Report on the Dagstuhl Seminar "Data Quality on the Web"

Michael Gertz
U of California at Davis, U.S.A.
gertz@cs.ucdavis.edu

M. Tamer Özsu
U of Waterloo, Canada
tozsu@uwaterloo.ca

Gunter Saake
U of Magdeburg, Germany
saake@iti.cs.uni-magdeburg.de

Kai-Uwe Sattler
TU Ilmenau, Germany
k.sattler@computer.org

1 Introduction

Over the past few years, techniques for managing, querying, and integrating data on the Web have significantly matured. Well-founded and practical approaches to assess or even guarantee a required degree of quality of the data in these frameworks, however, are still missing. This can be contributed to the lack of well-defined data quality metrics and assessment techniques, and the difficulty of handling information about data quality during data integration and query processing. Data quality problems arise in many settings, such as the integration of business data, in Web mining, data dissemination, and in querying the Web using search engines. Data quality (DQ) addresses various forms of data, including structured and semistructured data, text documents, multimedia, and streaming data. Different forms of metadata describing the quality of data is becoming increasingly important since they provide applications and users with information about the value and reliability of (integrated) data on the Web.

The Dagstuhl Seminar "Data Quality on the Web", organized by Michael Gertz, Tamer Özsu, Gunter Saake, and Kai-Uwe Sattler, took place between August 31st and September 5th 2003 at Schloss Dagstuhl, Germany. The objective of the seminar was to (1) foster collaboration among researchers that deal with DQ in different areas, (2) assess existing results in managing the quality of data, and (3) establish a framework for future research in the area of DQ. The application contexts considered during the seminar included in particular (Web-based) data integration and information retrieval scenarios, scientific databases, and application domains in the computational sciences and Bioinformatics. In all these areas, data quality plays a crucial role and therefore different, tailored solutions have been developed. Sharing and exchanging this knowledge could result in significant synergy effects.

2 The Meaning of Data Quality

Compared to core database concepts such a database integrity and security, the notion of *data quality (DQ)* has only emerged during the past 10 years and shows a steadily increasing interest. The 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 technologies. More than ever before, businesses, governments, and research organizations rely on the exchange and sharing of data. It is widely recognized that dealing with data quality problems is 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 [1], information quality (or data quality) is "an inexact science in terms of assessments and benchmarks". Oftentimes, high-quality data is simply "data that is fit for use by data consumers" [3]. An important objective of this seminar was to analyze and develop precise data quality definitions for data management scenarios and applications that are of practical relevance. A few conventional characterizations of data quality, as they can frequently be found in the literature were suggested, were used to develop such definitions.

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

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

Timeliness. The degree to which the recorded data are up-to-date.

Consistency. The degree to which the data managed in an information system satisfy specified constraints and business rules.

There are several works that deal with additional data quality aspects, in particular in the context of management information systems. For example, in [3] DQ dimensions are organized according to DQ categories

(such as intrinsic, contextual, accessibility, and representational). The most extensive summary of data quality aspects suggested by various data consumers can be found in [4], where 179 data quality dimensions are surveyed.

In order to (precisely) define data quality aspects, it is important to recognize that data quality cannot be studied in isolation, for example, in the context of 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* (entities that generate data), (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 flows underlying such systems. It was an important objective of the seminar to characterize such settings in specific application domains and to develop precise quality dimensions in these settings.

3 Seminar Approach

During the first day of the seminar, participants gave short presentations of their research and interest in data quality. Most participants also illustrated specific application domains they are working with. Based on problem statements suggested by the participants, core topics on data quality were identified as follows:

- *Criteria and assessment techniques* for the quality of data residing on the Web (including database systems),
- *representation, exchange, and maintenance* of data quality information as metadata, and
- *usage and maintenance* of data quality measures in data integration scenarios and querying the Web.

Working groups were built, each of which studied one of the topics in detail for a day. The findings of each group were presented in plenary sessions, which were followed by detailed discussions on the proposed solutions, open problems, and research agendas. The groups were then shuffled and different working groups studied the issues, starting from where the first ones left off, and the process was repeated. In some cases, additional working groups were formed based on the findings of the first set of working groups and plenary discussions.

A sub-task for each working group was to clarify terminologies and models, analyze the state of the art in the areas related to the specific topic, discuss problems, approaches and applications of quality-aware (Web) data management infrastructures, and to identify future trends and research directions regarding the topic.

The questions and application scenarios illustrated in the previous section were supposed to serve as starting points for developing (formal) models, techniques, and approaches to DQ in these 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 fundamental questions. A particular goal was to develop complete models that cover data producers, custodians, and consumers.

4 Results from Working Groups

4.1 Metadata and Modeling

The objective of the working group "Metadata and Modeling" was to identify and classify metadata related to data quality from a data processing and workflow point of view. This was motivated by the fact that metadata plays an important role in supporting quality assessment and interpretation.

For the discussion, a basic model consisting of data sources, transformation components, and clients (users/applications) was assumed. Transformation components represent data processing units such as schema mapping and data integration services, which either influence the quality of the processed data or where the quality of result data depends on the quality of the input data. For both sources and transformation components, it was assumed that an assessment of data quality at a component is possible. The interpretation of a resulting quality measures is then done by the user or application that consumes the processed data.

Based on the above model, a coarse classification of metadata has been developed. This classification includes metadata about the following components.

- Sources, that is, information about the coverage and completeness of the source; source schemas (data model and granularity), individual objects (timeliness, lineage, accuracy etc.), and the intended usage or the domain of sources.
- Transformation services, in particular metadata that describes the signatures, parameters, and requirements of operators, compositions of transformation services, and metadata capturing trace information about objects processed in the model.

- Users, e.g., data expectation or requirements of a user, or simply uninterpreted query/answer pairs.

Assuming that it is possible to assess data quality based on metadata of the forms described above, the main question then is: How to provide the user with data quality information in addition to query results? Based on the model of sources and transformation services introduced above, a *data quality algebra* was suggested and outlined. In this algebra, for each transformation operator, a "shadow" operator computes the quality of the output based on the quality values of the input data and available quality-related metadata. Such a data quality algebra can be used in two ways:

- In a *data-driven model*, the user queries the source(s) as usual, but additional information about the quality of the query result is provided. This is similar to a cost model approach for query processing where the cost (or result set cardinalities) for each query plan is computed using estimates. In a data quality-centric setting, operators of the data quality algebra compute the quality of the result while a query is processed (or planned).

- In a *quality-driven model*, the user can specify data quality requirements. During query processing, the system decides what sources and transformation operators to use, if alternatives are available in order to meet the specified requirements. This approach resembles the query optimization problem by treating quality values as costs.

It is important to note that metadata from all levels can serve as input for data quality operators in such a special purpose algebra. The metadata is either explicitly given or has to be assessed. Furthermore, in order to obtain meaningful data quality measures, (almost) complete quality related metadata is needed. Finally, in this working group, an implementation of a data quality algebra was outlined. While a general and comprehensive algebra is difficult to develop, it was observed that an implementation is feasible in some application domains.

In several discussions related to DQ metadata and modeling, it became evident that trust plays in important role in data quality. Thus a working group "Trust and Data Quality" was formed with the objective to explore the relationship between trust and data quality in the context of different data management architectures.

Based on the notion of data in an information system and the verifiability of the information represented by the data, it was observed that the "verifiability" of data forms the central axis along which various notions of trust in data can be defined. In particular, one then can distinguish between "fact" and "belief". If the information represented by data can be verified by some means, it is considered a fact. The verification can either be based on established theories or on the basis of "trust" in an authority assessing the information. If the information represented by the data cannot be verified, then there is certain "belief" in that piece of information. Belief may be created based on some evidence of past behavior or by some other means. This is captured by the notion of "reputation", which is the memory and summary of behavior based on past transactions. Thus, trust that relies on evidence-based belief, or reputation is one measurable and objective component of trust.

As a working model that captures the role of trust and its relationship to data quality in information systems, a set of data providers (sources) and a set of data consumers was considered. Consumers are simply represented by a collection of queries. Data providers are divided into two groups representing the set of trusted and non-trusted (as opposed to mis-trusted) providers, respectively. The data quality problem thus can be formulated as follows: What is the relationship between the query result obtained from a trusted source compared to the result obtained from a non-trusted source, and how can a relationship be used to measure or estimate the quality of query results?

It turns out that data quality is a composite of different dimensions, which have different evaluations based on the data management framework and the absence or presence of trusted data sources. There is a co-dependency between data quality and trust, and values and metrics of both have an impact on each other. Trust measurements themselves have a subjective and objective components. Even though, over a long period of time, based on the history of interactions between users and data sources, the subjective component of trust might fade away, it might be useful to view trust as a having two components, a subjective and an objective one.

4.2 Data Quality Assessment

In the working group "Data Quality Assessment and Measurement", issues regarding the acquisition and interpretation of metadata related to data quality were studied. Assessment is defined as the process of assigning numerical or categorical values (quality scores) to quality criteria in a given data setting. Measurement can be considered as a special form of assessment where well-defined DQ metrics are used and therefore resulting DQ metadata are more objective. In general, however, quality assessment is a difficult task due to the following reasons. First, many DQ criteria are subjective and therefore cannot be assessed automatically (e.g., trust or understandability). Second, many sources do not publish quality related metadata. Finally, for sources with large amounts of data, DQ assessment can be performed only on a sample of the data, thus result-

ing in a decreased precision of quality scores.

One can further distinguish between DQ assessment and DQ interpretation. Assessment produces "raw" and unweighted quality metadata, independent from a later usage. Interpretation, on the other hand, uses the assessment output in order to perform a reasoning on DQ using additional information such as user input and DQ requirements. Consequently, interpretation is user/application specific. Users and applications can view and interpret quality scores in different ways.

Considering an information integration scenario based on a mediator-wrapper-architecture as one of the main applications of quality assessment, one can identify several levels where quality assessment and interpretation occurs. Assessment is typically based on the data sources, e.g., by analyzing the data stored in a source or accessible through wrappers. However, not only the data have to be assessed but also the quality of the associated metadata. Clearly metadata of poor quality result in poor quality of the integration results. Furthermore, the quality of the integrated data depends not only on the quality of the base data. If integration is performed by applying data aggregation or transformation operations, the semantics of these operations have to be taken into account, too.

Quality interpretation is performed after assessment, that is, "above" the mediated schema of an integration architecture. Here, users can express their quality requirements in user profiles or as quality feedback during data usage. Given these requirements, the quality scores obtained by DQ assessment steps can be used as input for a quality interpretation. As a result of quality interpretation, a broad range of actions are possible to improve the quality of (integrated) data, ranging from changing the set of sources (excluding poor quality sources and looking for higher quality sources) to changing query parameters or even queries.

Results of assessment and interpretation can be represented in different forms. *Numbers* are used in quality vectors, representing dimension values for a certain quality criteria (e.g., a completeness of 80%). Such numbers can be aggregated to single scores, allowing a comparison of the quality of different sources or even a quality ranking. In contrast, assessment *categories* provide only a few values (e.g., accurate versus not accurate). They are easy to use and to interpret by the user, but are difficult to aggregate in a single score. Finally, *explanations* provide more details on the data or query results, e.g., about the lineage and trace of a data object resulting from transformation operators.

4.3 Data Integration

Data integration is the problem of combining data residing at different sources, and providing users with a unified view of the data. Data managed by different sources are typically overlapping, and data can be incorrect, incomplete, out-of-date, that is, it may be data of *poor quality*. The objective of this working group was to investigate how the integration process is affected by poor quality of data at the sources. To study this problem, two architectural approaches were considered: *materialized data integration* and *virtual data integration*.

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 this context, the following quality dimensions are important at both local sources level and global data integration result level.

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 that is already present at sources, the integration process can introduce additional duplicates due to data replication among local sources. The materialized data integration solution better supports the elimination of duplicates in comparison to the virtual approach by applying different duplicate elimination techniques.

Degree of granularity. A high degree of object granularity at sources implicitly improves the quality of data since properties of data can be described in more detail. This in turn increases the expressiveness of queries. The integration process can force the choice of the lowest common granularity. For instance, a monthly versus a daily data representation may force results to be upward aggregated to a monthly representation at the integration level for all query results.

Completeness. Completeness at sources can be considered at the object level and at the 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. However, this metric is very difficult to evaluate as it requires knowledge about source coverage with respect to a reference domain.

Currency. Currency can be viewed in different ways, e.g., as timeliness or freshness of the data managed by the sources. The maintenance of currency of an integrated result depends on the specific architectural solution adopted for data integration. The materialized solution caches results and may have a negative impact on currency whereas the virtual integration solution typi-

cally supports a higher degree of currency.

Ontology. The presence of an ontology is likely to improve the data integration process. In this context, an ontology focuses on the description of source schemas. Knowledge about schema structures is extremely important in order to drive the integration process. Considering the quality of schemas with respect to the quality of data, one can claim that in many cases, high quality schemas are an important prerequisite for high quality data.

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

4.4 Data Quality Dynamics

The objective of the working group "Data Quality Dynamics" (DQ dynamics, for short) was to identify the impact of dynamics on data quality. The term 'dynamics' was chosen as a neutral term that generalizes various specific aspects in handling data quality and which are connected to the flow of time and evolution of data and software components. DQ dynamics therefore touches areas like history of data, data lineage, application evolution, and change detection.

The "dynamics of the world" influences the assessment of data quality aspects in several ways. First, the history of data and data sources influences the quality of results. For example, one may rate some quality aspects, such as timeliness or trust, depending on the history of data updates of a source. Additionally, there is an evolution of DQ ratings depending on the application domain and technical realization of the underlying data management framework.

One important aspect of DQ dynamics is the *detection of change*. Changes may occur in data sources or in components used in data integration and aggregation processes. All these changes can result in feedback to DQ assessment and processing and have to be detected automatically, if possible. An additional aspect, which is hard to capture using automated tools, is the dynamics of data usage and interpretation in a given application scenario.

A general definition of data quality dynamics can be given as *the influence of changes to data quality*. Such changes include the change of data and associated data sources, change of the data integration and aggregation framework, change of quality characteristics, and change of data usage and interpretation. The DQ characteristics are defined by the parameters (metadata schema), the actual parameter values (metadata), and the chosen DQ metrics.

Besides these fundamental notions, one has to be aware of the two dimensions of data quality dynamics. The *passive view* can be characterized as tracing and observing relevant data and software components over time. This passive view is closely related to change detection. The second dimension results from an *active view* on DQ dynamics and is concerned with influencing the resulting DQ by enforcing changes in data and components. Thus, an active view aims to enforce some kind of required *DQ life cycle*.

During the study of DQ dynamics, a general model of the processes related to DQ dynamics was developed. In this model, a static integration framework is assumed, where data objects O_i from data sources s_j are integrated, cleaned, transformed and aggregated into a result N that is rated by data quality values. As transforming operations, data transformers and data mergers are used. In this setting, one can identify two types of traces that influence the quality of a query or data integration result.

- The *functional trace* records the modifications and transformations of the data at a certain run of the integration process. Although such a run needs some time, one can consider it as a static trace. In DQ frameworks, these kinds of traces are often called *lineage* of the resulting data.

- The *temporal trace* records changes over time over static structures, for example, changes on data objects or data transformers. This is called the *life cycle* of the system.

The temporal trace is of primary interest in the context of DQ dynamics. Sources of changes to the whole system can be manifold. The most interesting cases include: operators that directly manipulate the result N and transform it into a new state, changes on the input data O_i may occur and are detected by comparing old versions of O_i with the current version O_i', or an operation on the source data induces a change on O_i.

Tracing changes is essential for DQ dynamics. Therefore, a central question is how to model and represent trace information. This question was discussed in some detail and resulted in a trace modeling framework. Besides this modeling, the representation of metadata for traces is essential. Reporting and detecting changes is, of course, a prerequisite for building traces. Because of the heterogeneity of changes in a general framework, several models and techniques need to be combined to be applicable for practical application scenarios. For such scenarios, the *integration of traces* causes an additional problem where fundamental research is necessary.

5 Conclusions

There is no question that in today's information-centric society the quality of data plays an important role. Due to the rapid development of new Web technologies that allow sharing and exchanging data in an easy and transparent fashion, however, it is likely that data quality problems will become worse before they get better. It is interesting to relate the aspect of data quality to other core database concepts such as security and semantic integrity, for which models, techniques, and architectures have been developed and are successfully used in practice.

As a research agenda, the seminar's participants suggested to first concentrate on general and expressive models that can be used for modeling DQ metrics, DQ metadata, and DQ assessment and interpretation techniques. Existing models for modeling semantic rich data schemas, workflows, systems, and information flows might provide important concepts in developing models that deal with data quality. Once such models are in place, tools need to be developed that help users and IT managers to capture and analyze the state of data quality in an information system infrastructure within these models. Although it was recognized that several approaches to dealing with DQ aspects have been proposed in the research literature, most notably Nauman's work on DQ in data integration [2], more comprehensive frameworks are necessary. In particular, existing approaches that deal with data quality in management information systems and scientific databases, which typically employ data integration concepts, might provide further insights into DQ requirements and current best-practice rules for various data quality aspects.

Acknowledgments. We would like to thank the staff of the Dagstuhl Office for their professional support for the seminar and for providing an excellent work and discussion environment. Furthermore, we thank all participants for a week of fruitful and inspiring discussions. We are also thank Vipul Kashyap, Felix Naumann, and Monica Scannapieco for contributing to this report.

References

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

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

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

[4] 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,