Data-Centric Systems and Applications

Carlo Batini · Monica Scannapieca

Data Quality

Concepts, Methodologies and Techniques

With 134 Figures

Springer

Carlo Batini

Università di Milano Bicocca
Dipartimento di Informatica
Sistemistica e Comunicazione
Piazza dell'Ateneo Nuovo
20126 Milano
Italy
batini@disco.unimib.it

Monica Scannapieca

Dipartimento di Informatica e Sistemistica "A. Ruberti"
Università di Roma "La Sapienza"
Via Salaria 113
00198 Roma
Italy
monscan@dis.uniroma1.it

Library of Congress Control Number: 2006931550

ACM Computing Classification (1998): H.2, H.1

ISBN-10 3-540-33172-7 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-33172-8 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media
springer.com

Cover Design: KünkelLopka, Heidelberg
Typesetting: by the Authors
Production: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig

Printed on acid-free paper 45/3100/YL 5 4 3 2 1 0

To my wonderful children, Chiara, Laura, and Giulio.
Carlo

To Massimo and to my "Ernania" world.
Monica

Preface

Motivation for the Book

Electronic data play a crucial role in the information and communication technology (ICT) society: they are managed by business and governmental applications, by all kinds of applications on the Web, and are fundamental in all relationships between governments, businesses, and citizens. Because electronic data is so widely diffused, the "quality" of such data and its related effects on every kind of activity of the ICT society are more and more critical.

The relevance of data quality in both decisional and operational processes is recognized by several international institutions and organizations. As an example, the importance of data quality in decisional processes is clearly stated in the quality declaration of the European Statistical System [72], in which its mission is identified as follows: "We provide the European Union and the world with high quality information on the economy and society at the European, national, and regional levels and make the information available to everyone for decision-making purposes, research, and debate."

Furthermore, quality of data is also a significant issue for operational processes of businesses and organizations. The Data Warehousing Institute in a 2002 report on data quality (see [52]) shows that there is a significant gap between perception and reality regarding the quality of data in many organizations, and that data quality problems cost U.S. businesses more than 600 billion dollars a year.

The "Year 2000 problem", which led to modify software applications and databases using a two-digit field to represent years, has been a data quality problem. The costs to modify such software applications and databases have been estimated to be around 1.5 trillion US dollars (see [68]).

Some disasters are due to the presence of data quality problems, among them the use of inaccurate, incomplete, out-of-date data. For example, the explosion of the space shuttle Challenger is discussed in [78] according to a data quality perspective; the analysis reports more than ten different categories of data quality problems having a role in the disaster.

Such errors are motivations at the basis of the several initiatives that are being launched in the public and private sectors, with data quality having a leading role, as detailed in Chapter 1; the initiatives include, for instance, the Data Quality Act effected by the United States government in 2002 [149].

Electronic data are only to a certain extent of better quality than data stored in paper documents. Indeed, electronic data benefit from a defined and regulated representation, but processes that originate such data are often out of control, and consequently errors in data proliferate.

In the last decades, information systems have been migrating from a hierarchical/monolithic to a network-based structure, where the potential sources that organizations can use for the purpose of their businesses is dramatically increased in size and scope. Data quality problems have been further worsened by this evolution. In networked information systems, processes are involved in complex information exchanges and often operate on input obtained from other external sources, frequently unknown a priori.

As a consequence, the overall quality of the information that flows between information systems may rapidly degrade over time if both processes and their inputs are not themselves subject to quality control. On the other hand, the same networked information system offers new opportunities for data quality management, including the possibility of selecting sources with better quality data, and of comparing sources for the purpose of error localization and correction, thus facilitating the control and improvement of data quality in the system.

Due to the described above motivations, researchers and organizations more and more need to understand and solve data quality problems, and thus need answering the following questions: What is, in essence, data quality? Which techniques, methodologies, and data quality issues are at a consolidated stage? Which are the well-known and reliable approaches? Which problems are open? This book is an attempt to respond to all these questions.

Goals

The goal of this book is to provide a systematic and comparative description of the vast number of research issues related to quality of data, and thus to illustrate the state of the art in the area of data quality. While being a real problem in a vast number of activities in the private and public sectors, data quality recently resulted in a significant number of contributions to the research community. There are several international conferences promoted by the database and information system communities that have data quality as their main topic; the International Conference on Information Quality (ICIQ) [95], organized traditionally at the Massachusetts Institute of

Technology (MIT) in Boston, started in 1996; the International workshop on Information Quality in Information Systems (IQIS) [99], held in conjunction with the SIGMOD conference since 2004; the international workshop on Data and Information Quality (DIQ), held in conjunction with the Conference on Advanced Information Systems Engineering (CAiSE) since 2004 [98]; and the international workshop on Quality of Information Systems (QoIS), held in conjunction with the Entity Relationship (ER) conference since 2005 [100]. There are also national conferences, held in France, Germany, and the US.

On the practical side, many data quality software tools are advertised and used in various data-driven applications, such as data warehousing, and to improve the quality of business processes. Frequently, their scope is limited and domain dependent, and it is not clear how to coordinate and finalize their use in data quality processes.

On the research side, the gap, still present between the need for techniques, methodologies, and tools, and the limited maturity of the area, has led so far to the presence of fragmented and sparse results in the literature, and the absence of a systematic view of the area.

Furthermore, in the area of data quality we highlight the existence of a dichotomy, typical of many other research areas that have a deep impact on real life, between practice-oriented approaches and formal research contributions. This book tries to address such a dichotomy, providing not only comparative overviews and explanatory frameworks of existing proposals, but also original solutions that combine the concreteness of practical approaches and the soundness of theoretical formalisms. By understanding the motivations and the different backgrounds of solutions, we have figured out the paradigms and forces contributing to the data quality environment.

Our main concern in this book is to provide a sound, integrated, and comprehensive picture of the state of the art and of future evolutions of data quality, in the database and information systems areas. This book includes an extensive description of techniques which constitute the core of data quality research, including record matching, data integration, error localization, and correction; such techniques are examined in a comprehensive and original methodological framework. Quality dimension definitions and adopted models are also deeply analyzed, and differences between the proposed solutions are highlighted and discussed. Furthermore, while systematically describing data quality as an autonomous research area, we highlight the paradigms and influences deriving from other areas, such as probability theory, statistical data analysis, data mining, knowledge representation, and machine learning. Our book also provides very practical solutions, such as methodologies, benchmarks for the most effective techniques, case studies, and examples.

The rigorous and formal foundation of our approach to data quality issues, presented with practical solutions, renders this book a necessary complement to books already published. Some books adopt a formal and research-oriented approach but are focused on specific topics or perspectives. Specifically, Dasu and Johnson [50] approach data quality problems from the perspective of data mining and machine learning solutions. Wang et al. [206] provide a general perspective on data quality, by compiling a heterogeneous collection of contributions from different projects and research groups. Jarke et al. [104] describe solutions for data quality issues in the data warehouse environment. Wang et al. [203] is a survey of research contributions, including new methods for measuring data quality, for modeling quality improvement processes, and for organizational and educational issues related to information quality.

Some other books give much more room to practical aspects rather than to formal ones. In particular, leading books in the practitioners field are Redman' [167] and [169], and English' [68]. The two Redman' books provide an extensive set of data quality dimensions, and discuss a vast set of issues related to management methodologies for data quality improvement. English's book provides a detailed methodology for data quality measurement and improvement, discussing step-by-step issues related to data architectures, standards, process- and data-driven improvement methodologies, costs, benefits, and managerial strategies.

Organization

The book is organized into nine chapters. Figure 0.1 lists the chapters and details interdependencies.

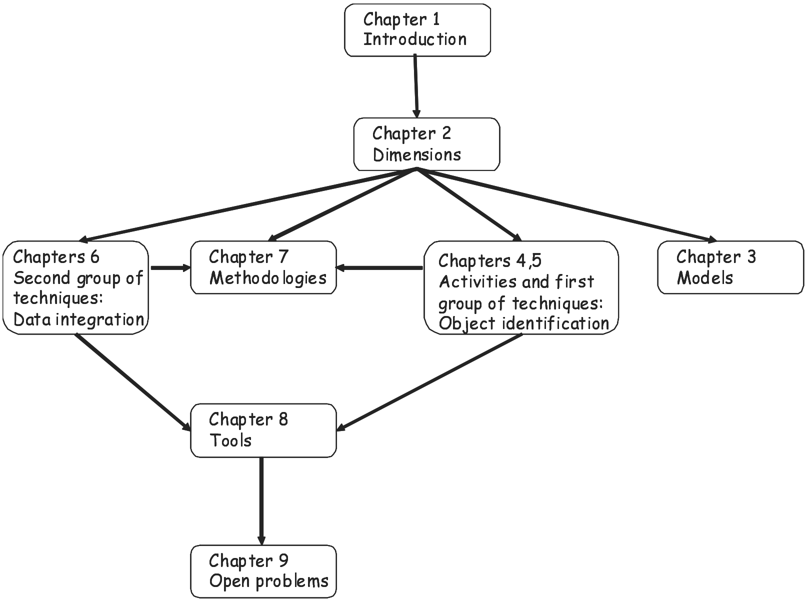

Fig. 0.1. Prerequisities among chapters

We initially provide basic concepts and establish coordinates to explore the area of data quality (Chapter 1). Then, we focus on dimensions that allow for the measurement of the quality of data values and data schemas (Chapter 2). These two chapters are preparatory to the rest of the book.

Models to express the quality of data in databases and information systems are investigated in Chapter 3. Chapter 4 describes the main activities for measuring and improving data quality. Some activities, such as error localization and correction, are introduced and fully described in Chapter 4; two specific chapters are dedicated to the most important activities and related research areas, namely object identification (Chapter 5) and data integration (Chapter 6), which are extensively investigated from the perspectives of relevant research paradigms and available techniques. Dimensions, models, activities, and techniques are the ingredients of any methodology for data quality measurement and improvement, and methodologies are the subject of Chapter 7. Specifically, in this chapter existing methodologies are examined and compared, and an original, comprehensive methodology is proposed, with an extensive case study. Tools, frameworks, and toolboxes proposed in the research literature for the effective use of techniques are described in Chapter 8. The book ends with Chapter 9, which puts all the ideas discussed in previous chapters in perspective and speculates on open problems and possible evolutions of the area.

Intended Audience

The book is intended for those interested in a comprehensive introduction to the wide set of issues related to data quality. It has been written primarily for researchers in the fields of databases and information systems interested in investigating properties of data and information that have impact on the quality of processes and on real life. This book introduces the reader to autonomous research in the field of data quality, providing a wide spectrum of definitions, formalisms, and methods, with critical comparisons of the state of the art. For this reason, this book can help establish the most relevant research areas in data quality, consolidated issues and open problems.

A second category of potential readers are data and information system administrators and practitioners, who need a systematization of the field. This category also includes designers of complex cooperative systems and services, such as e-Business and e-Government systems, that exhibit relevant data quality problems.

Figures 0.2 and 0.3 suggest possible paths, which can be followed by the above audiences.

The *researcher path*, for researchers interested in the core research areas in data quality, skips chapters on methodologies (Chapter 7) and tools (Chapter 8). The *information system administrator path* skips models (Chapter 3), data integration issues (Chapter 6) and open problems (Chapter 9).

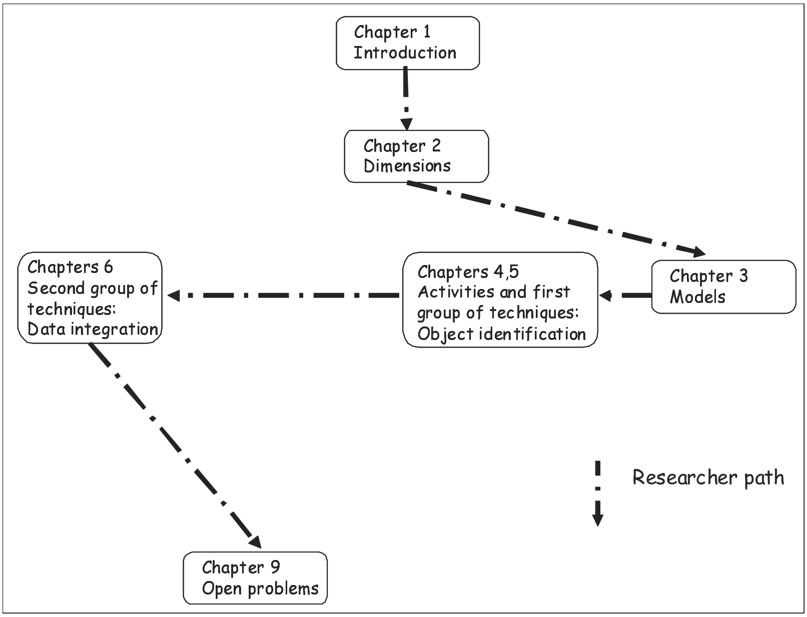

Fig. 0.2. Reading path for the researcher

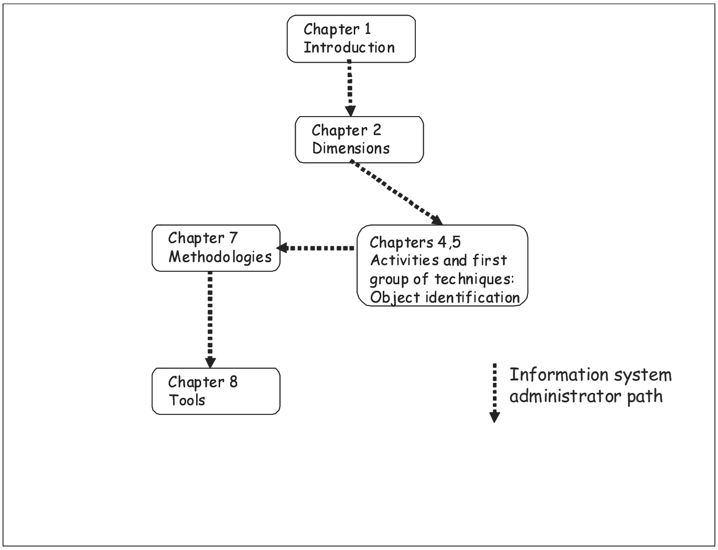

Fig. 0.3. Reading path for the information system administrator

Guidelines for Teaching

To the best of our knowledge, data quality is not a usually considered topic in undergraduate and graduate courses. Several PhD courses include data quality issues, while the market for professional, often expensive courses is rapidly increasing. However, recent initiatives are in the direction of introducing data quality in undergraduate and graduate courses [1]. We have organized the book to be used in an advanced course on the quality of databases and information systems. The areas of databases and information systems are currently lacking consolidated textbooks on data quality; we have tried to cover this demand. Although this book cannot be defined a textbook, it can be adopted, with some effort, as basic material for a course in data quality. Due to the undeniable importance of these topics, what happened in the 1980's for other database areas, e.g., database design, could happen for data quality: the plethora of textbooks which favored the introduction of this area in university courses.

Data quality can be the topic of self-contained courses, or else of cycles of seminars in courses on databases and information systems management. Data integration courses would also benefit from data quality seminars. With regards to information systems management, data quality can be taught in connection with topics such as information management, information economics, business process reengineering, process and service quality, and cost and benefit analysis. Data quality techniques can be offered also in specific courses on data warehousing and data mining.

The material of this book is sufficiently self-contained for students who are able to attend a course in databases. As students' prerequisites, it is useful, but not mandatory, to have notions of mathematics and, to some extent, probability theory, statistics, machine learning, and knowledge representation.

The book provides enough material to cover all the necessary topics without the need for other textbooks. In the case of a PhD course, the references are a good starting point for assigning students in-depth analysis activities on specific issues.

In terms of exercises, a useful approach for students is to develop a complex data quality project that can be organized into two parts. The first part could be devoted to the assessment of the quality of two or more databases jointly used in several business processes of an organization. The second part could focus on the choice and application of methodologies and techniques described in Chapters 4, 5, 6, and 7 to improve data quality levels of the databases to a fixed target. This approach gives students a taste of the problems to face within a real-life environment.

[1] As an example, in 2005 the University of Arkansas at Little Rock promoted a Master of Science in Information Quality (MS IQ).

Acknowledgements

First of all, we would like to thank several friends and researchers whose long-term frequentation and fruitful discussions on the topics covered in this book have positively influenced us; they are Daniele Barone, Laure Berti-Equille, Elisa Bertino, Paola Bertolazzi, Federico Cabitza, Cinzia Cappiello, Tiziana Catarci, Fabrizio De Amicis, Luca De Santis, Ahmed K. Elmagarmid, Markus Helfert, Domenico Lembo, Maurizio Lenzerini, Andrea Maurino, Massimo Mecella, Diego Milano, Paolo Missier, Felix Naumann, Davide Orlandi, Barbara Pernici, Louiqa Rashid, Ruggero Russo, Divesh Srivastava, Carlo Vaccari, and Richard Wang. Among students of our courses concerning data quality, we are particularly grateful to Matteo Strazzullo for his contributions on cost and benefit classifications.

Special mention should be given to the editors of this book series, Stefano Ceri and Mike Carey. We would also like to thank Ralf Gerstner of Springer-Verlag, who initially had the idea of considering data quality as a consolidated issue to be described in a book.

Carlo Batini
July 2006 *Monica Scannapieco*

Contents