quality Algebra for Estimating Query Result Quality.*

The MIT Context Interchange Project in Chapter 6 has been contributed by Stuart Madnick based on the paper, Metadata Jones and the Tower of Babel: The Challenge of Large-Scale Semantic Heterogeneity, published in the Proceedings of 1999 IEEE Meta-Data Conference. The European Union Data Warehouse Quality Project in Chapter 7 has been contributed by Matthias Jarke and Yannis Vassiliou based on their paper "Data Warehouse Quality: A Review of the DWQ Project", published in the Proceedings of the 1997 Conference on Information Quality. The Purdue University Data Quality Project in Chapter 8 has been written by Vassilios S. Verykios under the direction of Ahmed Elmagarmid. It profiles one aspect of the on-going Purdue Univesity's data quality initiative. Originally, additional chapters based on the work conducted at Telcordia (formerly Bellcore), Georgia Mason University, The University of Queensland, Australia, the University of St. Gallen in Switzerland, Universidad de Buenos Aires in Argentina, and other institutions were planned. Unfortunately, we were unable to incorporate them into the current edition, but we certainly hope to do so in a future publication.

We appreciate Elizabeth Ziad for her love, support, and understanding, our children, Fori Wang, and Leyla, Ahmed, and Nasim Ziad who bring so much joy and happiness to our lives, as well as Abdallah, Abdelkrim, Ali and the Ziad family. In addition, we thank A. Russell and Arlene Lucid, Dris Djermoun, Kamel Youcef-Toumi, Hassan Raffa, Said Naili, Youcef Bennour, Boualem Kezim, Mohamed Gouali, Ahmed Sidi Yekhlef, Youcef Boudeffa, and Nacim Zeghlache for their help, support, and generosity. We also thank Jack and Evelyn Putman, and John, Mary Lou, Patrick, and Helen McCarthy for their love and support.

Last, but not least, we would like to thank our parents who instilled in us the love of learning.

Richard Y. Wang
Boston University, Boston, Massachusetts
& MIT TDQM program (http://web.mit.edu/tdqm/)
rwang@bu.edu

Mostapha Ziad
Suffolk University, Boston, Massachusetts
mziad@acad.suffolk.edu

Yang W. Lee
Northeastern University, Boston, Massachusetts
y.lee@nunet.neu.edu