Prologue: Research and Practice in Data Quality Management

Shazia Sadiq

Abstract This handbook is motivated by the presence of diverse communities within the area of data quality management, which have individually contributed a wealth of knowledge on data quality research and practice. The chapter presents a snapshot of these contributions from both research and practice, and highlights the background and rational for the handbook.

1 Introduction

Deployment of IT solutions, often following from strategic redirections, upgrades, mergers and acquisitions, is inevitably subjected to an evaluation of return on investment (ROI), which includes evaluation of the costs of sizable installations as well as the cost of changing the culture and work practice of all involved. It is often observed that the results of such analyses frequently indicate a failure to achieve the expected benefits [2]. A range of factors contributes to dismal ROIs, including significant factors rooted externally to the technological sophistication of the systems and often residing in the quality of the information the system manages and generates.

The issue of data quality is as old as data itself. However, it is now exposed at a much more strategic level, e.g. through business intelligence (BI) systems, increasing manifold the stakes involved for corporations as well as government agencies. For example, the Detroit terror case triggered an overhaul of the nation-wide watch list system, where lack of data propagation/consistency and issues with data freshness can be observed. The issue is equally important for scientific applications where lack of knowledge about data accuracy, currency or certainty can lead to catastrophic results. For example, the hurricane protection system in

S. Sadiq (✉)
The University of Queensland, Brisbane, Australia
e-mail: shazia@itee.uq.edu.au

S. Sadiq (ed.), *Handbook of Data Quality*, DOI 10.1007/978-3-642-36257-6_1,

New Orleans failed because it was "inadequate and incomplete", having been built disjointedly over several decades using outdated elevation data (New York Times, June 1, 2006). Further, the proliferation of shared/public data as on the World Wide Web and growth of the Web community has increased the risk of poor data quality usage for individuals as well. This is particularly alarming due to the diversity of the Web community, where many are unaware of data sources and data credentials. The situation is further complicated by presence of data aggregations and assimilations, e.g. through meta-search engines where source attribution and data provenance can be completely hidden from the data consumers.

One can also observe the changing nature of data quality management over the last decade or more. First, there are clear implications that relate to the sheer volume of data produced by organizations today. Second, recent years have seen an increase in the diversity of data. Such diversity refers to structured, unstructured and semi-structured data and multimedia data such as video, maps and images. Data also has an increasing number of sources. The use of various technologies, for example, sensor devices, medical instrumentation and RFID readers, further increases the amount and diversity of data being collected. More subtle factors also exist—such as the lack of clear alignment between the intention of data creation and its subsequent usage. A prime example of such lack of alignment is the vast amount of data collected from social networks that can then be used, without assessment of quality, as a basis for design and marketing decisions. Accordingly, a related factor exists that relates to difficulties in defining appropriate data quality metrics.

As these changes occur, traditional approaches and solutions to data management in general, and data quality control specifically, are challenged. There is an evident need to incorporate data quality considerations into the whole data cycle, encompassing managerial/governance as well as technical aspects. Currently, data quality contributions from research and industry appear to originate from three distinct communities:

Business analysts, who focus on ***organizational*** solutions. That is, the development of data quality objectives for the organization as well as the development of strategies to establish roles, processes, policies and standards required to manage and ensure the data quality objectives are met.

Solution architects, who work on ***architectural*** solutions. That is, the technology landscape required to deploy developed data quality management processes, standards and policies.

Database experts and statisticians, who contribute to ***computational*** solutions. That is, effective and efficient IT tools, and computational techniques, required to meet data quality objectives. Techniques in this regard can include record linkage, lineage and provenance, data uncertainty, semantic integrity constraints as well as information trust and credibility.

For the research community to adequately respond to the current and changing landscape of data quality challenges, a unified framework for data quality research is needed. Such a framework should acknowledge the central role of data quality in future systems development initiatives and motivate the exploitation of synergies across diverse research communities. It is unclear if synergies across the three contributing communities have been fully exploited. The sections below substantiate

this observation through an analysis of last 20 years of literature on data quality [14]. We argue that a unified framework for data quality management should bring together organizational, architectural and computational approaches proposed from the three communities, respectively.

2 Related Studies

A number of studies have addressed the issue of defining and analysing the scope of data quality research in the past. Owing to the cross-disciplinary needs of this area, identifying the central themes and topics and correspondingly the associated methodologies has been a challenge. In [10], a framework is presented that characterizes data quality research along the two dimensions of topics and methods, thereby providing a means to classify various research works. Previous works have also assisted by developing frameworks through which data quality research could be characterized, including a predecessor framework by the above group [17] that analogized data quality processes with product manufacturing processes. Some key research aspects such as data quality standardization, metrics/measurements and policy management emerged from these earlier works.

Other more recent studies have also provided valuable means of classification for data quality research. Ge and Helfert [5] have structured their review of the literature as IQ Assessment, IQ Management and Contextual IQ. Lima et al. [8] classify the literature between theoretical (conceptual, applied, illustrative) and practical (qualitative, experimental, survey, simulation) aspects. Neely and Cook [12] present their classification as a cross tabulation of Wang's framework [17] and Juran's original fitness for use factors [7].

The above studies provide various angles through which the body of knowledge can be classified and thus provide an essential means of understanding the core topics of data quality. However, understanding the intellectual corpus of a discipline requires not only an understanding of its core but also its boundaries [1]. As the realm of data quality has grown, so has the scope of its reference disciplines. With these factors in mind, we focused our study on understanding the interconnections and synergies across the various communities that contribute to data quality, rather than an identification of its central themes. The sections below substantiate this observation through an analysis of last 20 years of literature on data quality [14]. We argue that addressing the current challenges in data quality warrants such an understanding so synergies would be better exploited and holistic solutions may be developed.

3 Results of Literature Analysis

As a first step towards understanding the gaps between the various research communities, we undertook a comprehensive literature study of data quality research published in the last two decades [14]. In this study we considered a broad range of

Table 1 Considered publication outlets

	Includes[a]	Total
CS Conferences	BPM, CIKM, DASFAA, ECOOP, EDBT, PODS, SIGIR, SIGMOD, VLDB, WIDM, WISE	7,535
IS Conferences	ACIS, AMCIS, CAiSE, ECIS, ER, HICSS, ICIQ, ICIS, IFIP, IRMA, PACIS	13,256
CS Journals	TODS, TOIS, CACM, DKE, DSS, ISJ (Elsevier), JDM, TKDE, VLDB Journal	8,417
IS Journals	BPM, CAIS, EJIS, Information and Management, ISF, ISJ (Blackwell), ISJ (Sarasota), JMIS, JAIS, JISR, MISQ, MISQ Executive	2,493

[a]Due to space limitation, widely accepted abbreviations have been used, where full names are easily searchable via WWW

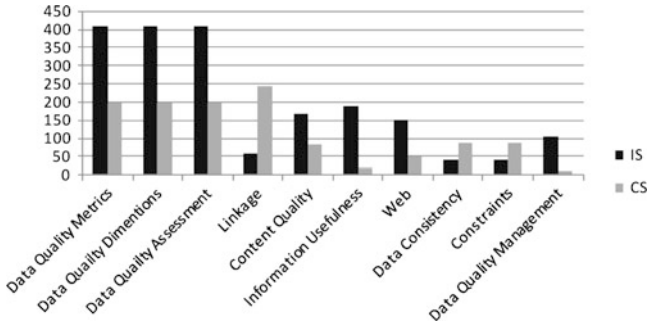

Fig. 1 Keyword frequency between IS and CS outlets

Information System (IS) and Computer Science (CS) publication (conference and journal) outlets (1990–2010) so as to ensure adequate coverage of organizational, architectural and computational contributions (see Table 1).

The main aims of the study were to understand the current landscape of data quality research, to create better awareness of (lack of) synergies between various research communities and, subsequently, to direct attention towards holistic solutions that span across the organizational, architectural and computational aspects (thus requiring collaboration from the relevant research communities).

In this section we present brief excerpts of the literature analysis conducted in [14] and [15] to provide a snapshot of the current research landscape in data quality management. As a consequence of the above studies, from the original data set of over 30,000 articles, a bibliographical database of almost 1500 publications (together with related keywords) was created through a rigorous analytical and reproducible methodology as detailed in [14].

The analysis revealed ***topics*** and ***venues*** of highest frequency as shown in Fig. 1 and Table 2, respectively. From the above, there is a clear indication that data quality themes are spread between IS and CS outlets. The overall distribution of papers between IS and CS outlets is summarized in Fig. 1. Clearly there are some topics

Table 2 Top publication frequencies with respect to publication venue

Publication outlet	# Pubs
International Conference on Information Quality (ICIQ)	241
Americas Conference on Information Systems (AMCIS)	152
International Conference on Very Large Databases (VLDB)	148
IEEE Transactions on Knowledge and Data Engineering (DKE)	120
ACM SIGMOD International Conference on Management of Data (SIGMOD)	116
ACM Transactions on Information Systems (TOIS)	51
Communication of the ACM (CACM)	49
Pacific Asia Conference on Information Systems (PACIS)	45
Hawaii International Conference on System Sciences (HICSS)	44
Symposium on Principles of Database Systems (PODS)	36
ACM Transactions on Database Systems (TODS)	35
International conference on Information Systems (ICIS)	34
European Conference on Information Systems (ECIS)	33
Australasian conference on Information Systems (ACIS)	33
Journal of Information & Management (IM)	27
ACM Special Interest Group on Information Retrieval (SIGIR)	27
International Conference on Extending Database Technology (EDBT)	22
International Conference on Database Systems for Advanced Applications (DASFAA)	20
Journal of Management Information Systems (MIS)	19
International Workshop on Information Quality in Information Systems (IQIS)	18
Journal of Information Systems Research (ISR)	12
Management Information Systems Quarterly (MISQ)	12
International Conference on Advanced Information Systems Engineering (CAISE)	10

where the overlap is greater (e.g. *Data Quality Metrics*) than others (e.g. *Linkage* and *Information Usefulness*).

Table 2 provides an alternative view for observing research activity in relation to prominent IS and CS publication venues. Obviously the International Conference on Information Quality (ICIQ) has the highest number of publications that span across a large number of keywords, with *Data Quality Assessment*, *Metrics* and *Dimensions* being the dominant ones. For AMCIS, in addition to the above keywords, *Information Usefulness* and *Content Quality* were also observed. Similarly, for VLDB as well as DKE journal, *Linkage* was the dominant keyword, closely followed by *Data Consistency* and data *Uncertainty*.

We further conducted a ***thematic*** analysis of the papers through a text-mining tool called Leximancer (www.leximancer.com). Leximancer performs a full text analysis both systematically and graphically by creating a map of the concepts and themes reappearing in the text. The tool uses a machine-learning technique based on a Bayesian approach to prediction. Leximancer uses concept maps to visualize the relationships. Each of the identified concepts is placed on the map in proximity of other concepts in the map through a derived combination of the direct and indirect relationships between those concepts (see Fig. 2). Concepts are represented by labelled and colour-coded dots. The size and brightness of a dot representing a

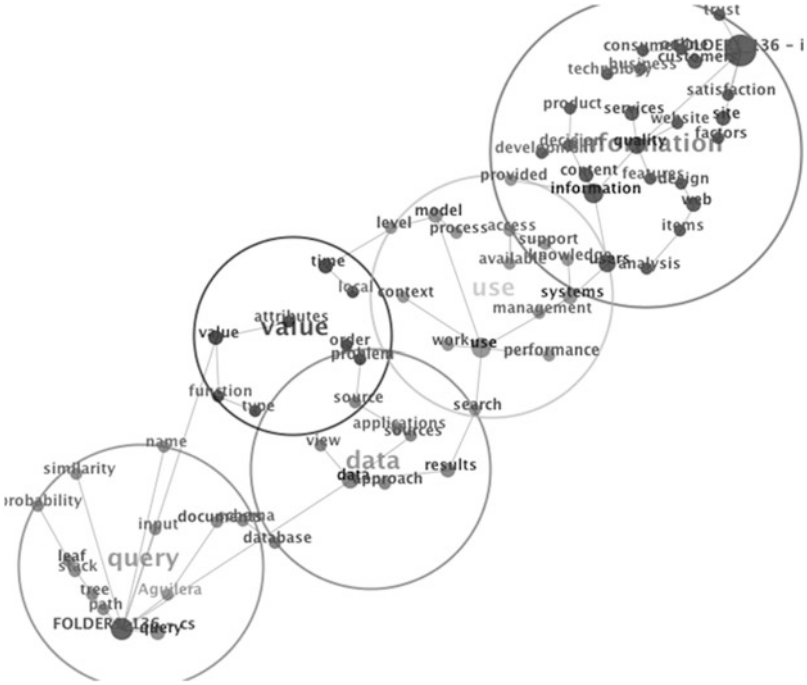

Fig. 2 Content quality—Information Systems vs. Computer Science publication focus

given concept on the map, is indicative of the concept's strength within the body of analysed text. The thickness and the brightness of connections between concepts are indicative of the frequency of co-occurrence of the two concepts. The relative distance of concepts on the map is indicative of similar conceptual contexts—i.e. the shorter the distance between the two concepts, the closer in context they are. Thematic clusters of highly connected concepts are indicated through coloured circles, called themes.

To explore synergies and differences between data quality research in the CS and IS disciplines, we conducted a series of Leximancer analyses for each of the top 10 keywords listed in Fig. 1. For each of the keywords, data was analysed considering the CS publications in isolation, then considering the IS publications in isolation, followed by a joint analysis of both data sets to gain a better understanding of the common focus of the two disciplines.

Due to space limitations, a detailed analysis is omitted here. However, as an example consider Fig. 2, where we selected the set of "Content Quality"-related publications in the CS and IS publication outlets for a joint analysis. Here the collective Information Systems data set related to the "Content Quality" topic is indicated by a "FOLDER-136-IS" concept. Likewise, the Computer Science data set is represented by "FOLDER-136-CS" concept. Specifically, it shows the relationships of concepts related to Content Quality across all considered

Table 3 Authors with more than 1,000 citations

Author	Citations	Author	Citations
Wang, R. Y.	4,364	McLean, E. R.	1,373
Widom, J.	2,774	Halevy, A.	1,308
Strong, D.	1,986	Lenzerini, M.	1,299
Ng, R. T.	1,894	Lee, Y. W.	1,183
Motwani, R.	1,847	Gibbons, P. B.	1,105
Datar, M.	1,739	Knorr, E. M.	1,071
Babcock, B.	1,685	Koudas, N.	1,061
Babu, S.	1,607	Chaudhuri, S.	1,056
Garofalakis, M. N.	1,428	Shim, K.	1,051
Rastogi, R.	1,378	Hellerstein, J. M.	1,014
DeLone, W.	1,373		

publication years and how the data set relates to concepts that were identified to be the strongest common concepts across the two data sets. Our analysis indicates that, while there are concepts that are common to both data sets, the strength of the connection is weak (while this is not visible in Fig. 2, due to resolution, the weakness is indicated in the Leximancer tool environment by the relative lack of thick, bright connections between both folder concepts and any one of the Content Quality concepts).

Indeed, the analysis uncovers strong evidence that the Information Systems set of papers is strongly focused on information quality, issues relating to satisfaction and business value in general, yet it is not as strongly focused (as indicated by the relative distance of the themes from each other and the relative closeness of the themes to each of the two publication sets) on approaches for ensuring content quality. While this is not surprising in itself, given that Information Systems is less technically oriented, we see a weakness in a situation where the communities that should be collaborating together appear to lack a strong collaboration and common focus.

We also conducted a ***citation*** analysis. For this purpose, we wrote a crawler script that searches all papers in the database within Google scholar and collects information regarding number of citations for the paper. In Table 3 we list the top cited authors. It is important to note that the citation counts are entirely based on the publications which are part of our collection and thus do not reflect the overall count for authors.

Some of the earliest contributions came from Wang, R. Y., Strong, D. and associates on the identification of *Data Quality Dimensions* and *Data Quality Assessment*. These contributions have been heavily utilized by later researchers as is evident from the high citation count above. Widom, J. and co-authors have contributed substantially to the body of knowledge on *data lineage* and *uncertainty* especially through the Trio system (see infolab.stanford.edu/trio). Similarly works of Ng, R. T. on identification of outliers in large data sets have applications in *error detection*, *entity resolution* and a number of data quality-related problems. Although it is not possible to summarize the contributions of all highly cited authors, it is safe to conclude that the contributions of these influential contributors are indicative of the wide span of data quality research.

4 The Three Pillars of Data Quality Management

The diversity and span of data quality research evident from the above-presented analysis of research literature from CS and IS publications is further exaggerated when we consider the vast experiential knowledge found in the practitioner and professional community within the information industry. Data quality management has been supported for last several decades by a number of highly active and experienced practitioners, including but not limited to [3, 9, 11, 13].

There have also been some industry-led initiatives that have attempted to identify key requirements or demands from industry in terms of data quality management [6]. The most relevant and recent of which is a job analysis report published by the International Association for Information and Data Quality (iaidq.org). The report provides data that assists in understanding and establishing the roles of data quality professionals in industry. Additionally, the report also identifies the body of knowledge required by those professionals to provide information/data quality services across various roles of an organization [18].

The contributions from various industry sources as above are inclined towards the *organizational* aspects of data quality management. For example, the industry-driven Information Quality Certification Program (www.iaidq.org/iqcp) covers six domains of (1) Information Quality Strategy and Governance, (2) Information Quality Environment and Culture, (3) Information Quality Value and Business Impact, (4) Information Architecture Quality, (5) Information Quality Measurement and Improvement and (6) Sustaining Information Quality. Although the organizational issues are an essential aspect of the overall space for data quality, it is also evident that lack of appropriate tools and systems to support organizational initiatives on data quality will undermine the best efforts of a dedicated team. This becomes especially apparent in the presence of large data sets, on one hand, and the complex dynamics of IT systems, enterprise software and legacy applications, on the other. There is a substantial body of knowledge that exists in support of such challenges, such as advanced record linkage, entity resolution, duplicate detection, managing uncertain data and data tracking and lineage. Most of these solutions are based on advanced *computational* techniques.

Finally, to state the obvious, there is a multibillion dollar data management market and commercial products and solutions that provide technology-related products and services across data(base) management, data integration and data analytics (including data warehousing and business intelligence solutions). Many of these vendors provide solutions directly related to data quality management [4]. These solutions provide the space in which many data quality solutions are deployed. Alignment between the organizational objectives and the technology *architecture* of deployed solutions is imperative.

In spite of the large body of knowledge stemming from research, practitioner and vendor communities, recent studies of data professionals indicate that a resounding 68 % of data quality problems are still detected due to complaints and/or by chance [16]. We argue that a key contributing issue is the segregated nature of the body of

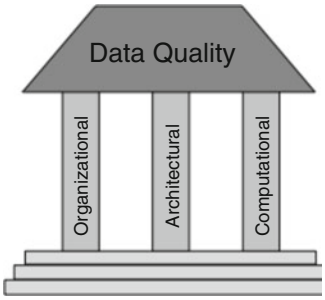

Fig. 3 The three pillars of data quality. *Organizational:* the development of data quality objectives for the organization, as well as the development of strategies to establish roles, processes, policies and standards required to manage and ensure the data quality objectives are met. *Architectural:* the technology landscape required to deploy developed data quality management processes, standards and policies. *Computational:* effective and efficient IT tools, and computational techniques, required to meet data quality objectives. Techniques in this regard can include record linkage, lineage and provenance, data uncertainty, semantic integrity constraints, as well as information trust and credibility

knowledge for data quality management and technology solutions. Next-generation solutions need to embrace the diversity of the data quality domain and build upon the three foundation pillars relating to organizational, architectural and computational aspects of data quality as depicted in Fig. 3.

As a simple example to illustrate the necessity of the three pillars, consider the following scenario:

> A large distribution company (LDC) that acquires two other distribution establishments, which will now form part of LDC operations as subsidiaries while maintaining their individual brandings. Each of the subsidiaries may have their own partner suppliers along with item catalogs. Consider the case that there is a large overlap of business with a particular supplier group, which may put LDC into a favorable bargaining position to negotiate significant discounts. However, data differences do not reveal this position, and thus directly impact on the bottom line for LDC.

In its simplest form a solution for the above scenario may be:

1. Create a reference (synonym) table for suppliers
2. Load supplier data from all subsidiaries into the reference table
3. Use matching techniques to identify potential overlaps
4. Extract a master table for suppliers—represents a single version of truth
5. Retain original representations—represent multiple versions of truth
6. Allow access for subsidiaries to reference master data in all new (or update) transactions involving supplier data
7. Ensure data managers are accountable for continued master data checks
8. Introduce a periodic monitoring scheme

Steps 1–5 require management intervention but at the same time require computational expertise specifically for step 3 and at the very least IT support for the

Table 4 Topics covered in the handbook

Prologue: Research and Practice in Data Quality Management			
Part I Organizational	Part II Architectural	Part III Computational	Part IV Data Quality in Action
Data Quality Management: history, frameworks, DQ projects and DQ programs (1, 2) Data Quality Costs (3) Governance and Maturity (4)	DQ issues for Data Warehouses (5) Role of Semantics and Ontologies for DQ (6) Data Quality Assessment and Error Detection (7)	DQ Rules and Constraints (8) Record Linkage, Duplicate Detection and Entity Resolution (9, 10) Managing Data Uncertainty (11) Data Provenance and Lineage Tracking (12)	Case Study presenting successful Data Integration (13) Case Study presenting a longitudinal process-oriented DQ initiative (14) Case Study focusing on creating a culture of Information Management (15)
Epilogue: The Data Quality Profession			

other steps. Step 6 demands that the company put in place the requisite technology (systems, networks, etc.) to enable appropriate access mechanisms and transactional consistency. Steps 7–8 are primarily management related although it is clear that previous steps will also need support of management, business teams and data owners.

5 Handbook Topics

The rationale for this handbook is motivated by the above analysis of research and practice in data quality management. The handbook is accordingly structured in to three parts representing contributions on organizational, architectural and computational aspects of data quality management, with the last part devoted to case studies of successful data quality initiatives that highlight the various aspects of data quality in action. The book concludes with a chapter that outlines the emerging data quality profession, which is particularly important in light of new developments such as big data, advanced analytics and data science. The four parts of the handbook and constituent topics (chapters) are summarized in Table 4.

Most chapters that focus on specific topics present both an overview of the topic in terms of historical research and/or practice and state of the art as well as specific techniques, methodologies or frameworks developed by the individual contributors.

Researchers and students from Computer Science, Information Systems as well as Business Management can benefit from this text by focusing on various sections relevant to their research area and interests. Similarly data professionals and practitioners will be able to review aspects relevant to their particular work. However, the biggest advantage is expected to emerge from wider readership of chapters that may not be directly relevant for the respective groups.

References

1. Benbasat I, Zmud RW (2003) The identity crisis within the IS discipline: designing and communicating the discipline's core properties. MIS Q 27(2):183–194
2. Carr N (2004) Does IT matter? Information technology and the corrosion of competitive advantage. Harvard Business School Press, Boston
3. English LP (2009) Information quality applied: best practices for improving business information processes and systems. Wiley, Indiana
4. Gartner Magic Quadrant for Data Quality. http://www.citia.co.uk/content/files/50_161-377.pdf. Accessed 15 Oct 2012
5. Ge M, Helfert M (1996) A review of information quality research. In: The 12th international conference on information quality. MIT, Cambridge, pp 1–9
6. Harte-Hanks Trillium Software 2005/6 Data Quality Survey (2006) http://infoimpact.com/Harte-HanksTrilliumSoftwareDQSurvey.pdf. Accessed 15 Oct 2012
7. Juran JM (1962) Quality control handbook
8. Lima LFR, Macada ACG, Vargas LM (2006) Research into information quality: a study of the state of the art in IQ and its consolidation. In: 11th international conference on information quality. MIT, Cambridge
9. Loshin D (2006) Monitoring data quality performance using data quality metrics. Informatica Corporation, Redwood City
10. Madnick SE, Wang RY, Lee YW, Zhu H (2009) Overview and framework for data and information quality research. J Data Information Qual 1(1):1–22
11. McGilvray D (2008) Executing data quality projects: ten steps to quality data and trusted information. Morgan Kaufmann, Burlington
12. Neely MP, Cook J (2008) A framework for classification of the data and information quality literature and preliminary results (1996–2007). AMCIS Proceedings. Accessed from http://www.citia.co.uk/content/files/50_161-377.pdf. Accessed 15 Oct 2012
13. Redman TC (1996) Data quality for the information age. Artech House, Boston
14. Sadiq S, Yeganeh NK, Indulska M (2011) 20 years of data quality research: themes, trends and synergies. In: Proceedings of the 22nd Australasian Database Conference (ADC 2011),, Perth, WA, Australia. 17–20 January 2011, pp 1–10
15. Sadiq S, Yeganeh NY, Indulska M (2011) An analysis of cross-disciplinary collaborations in data quality research. In: European conference on information systems (ECIS 2011), Helsinki, Finland, 2011
16. Sadiq S, Jayawardene V, Indulska M (2011) Research and industry synergies in data quality management. In: International conference on information quality (ICIQ2011), Adelaide, Australia, 18–20 November, 2011
17. Wang RY, Storey VC, Firth CP (1995) A framework for analysis of data quality research. IEEE Trans Knowledge Data Eng 7(4):623–640
18. Yonke CL, Walenta C, Talburt JR (2011) The job of the information/data quality professional. International Association for Information and Data Quality. Available from http://iaidq.org/publications/yonke-2011-02.shtml. Accessed 15 Oct 2012

http://www.springer.com/978-3-642-36256-9

Handbook of Data Quality
Research and Practice
Sadiq, S. (Ed.)
2013, XII, 438 p., Hardcover
ISBN: 978-3-642-36256-9