

**Dagstuhl Seminar No. 03362**

**Report No. 0000**

## **Data Quality on the Web**

**31.08.-05.09.2003**

**organized by**

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, Technical University of Ilmenau, Germany

## **Participants**

Susanne Boll, Universität Oldenburg, Germany

Alejandro Buchmann, Technische Universität Darmstadt, Germany

Kasim Selcuk Candan, Arizona State University, U.S.A.

Cinzia Cappiello, Politecnico di Milano, Italy

Stefan Conrad, Universität Düsseldorf, Germany

Terence Critchlow, Lawrence Livermore National Lab, U.S.A.

Michael Gertz, University of California, Davis, U.S.A.

Vipul Kashyap, National Center for Biotechnology Information, U.S.A.

Frank Köster, Universität Oldenburg, Germany

Wolfgang Lehner, TU Dresden, Germany

Felix Naumann, Humboldt-Universität zu Berlin, Germany

Vincent Oria, New Jersey Institute of Technology, U.S.A.

Louiqa Raschid, Univ. of Maryland at College Park, U.S.A.

Gunter Saake, Universität Magdeburg, Germany

Kai-Uwe Sattler, TU Ilmenau, Germany

Monica Scannapieco, Università di Roma “La Sapienza”, Italy

Gerd Stumme, Universität Karlsruhe, Germany

Can Türker, ETH Zürich, Switzerland

Cai Ziegler, Universität Freiburg, Germany

M. Tamer Özsu, University of Waterloo, Canada

## Preface

Although techniques for managing, querying, and integrating data on the Web have significantly matured over the last few years, well-founded and applicable approaches to determine or even to guarantee a certain degree of quality of the data are still missing. Reasons for this include in particular the lack of common, agreed-upon models of quality measurements and the difficulty of handling quality information during data integration and query processing. The problem of data quality arises in many scenarios, e.g., during the integration of business or scientific data, in Web mining, data dissemination, and in particular in querying the Web using search and meta-search engines. Furthermore, it affects various kinds of data, such as structured and semistructured data, text documents as well as streaming data. Information about data quality is becoming more and more important since it provides some kind of yardstick describing the value and reliability of (possibly heterogeneous) forms of distributed or integrated data.

The aim of this seminar was to foster collaboration among researchers from different areas working on problems related to data quality. This included but was not limited to data integration, information retrieval (particularly search engines), scientific data warehousing and applications domains from the computational sciences and bioinformatics. In all these areas, data quality plays a crucial role and therefore different specific solutions have been developed. Sharing and exchanging this knowledge could result in significant synergy effects.

The seminar focused on the following major issues:

- Criteria and measurements for quality of Web data
- Representation and exchange of quality information as metadata
- Usage and maintenance of data quality in Web querying and data integration

These general issues have been discussed in terms of the following specific topics:

- Integrity and quality in Web data integration
- Data quality in search engines
- Monitoring of data quality, Web dynamics
- Methodologies for data quality assessment and management
- Quality-driven integration and query processing
- Quality of scientific data

The intention was to clarify terminologies and models, analyze the state of the art in the different areas, discuss problems, approaches and applications of quality-aware Web data management and to identify future trends and research directions in the above mentioned areas.

# 1 Introduction

## 1.1 Aims and Scope of the Seminar

The aim of this section is to provide the reader with basic 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 reader with an understanding of the seminar, outline specific foci, and raise questions and problems to be addressed in research related to data quality.

In the following Section 1.2, 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 1.3, we detail a list of questions that arise when data quality is of concern. In Section 1.4, 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 1.5, we summarize the objectives and outcomes that have been taken into account by the working groups, whose reports can be found in the subsequent sections of this report.

## 1.2 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 was 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 suggested 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
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 was an important 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.

### 1.3 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 formulate some general questions that were to be studied 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 during the seminar, other important and challenging questions came up. As outlined in the following Sections 2-5, for some application scenarios and settings, working groups were 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, through comparing solutions for the above problems in different application settings, we were able to identify new aspects, principles, and general models that can be adopted to a wider range of application scenarios.

#### **1.4 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 was our aim to have very specific application scenarios in place. That is, we wanted to develop specific DQ solutions rather than general frameworks or just recycle topics already completed.

The questions formulated in the previous section were studied in working groups, each working group focusing on a specific application domain and type of application or data management setting. The following list was proposed prior to the seminar.

*(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 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 the 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) 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?

### *(5) 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?

The following data quality settings were not fully covered during the seminar but addressed in the context of the above settings.

### *(6) 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?

### *(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 could have been suggested. The above list only illustrates some of the areas we envisioned to be covered during the seminar. During the seminar participants contributed other areas and developed more specific application settings and scenarios regarding the above areas.

## 1.5 Summary and Work-plan

The base questions and application scenarios illustrated in the previous sections were 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 was 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 1.3) were addressed during the seminar.

We were in particular interested in developing complete models that cover data producers, custodians, and consumers. At the beginning of the seminar and during the first two days of the seminar, participants posed 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) provided 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/Dagstuhl103/](http://www.db.cs.ucdavis.edu/Dagstuhl103/).

## 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

## 2 Working Group “Metadata and Modeling”

Prepared by Kai-Uwe Sattler

The objective of the working group “Metadata & Modeling” was to identify and classify metadata related to data quality from an architectural data and workflow point of view. This was motivated by the fact that metadata play an important role in supporting quality assessment and interpretation.

For the discussion, we assumed a basic model consisting of data sources, transformation components, and clients (users/applications). Here, transformation components represent data processing units like integration services, data mining tools etc., which either influence the quality of the processed data or whose output quality depends on the quality of the input data. Both for sources and transformation components we assume that an assessment of data quality is possible. Finally, the task of the user or the application is the interpretation of quality measures.

Based on this simple model a coarse classification of metadata can be derived. We can distinguish between quality-related metadata about

- sources, e.g., information about schemas and the individual objects as well as the intended usage or the domain of sources,
- transformation services, e.g. metadata describing operators as well as object-specific metadata for capturing trace information,
- users such as the “character” and experience or simply query/answer pairs.

For the source-specific metadata, we can further classify four levels:

- The *source level* comprises metadata like coverage and/or completeness of a source wrt. a certain domain.
- *Schema level* information describes the used data model, the granularity of the schema and the attribute domains or maintenance information such as the frequency of schema changes.
- Typical quality-related metadata at the *object level* are timeliness (esp. for temporal or time-critical data), lineage, correctness and accuracy.
- The fourth level comprises *usage-based* metadata, e.g. domain aspects or domain granularity.

Transformation specific metadata describe properties of the data transformation process which affect data quality. At the lowest level these metadata are *operator-specific*, representing information such as the signature and parameters of the transformation operator, the implementation version and requirements wrt. input data. Based on individual operators composite transformation services can be realized through a set or sequence of operators. Metadata at this *middle-ware level* describe these services and their composition.

Finally, if new objects are created by a transformation service we need *object-specific* metadata in order to assess and/or interpret data quality. These metadata describe the sources and the supporting objects of the new object as well as the service instance used for creating this object, e.g., by parameter settings, service version etc.

For a classification of user-specific metadata, we assume a scenario where a user executes a query returning a result. Here, the following kinds of quality-related metadata can be identified:

- a user profile characterizing the user’s experience,
- user requirements and preferences with regard to a specific query or class of queries,
- the user’s ratings of previous results.

Assuming we are able to assess data quality based on metadata of the various forms described above, the main question is: How to provide the user with data quality information in the context of query results?

Based on the basic model of sources and transformation services introduced above, we need a *data quality algebra* where for each transformation operator a “shadow” operator computes the output quality based on the quality values of the input data and the available quality-related metadata. That is, for each operator

$$op : p_1 \times p_2 \times \dots \times p_n \rightarrow r$$

the algebra provides an operator  $op_{dq}$  such that

$$op_{dq} : dq(p_1) \times dq(p_2) \times \dots \times dq(p_n) \rightarrow dq(r)$$

where  $dq(d)$  denotes the quality of the dataset  $d$ .

Such a data quality algebra can support two different usage scenarios:

- In a *data-driven model* the user queries the data as usual but gets additional information about the quality of the results. This is similar to a cost model approach in query processing where costs (or result set cardinalities) are computed using estimates for each plan. In our case, however, we have to use the operators of the data quality algebra for determining the output quality.
- In a *quality-driven model* the user can specify quality requirements. During data processing, the system has to decide about choosing appropriate sources and transformation operators if alternatives are available in order to meet the specified requirements. This corresponds to the query optimization problem by treating quality values as costs.

As the result of the discussion in the working group we observed that metadata from all levels can serve as input for data quality operators. These metadata are either explicitly given or have to be assessed. Furthermore, in order to get expressive data quality measurements more complete quality related metadata are needed. Finally, we discussed issues of implementing a data quality algebra and concluded that an implementation is feasible in certain application domains.

## 3 Working Group “Information Quality Assessment and Measurement”

Prepared by Felix Naumann

### 3.1 Introduction

Before any kind of information quality reasoning about data can be performed, information quality<sup>1</sup> needs to be quantified. Information quality (IQ) assessment is the process of assigning numerical values (IQ-scores) to IQ-criteria. An IQ-score reflects one aspect of information quality of a set of data items. IQ-assessment is rightly considered difficult for several reasons.

1. Many IQ-criteria are of subjective nature and can therefore not be assessed automatically, i.e., independently and without help of the user. Examples include trust, understandability, reputation, etc.
2. Information sources often do not publish useful (and possibly compromising) quality metadata. Many sources even take measures to hinder IQ-assessment, and one must assume that data sources actively find ways to improve the perceived information quality without improving the quality itself. Examples include the completeness of a search engine or the update-frequency of stock quotes.
3. If the amount of data in a source is large, assessment of the entire data set is impeded. In these cases, sampling techniques are necessary, decreasing the precision of the assessed scores [3]. Examples include life sciences data sources with billions of entries and well-known IQ deficiencies.
4. Information from autonomous sources is subject to sometimes drastic changes in content and quality.

A special form of IQ-assessment is IQ-measurement. While assessment describes the overall process of acquiring IQ related metadata, measurement is a more objective process, its results stemming from inside the data itself. While measurement makes use of well-defined metrics, the entire assessment process is much less formal, e.g., using subjective metadata obtained from users.

### 3.2 An Architecture for IQ-Assessment

We regard information quality assessment in an information integration setting. Information quality is of special interest when integrating data, because often the integrated data sources are autonomous and thus of dubious quality. Additionally, the integration process itself introduces new data errors that must be assessed (and cleansed). In this report we regard the mediator-wrapper-architecture of Wiederhold [5] for a federated information system (Figure 1). The conclusions found here are just as applicable to materialized integrated information systems, such as data warehouses.

In general, IQ-assessment can be performed at the following locations of the architecture.

1. Sources

---

<sup>1</sup>We use the terms *data quality* and *information quality* interchangeably: One person’s data is another’s information.

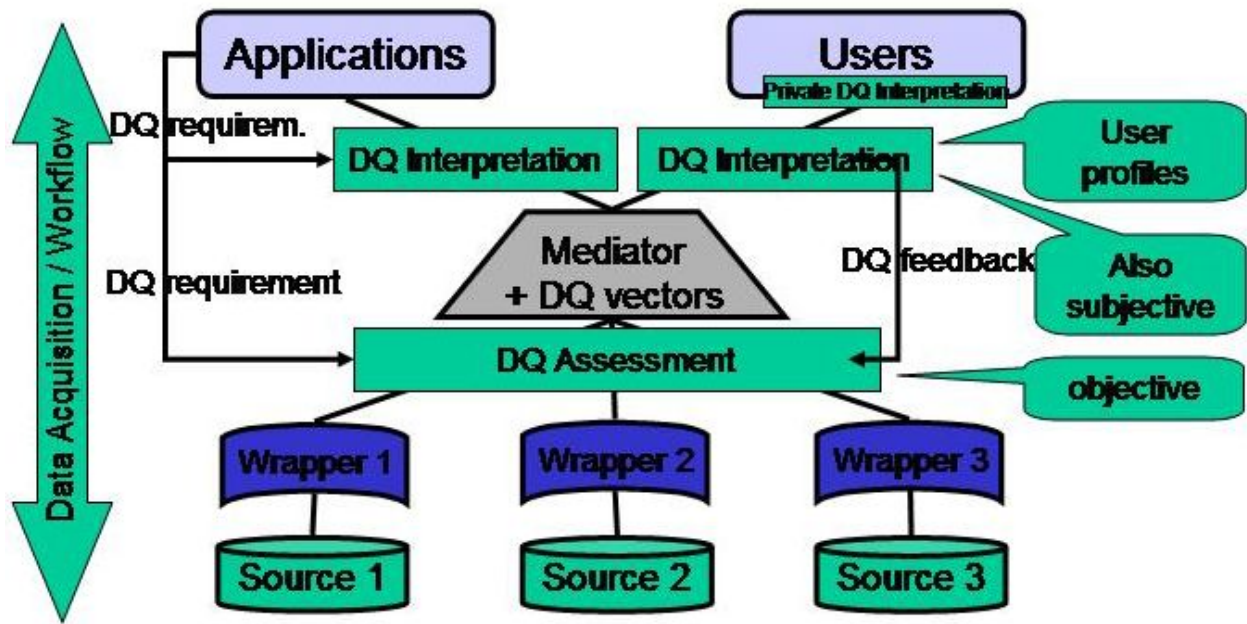


Figure 2: DQ Assessment in an integrating architecture

2. Wrappers
3. Mediated Schema
4. Mappings
5. Query decomposer
6. Result composer
7. Integrated result at user/application

In this report, we assume soundness and completeness for the processes of query decomposition and result composition. Thus, IQ assessment is not necessary at these stages.

The main finding of the working group is the distinction of IQ-assessment and IQ-interpretation. While IQ-assessment delivers the raw, untainted and unweighted metadata about information quality, IQ-interpretation uses those values to perform quality-reasoning based additional on user input, policies, weightings, etc. Thus, we postulate assessment-independence in analogy to data-independence: IQ-assessment is independent of its later use by users or applications. On the other hand, IQ-interpretation is application and user specific, as each views IQ differently and makes use of the scores in a different manner. In the following sections we describe both in more detail.

### 3.2.1 IQ-Assessment

IQ-assessment is typically based on a data source, i.e., the data stored at the source is analyzed. Lacking the ability to assess the source directly, assessment is performed on past query results obtained from the source. Assessment regards not only the data of the source, but also the quality of the metadata that is passed along with it. Poor metadata quality results in poor usage of the data and ultimately in poor quality of integrated results.

If the data is transformed and aggregated to include it in the integrated system, the processes of transformation and aggregation can heavily influence the quality of their results. Hence, the quality of these steps must be assessed as well.

IQ-assessment, as opposed to IQ-interpretation is data-oriented by nature. Results are mostly objective, because of its independence of further usage.

### 3.2.2 IQ-Interpretation

Conceptually, IQ-interpretation is performed after IQ-assessment. The IQ-scores obtained during assessment are used as input to IQ-interpretation. IQ-interpretation is typically based on individual users or applications. Users have differing requirements towards information quality. These requirements are expressed in user profiles and—during data usage—as IQ-feedback. More strict requirements towards IQ-interpretation are hard IQ standards, as expressed in ISO standards or company specifications. IQ-interpretation is performed online, i.e., during query processing and data interpretation.

As shown in Figure 1, we distinguish general and private IQ-interpretation. General interpretation is driven by general guidelines within a company, while private interpretation addresses individual users.

As a result of IQ-interpretation, several actions are possible to improve information quality: Poor quality sources can be excluded, more time and money can be invested to use additional sources, query parameters can be changed, or even new, higher quality information sources can be searched for.

### 3.2.3 Related work

IQ-assessment has been regarded in several projects. Existing assessment methods *solely* rely on users to provide IQ-scores. For instance, Wang et al. present an information quality assessment methodology called AIMQ, which is designed to help organizations assess the status of their organizational information quality and monitor their IQ improvements over time using questionnaires [4]. Bobrowski et al. present a methodology to measure data quality within an organization, again using questionnaires [1]: Following the goal-question-metric a user questionnaire is set up, which is based on samples of the database.

Both AIMQ and the approach of Bobrowski et al. rely on questionnaires to find IQ-scores. Although this assessment method is inevitable for some criteria, it is by no means the only choice for all criteria. For instance, an automated method is much more precise in assessing the average response time of a source.

Naumann and Rolker present a classification of IQ-criteria, distinguishing the ability to assess them automatically or with the help of users [2].

### 3.3 Assessment and Interpretation results

The result of IQ-assessment can have one of three flavors: numbers, categories, and explanations.

Numbers (or IQ-values) appear in vectors with one dimension per assessed IQ-criterion. To be able to compare the quality of different sources or data sets, the numbers in the vectors must be aggregated to a single IQ-score. This aggregation must perform all necessary scaling and weighting of the individual values. Based on the aggregated, overall IQ-scores, a quality-ranking can be generated.

Categories of IQ-assessment usually have only few dimensions, such as good and poor quality, or usable, partially usable, and unusable. Categories are difficult to aggregate but have the advantage of being more intuitive for data consumers, who are unaware of details of the IQ-assessment process.

Finally, explanations include more details about the information quality of a query result or a source itself. For instance, explanations show which of the sources contributed most to the poor quality, which transformation affected quality most, or even give hints on how to improve information quality. A helpful tool to generate explanations are traces of data items, as presented in the “Data Quality Dynamics” workgroup.

### 3.4 Discussion

We conclude with some open questions:

- How far should automation of IQ-assessment go? When will users begin to distrust results?
- Can all application- and user-specific aspects be considered during interpretation?
- Is a single, offline assessment of a source enough?
- If not: Is a simple parameterization of the assessment enough?
- Is the assumption of assessment independence correct? When must it be relaxed?

## References

- [1] Mónica Bobrowski, Martina Marré, and Daniel Yankelevich. A homogeneous framework to measure data quality. In *Proceedings of the International Conference on Information Quality (IQ)*, pages 115–124, Cambridge, MA, 1999.
- [2] Felix Naumann and Claudia Rolker. Assessment methods for information quality criteria. Technical Report 138, Humboldt-Universität zu Berlin, Institut für Informatik, 2000.

- [3] Frank Olken and Doron Rotem. Random sampling from database files: A survey. In *Proceedings of the International Conference on Scientific and Statistical Database Management (SSDBM)*, pages 92–111, Charlotte, NC, 1990.
- [4] Richard Y. Wang, Diane M. Strong, Beverly K. Kahn, and Yang W. Lee. An information quality assessment methodology. In *Proceedings of the International Conference on Information Quality (IQ)*, pages 258–265, Cambridge, MA, 1999.
- [5] Gio Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, 25(3):38–49, 1992.

## 4 Working Group “Do you Trust in Data Quality?”

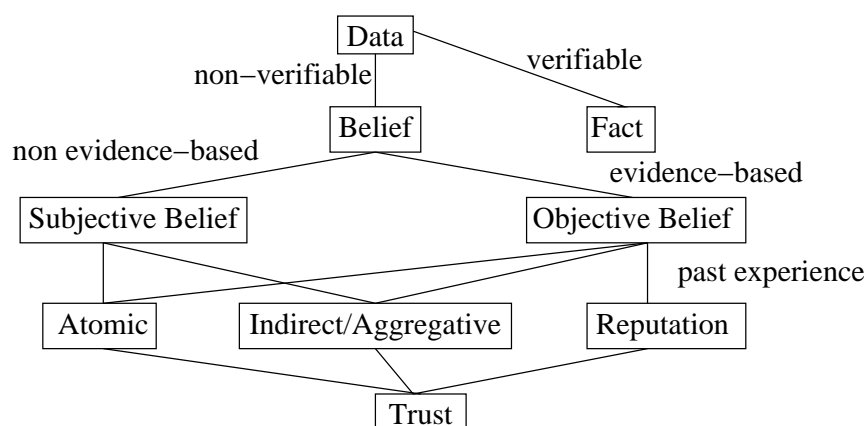
Prepared by Vipul Kashyap

### 4.1 Problem Statement and Objectives

The objective of this working group was to explore the relationship between trust and data quality, in the context of various data management settings. Given that trust has been studied by a large number of sub-disciplines of Computer Science, an attempt was made to define the notion of trust in a standardized way. Once the working definitions of trust were established, an attempt was made to identify and characterize the dimensions of data quality. A set of potential metrics to assess data quality and trust were also explored.

### 4.2 Data, Fact, Belief and Trust: Working Definitions

A taxonomy/flow chart of pragmatic and workable definitions (and not philosophical ones) for the notion of trust and other related concepts (e.g., fact, belief, reputation, etc) is proposed and illustrated below.



We start with the notion of data in an information system and the verifiability of the information captured (or interpreted) from the state of the data. It was observed that “verifiability” formed the central axis along which various notions of trust could be captured. Let us now step through the taxonomy illustrated in the figure above:

- If the information represented by the data can be verified by some means, it is denoted as a “fact”. It might be noted that the verifiability might either be based on established theories (e.g.,  $3 + 2 = 5$ ) or on the basis of “trust” in an authority (for e.g., the US Government website currently states that George W. Bush is the current US President). The notions of trust and belief involved in the latter are investigated below.
- If the information stored in the data cannot be verified by some means, then we need to have “belief” in that piece of information. Belief is a notion intimately related to trust, as according to the Webster’s

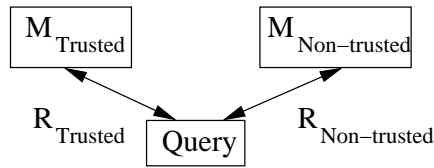
dictionary, it is defined as a *state or habit of mind in which trust or confidence is placed in some person or thing*.

- Belief may be created based on some evidence of past behavior or on some other means. This is captured in the notion of reputation, which is the *memory and summary of behavior based on past transactions*. Thus, trust based on evidence-based belief or reputation is one measurable and objective component of trust.
- Belief may also be based on some other means, i.e., non-evidence based belief. It might either be a “atomic belief”, as in the direct trust (or mis-trust) of the consumer of data on the information captured in the piece of the data.
- On the other hand, belief may either be indirect, as in depends on whether some other agent trusts the information captured in the data, or an aggregation of the trust values of other agents in the system. This is collectively referred to as “indirect belief”.

It was observed that reputation, being a summary of past transactions, built over time is an objective measure. On the other hand “non-evidence” based belief is a subjective measure, whose impact on the total trust measure is expected to fade over time. An interesting approach is to view the trust metric as having two components consisting of the objective and subjective measures. The subjective measure may only be used in situations where the values of the objective measures are unknown or are equal (as a “tie-breaker”).

### 4.3 Working Model

A working model that captures the role of trust in a distributed information system is illustrated below:



The main components of the working model are:

- There are a set of information providers or sources,  $M$  and a set of consumers. The consumers in our model are represented by a collection of queries.
- The information sources are divided into two groups,  $M_{Trusted}$  and  $M_{Non-Trusted}$ , which represent the set of trusted and non-trusted (as opposed to mis-trusted) data sources.  $M_{Trusted} \cap M_{Non-Trusted} = \emptyset$  and  $M = M_{Trusted} \cup M_{Non-Trusted}$

- Let  $Q$  be a query and  $R_{Trusted}$  be the result obtained by executing  $Q$  on the set of Information Sources in  $M_{Trusted}$  and  $R_{Non-Trusted}$  be the result obtained by executing  $Q$  on the set of Information Sources in  $M_{Non-Trusted}$ .

The Data Quality problem in this context can be framed as: *What is the relationship between  $R_{Trusted}$  and  $R_{Non-Trusted}$  and how can it be used to estimate the quality of the results?*

#### 4.4 Data Quality Dimensions and Data Management Settings

Three key dimensions of data quality, viz., completeness, correctness and timeliness are defined in terms of the relationship between  $R_{Trusted}$  and  $R_{Non-Trusted}$  defined in the working model discussed above.

- **Completeness** is the degree to which expected values are present in a data collection. There are two cases: In the case where the trusted and non-trusted data sources are known, the completeness measure can be defined as follows:

$$\frac{R_{Trusted} \cap R_{Non-trusted}}{R_{Trusted}}$$

In the case, where the trusted data sources are not known, the trusted result is based on the results available so far.

$$\frac{R_i \cap \bigcup_i R_i}{\bigcup_i R_i}$$

- **Timeliness** is the relationship between the validity of the data item and the time at which it is requested by a particular query. This is a verifiable notion and is independent of the trust-worthiness of the data source, under the assumption that validity intervals of a data item are trustworthy in this context.
- **Correctness** is the degree to which expected values are present in the results of a query and is defined in a manner dual to the completeness measure. In the presence of trusted data sources, the correctness measure can be defined as follows:

$$\frac{R_{Trusted} \cap R_{Non-trusted}}{R_{Non-trusted}}$$

In the absence of trusted data sources, the trusted result is built up incrementally based on the results available so far.

$$\frac{R_i \cap \bigcup_i R_i}{R_i}$$

Data Management settings that capture a wide variety of data quality scenarios are:

1. The traditional database setting, characterized by structured data and exact queries. The above measures are defined in this context.
2. The web data and textual documents setting, characterized by un-structured/semi-structured data and exact queries. The large number of information sources in this case make it more difficult to estimate the completeness of a given result.

3. The Information Retrieval/Multimedia setting, characterized by unstructured/semi-structured data and fuzzy queries. The results in this setting might have an associated ranking and correctness measures are impacted, e.g., in the case where the same results is returned with a different ranking order.
4. The Information Retrieval/Databases setting, characterized by structured data and fuzzy queries.

#### **4.5 Conclusions and Open Issues**

Data Quality is a composite of different dimensions which have a different evaluation based on the data management setting/scenario and the absence or presence of trusted data sources. There is a co-dependency between data quality and trust and values and metrics for both impact each other. Trust measurements, themselves have a subjective and objective component of measurement. Even though over a large period of time, based on a large history of interaction between the agents and the data sources, the subjective components of trust might "fade" away, it might be useful to view trust as a having two components, a subjective and an objective one. A list of open interesting issues and challenges in the context of trust and data quality are:

- The role of trust in the context of metadata quality
- Mechanisms for feedback between data quality and trust
- Models for combining values along different dimensions into a single data quality value
- Subjective and Objective Dimensions of Trust.

## 5 Working Group “Data Integration”

Prepared by Monica Scannapieco

### 5.1 Motivation

Data integration is the problem of combining data residing at different sources, and providing the user with a unified view of this data. Data stored by the different sources are typically overlapping data that may be incorrect, incomplete, out-of-date, in general it may be *poor quality* data.

We have investigated how the integration process is affected by poor quality of data at the sources. First, we defined some quality dimensions and discussed how quality of data at sources may be characterized in terms of such dimensions. Then, we have discussed the impact of poor quality data on the result of the integration process. Finally, we illustrated the issues using a case study involving two well known sources from the bibliographical domain.

To provide a context of the architectural setting for data integration, we distinguish the two main architectural approaches, namely:

- Materialized data integration where the (unified view of) data is materialized, for instance, in a data warehouse;
- Virtual data integration where the unified view is virtual and data resides only at sources. A reference architecture for virtual data integration is the mediator-wrapper architecture.

As a general consideration, the materialized data integration solution provides more opportunities for improving quality *off-line*. Specifically, having data materialized in the integration system supports the application of algorithms that enhance data quality by comparing across copies from different sources. Conversely, the virtual data integration solution must rely on *on-line* solutions for query improvement that are limited to instance-level reconciliation at query execution time.

In the following section, we describe quality dimensions both at local sources level and at the global data integration result level; in the latter case, we also distinguish cases in which the two described architectural solutions influence quality dimensions.

### 5.2 Quality Dimensions

A first quality dimension is the *degree of duplicates*. The degree of duplicates at sources is inherited by integrated results, if no solutions are adopted to detect and eliminate such duplicates. Furthermore, besides duplication already present at sources, the integration process also introduces additional duplicates due to data replication among local sources. The materialized data integration solution better supports the improvement (elimination) of the degree of duplicates in comparison to the virtual solution. Indeed, in the materialized solution, duplicate elimination algorithms can work on huge data sets that better allow to

accomplish some typical tasks in duplicate elimination such as sorting, clustering and duplicate decision making.

The second quality dimension we consider is the *degree of granularity*. We assume that the higher granularity at sources implicitly improves data quality since it supports a higher level of detail. This in turn increases the expressiveness of queries. The integration process can force the choice of the lowest common granularity measure. For instance, a monthly versus a daily representation may force results to be upward aggregated to a monthly representation at the integration level for all the results.

The third quality dimension is *completeness*. Completeness at sources can be considered at object level and at attribute level. Object completeness considers the number of object occurrences, while attribute completeness considers the number of attribute values describing the object occurrences. Integration typically improves the completeness of the integrated result. The degree of improvement depends on the degree of overlap at the object and attribute level. This metric is very difficult to determine evaluate as it requires knowledge about coverage with respect to a reference domain. Such information may not always be available.

The fourth quality dimension we take into account is *currency*. Currency can be referred to in different ways, including timeliness, freshness, up-to-dateness with respect to the data in sources. The maintenance of currency of the integrated result depends on the specific architectural solution adopted for integration. The materialized solution caches results and may have a negative impact on currency whereas the virtual integration solution typically supports a higher degree of currency since it must usually access the data sources.

As a fifth issue, we consider *ontologies*. The presence of ontologies makes the data integration process easier and better. We note that ontologies in this context will focus on the schema for data sources. The knowledge of schema structure is extremely important to drive the integration process, as well as a knowledge about, for example, *synonyms* and *homonyms* present in the integration setting. When considering the quality of schemas with respect to the quality of data, we can state that high quality schemas can be an important prerequisite for high quality data, in many cases. For instance, let us consider the structure of an *address*; it is more likely that a single-field *address* has a poor quality, with respect to the case in which an *address* is represented with multiple fields, such as *street*, *number*, *zipCode* and so forth.

Two further issues that may impact data quality assurance in a data integration architecture are *response time* and *availability* of sources. These two measure relate to the quality of service or access to the data. Indeed, it is very probable that there exists a trade-off between improving quality of data and the quality of service dimensions.

### 5.3 Quality Issues in Integrating Two Bibliographic Sources

We have considered two sources from a bibliographical application domain, namely CiteSeer (<http://citeseer.nj.nec.com/cs>) and DBLP

	<b>CiteSeer</b>	<b>DBLP</b>	<b>Comment</b>
Degree of Duplicates	Lower	Higher	tentatively checked
Granularity	Lower	Higher	DBLP includes subtype
Completeness	Higher	Lower	Difficult to evaluate because of different reference domains for CiteSeer (including all on-line CS papers) and DBLP (including all DB- and LP- and algorithmic papers)
Ontologies	Equal	Equal	both have schemas
Currency	possibly higher	possibly lower	DBLP performs manual updates

Table 1: Quality dimensions evaluation for CiteSeer and DBLP.

(<http://www.informatik.uni-trier.de/~ley/db/index.html>).

For each of the two sources we estimated a value or characteristic of the quality dimensions described in the previous section. This exercise was to illustrate issues in performing a quality evaluation for a real source. The results of the quality evaluation are described in Table 1.

## 5.4 Conclusions

When considering data quality in a data integration setting, both data quality at sources and data quality of the integrated result need to be considered. We have identified a set of data quality dimensions, and how they are impacted through integration. Some dimensions, such as completeness, typically improve. Other dimensions, such as duplicates or currency, *may* improve, and this depends on whether it is possible to exploit data replication or overlap among sources. A high granularity at sources typically implies higher quality, but this improvement may no longer be available in the integrated result, if the result is expressed based on the lowest common granularity measure. Finally, the use of ontologies supports the data integration process and high quality ontologies *may* have a positive impact on the quality of integrated data.