Data Quality on the Web

(Objectives and Goals of the Seminar)

Michael Gertz, University of California at Davis, U.S.A.
Tamer Özsu, University of Waterloo, Canada
Gunter Saake, University of Magdeburg, Germany
Kai-Uwe Sattler, University of Magdeburg, Germany

1 Aims and Scope

The aim of this paper is to provide participants of the seminar with some background information on the general notion of data quality (DQ), to illustrate a few agreed-upon and frequently used concepts and definitions, and to detail open problems to be addressed during the seminar. The paper is not meant as complete and comprehensive overview of all aspects related to data quality in the context of databases, information systems or the Web. It is rather intended to provide the participants with a basis for the seminar, outline specific foci, and raise questions and problems to be addressed extensively (and hopefully solved) during the seminar.

In the following Section 2, we will summarize some basic settings, definitions, and concepts commonly used in the context of data quality. We will also give some references to relevant literature that discusses these aspects in more detail. In Section 3, we detail a list of questions that arise when data quality is of concern. In Section 4, we then propose some application domains and scenarios in which these questions are to be studied and solutions are to be developed. Both the list of DQ questions and application domains and settings are by no means complete, but should illustrate the depth and breadth we expect data quality aspects in different settings to be covered. In Section 5, we summarize the objectives and outcomes to be taken into account by the working groups.

2 The Various Meanings of Data Quality

Compared to core database concepts such a database integrity and security, which have been studied in detail since the introduction of relational database technology, the notion of *data quality* has only emerged during the past 10 years and shows a steadily increasing interest. A major reason for this is the increase in interconnectivity among data producers and data consumers, mainly spurred through the development of the Internet and various Web-based technologies. More than ever before businesses, governments, and research organizations rely on the exchange and sharing of various forms of data. Oftentimes, data is the most valuable asset of an organization. It is also widely recognized that dealing with data quality problems can be very expensive and time consuming, leading to new IT technology branches that exclusively focus on the assessment of data quality in an organization and cleaning poor quality data.

One can probably find as many definitions for data quality as there are papers on data quality. As stated in [10], information quality (or data quality) is “an inexact science in terms of assessments and benchmarks”. Oftentimes, high-quality data is simply described as “data that is fit for use by data consumers” [21]. It will be a major objective of this seminar to analyze and develop precise (formal) data quality definitions for specific application scenarios that are of practical relevance and importance. As a guideline for developing such definitions, we suggest the use a few conventional characterizations of data quality as they can frequently be found in the literature.

Accuracy The degree of correctness and precision with which the real world data of interest to an application domain is represented in an information system

Completeness The degree to which all data relevant to an application domain has been recorded in an information system

Timeliness The degree to which the recorded data is up-to-date

Consistency The degree to which the data managed in an information system satisfies specified integrity constraints

Naturally the above characterizations are not definitions since they are hard to measure and not context independent. There are several works that deal with additional data quality aspects, in particular in the context of management information systems. For example, in [21], DQ dimensions are organized according to DQ categories (Figure 1). The above data quality aspects are also by no means complete, depending on the specific context (application domain) in which data quality is considered. For example, [28] present a survey of 179 data quality dimensions suggested by various data consumers. A more detailed discussion of some of these dimensions can be found in [16], Chapter 3.

DQ Category	DQ Dimensions
Intrinsic DQ	Accuracy, objectivity, believability, reputation
Accessibility DQ	Accessibility, access security
Contextual DQ	Relevancy, value-added, timeliness, completeness, among of data
Representational DQ	Interpretability, ease of understanding, concise/consistent representation

Figure 1: DQ categories and dimensions (taken from [21])

In order to better characterize (or even formally define) data quality aspects or dimensions, it is important to recognize that data quality cannot be studied in isolation, for example, in the context of just only one application. Underlying the management of data there are typically complex processes and workflows. It is thus imperative to study data quality aspects for the entire data management process. That is, one has to focus on three components: (1) *data producers*, (2) *data custodians* (entities that provide and manage resources for processing and storing data), and (3) *data consumers*. Depending on the application domain, these must not necessarily be different entities. For example, in the context of Web-based information systems, data producer and custodian are often the same entity. Complex information system infrastructures and in particular the Web comprise many producers, custodians, and consumers. In such a setting, the analysis and characterizations of data quality aspects naturally becomes more difficult because of the complex data (and feedback) flows underlying such systems. It will be a major objective of the seminar's working groups to precisely characterize such settings in specific application domains and to develop precise quality dimensions in these settings.

3 Fundamental Questions

Given a specific application scenario in which data producers, custodians, and consumers including their interactions have been characterized, several fundamental questions regarding DQ aspects can be posed. These questions serve the development of specific models and definitions for DQ dimensions in the scenario(s) considered. In the following, we will formulate some general questions we hope participants will study in working groups. In the following section, we then outline some application scenarios and domains that might be of particular interest for studying these questions and their answers.

(1) How is data quality assessed (DQ Assessment)?
It is natural to first ask this question data consumers and then work backwards to data providers, investigating the impact and propagation of poor quality data from data producers to consumers. The above question can also be formulated as "how is DQ measured"? Again, this question needs to be addressed from the viewpoint of the three components contributing to the scenario. For example, incomplete data can mean different things to data consumer and data producer in a given application scenario.

(2) How can data quality be model as metadata (DQ Metadata)?
Ideally, data should come with metadata describing the various processes the data went through until it reaches the data consumer. Of particular interest in this context is the notion of *data lineage (or data provenance)* that characterizes how data/information has been obtained [1, 5]. Embedding data lineage aspects as metadata into the data flow and workflow, of course, can lead to drastic improvements in dealing with DQ issues (given that the metadata is of "good quality").

(3) How to describe the DQ life-cycle?
Answers to the above questions might give valuable insights into the entire life-cycle of the quality of

data. Of particular interest in this context is the development of models that initially provide for some DQ measurements and then improve the quality of data through feedback mechanisms etc. Formalizing and modeling such DQ life-cycles for several application scenarios might help to better communicate DQ requirements and issues among data producers, custodians, and consumers.

(4) How are DQ aspects utilized at the data consumer side (DQ-based usage)?
To what extend do current application scenarios provide users with means to explicitly formulate DQ requirements and provide DQ feedback to data custodians and producers? What impact would such models have on processing and managing data?

(5) How to deal with poor quality data (DQ Improvement)?
Depending on the degree of autonomy of data producer, custodian, and consumer, what are appropriate means to improve the quality of data, either through improvements of the data management process or through explicit (observable/queryable) DQ measures and metrics.

(6) What are the relationships between data quality and trust?
In the context of Web-based information systems, the trustworthiness of data recently received quite a lot of interest. Some works simply consider trust as one DQ dimension whereas other works use the notion of trustworthiness of data as some kind of aggregation for multiple DQ dimensions. What are the exact relationships between the notions of DQ and trust? Is trust really a more abstract concept for DQ or are there specific application scenarios where the trust in data plays a more important role than DQ?

The above list of questions is by no means complete and we hope that during the first two days of the seminar, other important and challenging questions will arise. In general, we hope that for specific application scenarios and settings, working groups will be able to (formally) model interactions among data producers, custodians, and consumers and to extend these model in order to address and develop solutions to the above questions. In particular, we envision that through comparing solutions for the above problems in different application settings, we will be able to identify new aspects, principles, and general models that can be adopted to a wider range of application scenarios.

4 Questions in Context

Most of the work focusing on data quality has mainly been dealing with general application settings, primarily in the context of management information systems, Web-based information systems, or data integration. While these types of systems definitely can serve as a starting point for studying the above fundamental questions, it is our aim to have very specific application scenarios in place. That is, we would like to develop specific DQ solutions rather than general frameworks or just recycle topics already completed.
The questions formulated in the previous section will be studied in working groups, each working group focusing on a specific application domain and type of application or data management setting. The following

list suggests some of these settings.

(1) Scientific Databases

In numerous areas of the computational sciences such as biology, physics, chemistry, and astronomy, huge amounts of data are generated from experiments, observations, and simulations. Ensuring that the input to data analysis and explorations tools is of high quality poses major challenges in respective data management infrastructures.

(2) Data integration

There has been quite a lot of work on DQ issues in the context of data integration scenarios, in particular the usage of DQ in query formulation, processing (mediation) and optimization (e.g., [7, 14, 16, 17, 19]). Interesting aspects in general data integration scenarios are how the quality of data from different, perhaps heterogeneous and dynamic sources, is assessed and measured and how DQ dimensions are represented to users and applications. What are the specifics of the data flows? How is information about DQ dimensions captured and maintained as metadata? What feedback mechanisms exist or are desirable to improve DQ in different application and data integration settings? When and how should DQ aspects be visible (e.g., in query formulation) to the user?

(3) E-commerce

E-commerce can be considered as a special case of a data integration scenario. However, in E-commerce there are much more stringent requirements regarding the quality of the data provided to data consumers, primarily regulated through business rules or federal and government regulations. What are he specific DQ standards (if such exist) and how are they realized at the different components of data management infrastructures for E-commerce (including E-business, B2B etc).

(4) Streaming data

The management and processing of streaming data has become a very hot research area in the database community. Because of the characteristics of the data in, e.g., sensor networks, dealing with the quality of query results and quality of the data underlying the computation of these results is a non-trivial issue. What DQ principles, models, and techniques can be directly adopted to deal with DQ aspects in data stream management and what novel techniques and concepts need to be developed?

(5) Web data

We consider Web data as heterogeneous forms of data collected by a Web crawler. In particular, Web data must not necessarily correspond to data extracted from Web-accessible databases. In the context of “plain” Web pages collected by a Web crawler, what are the specific DQ dimensions of interest? How is the quality of, e.g., a Web page assessed, measured and represented in managing and querying Web data? Are current techniques employed by, e.g., Google, sufficient or are there more precise and better DQ measurements and techniques?

(6) Data Warehouses

Data warehouses typically contain data from multiple sources, aggregated over time. Although often data cleansing techniques are employed while data is loaded into a data warehouse, many reports from industry and government projects indicated that several "mission critical" data warehouses contain a huge amount of poor quality data. What are the specific problems regarding the assessment, measurement, and utilization of DQ dimension in data warehouses? What model and techniques should be employed in creating data warehouses that maintain high quality data and existing data warehouses that are polluted by poor quality data?

(7) Data Mining

Data exploration and knowledge discovery tools have become standard applications in the context of large-scale databases and data warehouses. The role of data mining (DM) techniques in these settings can be investigated from two perspectives: (1) standard usage of DM to discover patterns that are relevant to improve business practices, and (2) usage of DM to investigate the quality of the data managed in the database or data warehouse. What are the specific DQ assessment and measurement models in these settings? What impact does information about DQ dimensions have on managing and utilizing the data residing in such data stores?

Of course, many more application scenarios and settings can be given. The above list only illustrates some of the areas we envision to be covered during the seminar. We hope that participants can contribute other areas and develop more specific application settings and scenarios regarding the above areas.

5 Summary and Work-plan

The base questions and application scenarios illustrated in the previous two sections are supposed to serve as starting point for developing (formal) models, techniques, and approaches to DQ during the seminar in the context of working groups. It will be up to the working groups how to address these challenges, e.g., either bottom-up by starting with a very specific application scenario or top-down by initially focusing on specific fundamental questions. Eventually, some or all base questions (including those not listed in Section 3) should be addressed during this seminar.

We are in particular interested in developing complete models that cover data producers, custodians, and consumers. We hope that until the beginning of the seminar and during the first two days of the seminar, participants pose fundamental base questions and develop realistic and important application scenarios and settings in which DQ is of importance. The following references (some are not cited in the above sections) provide the interested reader with some more background material on data quality. Please note that online versions of most of the papers listed can be found at `www.db.cs.ucdavis.edu/Dagstuhl03/`.

Finally, any comments and additions on the above statements and coverage of questions and applications are welcome and should be addressed to the organizers of the seminar. We will then include respective statements into this documents and also make the statements available to other participants.

References

[1] Peter Buneman, Sanjeev Khanna, Wang Chiew Tan: Why and Where: A Characterization of Data Provenance. In Database Theory - ICDT 2001, 8th International Conference, LNCS 1973, Springer, 316-330, 2001.

[2] Donald P. Ballou, Giri Kumar Tayi: Enhancing Data Quality in Data Warehouse Environments. CACM 42(1): 73-78, 1999.

[3] Monica Bobrowski, Martina Marr, Daniel Yankelevich: A Homogeneous Framework to Measure Data Quality. In *MIT Conference on Information Quality (IQ)*, 115-124, 1999.

[4] InduShobha N. Chengalur-Smith, Donald P. Ballou, Harold L. Pazer: The Impact of Data Quality Information on Decision Making: An Exploratory Analysis. IEEE Transactions on Knowledge and Data Engineering 11(6): 853-864, 1999.

[5] Yingwei Cui, Jennifer Widom: Practical Lineage Tracing in Data Warehouses. In Proceedings of the 16th International Conference on Data Engineering, IEEE Computer Society, 367-378, 2000.

[6] Tamraparni Dasu, Theodore Johnson, S. Muthukrishnan, Vladislav Shkapenyuk: Mining database structure; or, how to build a data quality browser. In *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data*, 240-251, 2002.

[7] Michael Gertz: Managing Data Quality and Integrity in Federated Databases. In *Second Working Conference on Integrity and Internal Control in Information Systems: Bridging Business Requirements and Research Results*, 136 Kluwer, 211-230, 1998.

[8] Markus Helfert, Eitel von Maur: A Strategy for Managing Data Quality in Data Warehouse Systems. In *MIT Conference on Information Quality (IQ)*, 62-76, 2001.

[9] Theodore Johnson, Tamraparni Dasu: Data Quality and Data Cleaning: An Overview. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, 681, 2003.

[10] Beverly K. Kahn, Diane M. Strong, Richard Y. Wang: Information Quality Benchmarks: Product and Service Performance. CACM 45(4): 184-192 (2002).

[11] Dominik Lbbers, Udo Grimmer, Matthias Jarke: Systematic Development of Data Mining-Based Data Quality Tools. In *Proceedings of 29th International Conference on Very Large Data Bases*, 2003.

[12] Stuart E. Madnick, Richard Y. Wang, Frank Dravis, Xinping Chen: Improving the Quality of Corporate Household Data: Current Practices and Research Directions. In *MIT Conference on Information Quality (IQ)*, 92-104, 2001.

[13] Massimo Mecella, Monica Scannapieco, Antonino Virgillito, Roberto Baldoni, Tiziana Catarci, Carlo Batini: Managing Data Quality in Cooperative Information Systems. in *DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002*, LNCS 2519, Springer, 486-502, 2002.

[14] George A. Mihaila, Louiqa Raschid, Maria-Esther Vidal: Using Quality of Data Metadata for Source Selection and Ranking. In *Proceedings of the Third International Workshop on the Web and Databases, WebDB 2000*, 93-98, 2000.

[15] Amihai Motro, Igor Rakov: Estimating the Quality of Databases. In *Flexible Query Answering Systems, Third International Conference, FQAS'98*, LNCS 1495, Springer, 298-307, 1998.

[16] Felix Naumann: Quality-Driven Query Answering for Integrated Information Systems. LNCS 2261, Springer, 2002.

[17] Felix Naumann, Ulf Leser, Johann Christoph Freytag: Quality-driven Integration of Heterogenous Information Systems. In *Proceedings of 25th International Conference on Very Large Data Bases*, 447-458, 1999.

[18] Jack E. Olson: Data Quality: The Accuracy Dimension. Morgan Kaufmann 2003

[19] Barbara Pernici, Monica Scannapieco: Data Quality in Web Information Systems. In *ER 2002, 21st International Conference on Conceptual Modeling*, LNCS 2503, 397-413, 2002.

[20] Leo Pipino, Yang W. Lee, Richard Y. Wang: Data quality assessment. CACM 45(4): 211-218 (2002).

[21] Diane M. Strong, Yang W. Lee, Richard Y. Wang: Data Quality in Context. CACM 40(5): 103-110 (1997)

[22] Diane M. Strong, Yang W. Lee, Richard Y. Wang: 10 Potholes in the Road to Information Quality. IEEE Computer 30(8): 38-46 (1997)

[23] Bhavani M. Thuraisingham, Eric Hughes: Data quality: developments and directions. In *IFIP TC11/WG11.3 Fourth Working Conference on Integrity, Internal Control and Security in Information Systems*, Kluwer, 97-102, 2001.

[24] Sabrina Vazquez Soler, Daniel Yankelevich: Quality Mining: A Data Mining Based Method for Data Quality Evaluation. In *MIT Conference on Information Quality (IQ)*, 162-172, 2001.

[25] Yair Wand, Richard Y. Wang: Anchoring Data Quality Dimensions in Ontological Foundations. CACM 39(11): 86-95 (1996)

[26] Richard Y. Wang, Henry B. Kon, Stuart E. Madnick: Data Quality Requirements Analysis and Modeling. In, *Proceedings of the Ninth International Conference on Data Engineering*, 670-677, 1993.

[27] Richard Y. Wang: A Product Perspective on Total Data Quality Management. CACM 41(2): 58-65 (1998)

[28] Richard Y. Wang, Diane M. Strong: Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems 12:4, 5–34, 1996,

[29] Richard Y. Wang, Veda C. Storey, Christopher P. Firth: A Framework for Analysis of Data Quality Research. IEEE Transactions on Knowledge and Data Engineering 7(4): 623-640, 1995.

[30] Richard Y. Wang, Mostapha Ziad, Yang W. Lee: Data Quality. Kluwer 2001