Ontology-Based Data Quality Management (OBDQM)

Methodology, Costs, and Benefits

Christian Fuerber and Martin Hepp
Universität der Bundeswehr München, E-Business & Web Science Research Group
c.fuerber@unibw.de, mhepp@computer.org

E-BUSINESS+ WEB SCIENCE RESEARCH GROUP

Abstract

The competitiveness of today's businesses strongly depends on their data, and the degree of automation in business processes. The business performance of information systems is constrained by the quality of this data. Due to the multidimensional characteristics of data quality, identification and improvement of data quality problems are complex tasks which require knowledge about what are correct data in the relevant domain. Arisen from Semantic Web research, ontologies have been discussed as a means to provide such knowledge, and they may actually help mitigate the problem not only from technical perspective, but also from an organizational point of view. The construction and maintenance of ontologies, however, is a costly task. Thus, they can gain practical relevance for data quality improvement only if we manage to provide certainty about its efficient usage. This poster outlines a PhD research project, which aims at developing an architecture for ontology-based data quality management (OBDQM) and a matching ex ante efficiency estimation model.

What Data Quality Problems Are We Facing?

Single-Source Problems

Data Quality Problem	Instance	Schema	Cause type
Word transposition / Syntax violation	X		Inconsistency
Outdated values	X		Inconsistency
False values	X		Inconsistency
Misfielded values	X		Inconsistency
Meaningless values	X		Inconsistency
Missing values	X		Inconsistency
Out of range values	X		Inconsistency
Invalid substrings	X		Inconsistency
Mistyping / Misspelling errors	X		Inconsistency
Imprecise values	X		Ambiguity
Unique value violation	X		Inconsistency
Violation of a functional dependancy	X		Inconsistency
Referential integrity violation	X	X	Inconsistency
Incorrect reference	X	X	Inconsistency
Contradictory relationships	X	X	Inconsistency

Multi-Source Problems

Data Quality Problem	Instance	Schema	Cause type
Heterogeneity of syntaxes	X	X	Heterogeneity
Heterogeneity of units of measurement	X	X	Heterogeneity
Data precision conflicts	X		Heterogeneity
Heterogeneity in time reference	X		Heterogeneity
Default value conflicts	X		Heterogeneity
Source specific identifiers		X	Heterogeneity
Heterogeneity of integrity constraints		X	Heterogeneity
Heterogeneity in cardinality		X	Heterogeneity
Schema discrepancy		X	Heterogeneity
Schema isomorphism conflict		X	Heterogeneity
Existence of hypernyms		X	Heterogeneity
Overlapping concepts / Role conflicts		X	Heterogeneity

Problems of Both Scenarios

Data Quality Problem	Instance	Schema	Cause type
Existence of synonyms	X	X	Heterogeneity, Redundancy
Existence of homonyms	X	X	Heterogeneity, Ambiguity
Approximate duplicate tuples	X		Redundancy
Inconsistent duplicate tuples	X		Redundancy, Inconsistency
Business domain constraint violation	X	X	Inconsistency
Outdated conceptual elements		X	Inconsistency

The data quality problems shown above are based on findings by [1],[3],[4],[5], and [6]

Inconsistency:
Data values or schema elements usually refer to other data values, schema elements, or real-world concepts. Inconsistencies occur if data or schema elements differ significantly in meaning from their reference and, therefore, either the reference or the data or schema element on hand is perceived as incorrect.

Heterogeneity:
Heterogeneity predominantly occurs in multiple-source scenarios and can be specified into structural and semantic heterogeneity. In cases of structural heterogeneity, the same real-world domain is represented by different schema elements. Semantic heterogeneity also constitutes a difference in the intension of the compared schemata with overlapping elements [1].

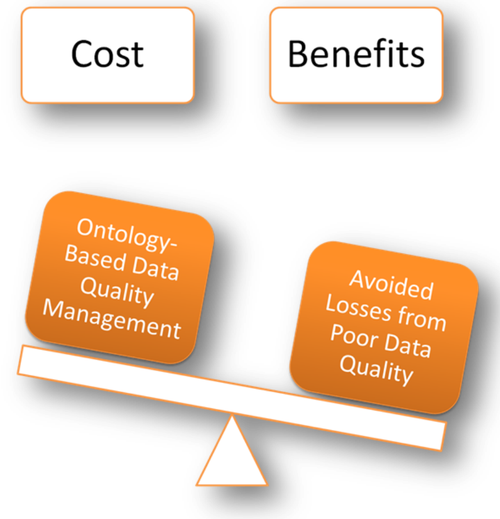
Redundancy:
Redundancy problems exist when the same real-world entity is represented at least twice. Redundancy problems are not constraint to multiple-source scenarios. A redundant tuple can also come along with inconsistency problems, if some of the attribute values differ significantly in meaning.

Ambiguity:
Ambiguous naming of data values or conceptual elements can also cause data quality problems. Due to the ambiguous naming, the intended meaning cannot be identified precisely. Hence, ambiguity can complicate identification of the real-world entities of data values or conceptual elements.

Related Work

Previous work on ontology-based data quality management focuses on data integration, data retrieval, and data cleansing. Ontology-based data integration and retrieval techniques mainly aim to reduce heterogeneity of multiple data sources, e.g. syntactic and semantic differences in data, without changing or correcting any data in the sources. Those approaches enable data retrieval and integration processes to access domain knowledge represented within an ontology with the intention to establish a common understanding about the retrieved data [7-12]. Ontology-based data cleansing approaches aspire to detect and remove data deficiencies, such as inconsistencies, data duplicates, or violation of domain constraints, directly in the source supported by domain and task ontologies [13-15]. In a nutshell, there exists preliminary research of applying ontologies to data quality issues. However, they are limited in scope and do not go beyond the stage of early prototypes.

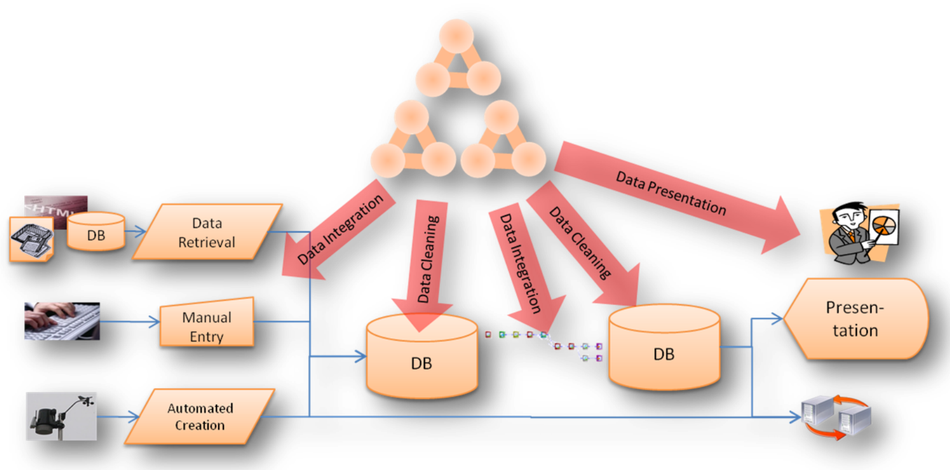
How can we estimate the costs for ontology construction, maintenance, population, and usage? By analysis of the manual processes related to ontology development and use, we aim to identify reference parameters that allow the application of a cost prediction model [17] to predict the costs of creating and using respective ontologies ex ante and from incomplete information. E.g. the number of concepts within the domain of interest might allow a quick estimation of the potential costs for respective ontologies. The identified technique will be validated within at least one case study.



How can we estimate the reduction in costs caused by poor data quality? Data quality problems usually affect one or more business processes in their performance. Ontologies are most likely not able to solve all data quality problems. Hence, for assessing the benefits of OBDQM, we first need to know what data quality problems can be solved by ontologies and, secondly, we must be able to quantify the avoided losses from data quality problems which can be solved with our approach.

Outlook on OBDQM

In OBDQM, we try to combine several existing techniques and create a single methodology and architecture to support a broad range of methods for improving data quality. A core part of OBDQM will be a task ontology (similar to [15]) which will contain typical data quality problems (see tables on the left) including appropriate algorithms for their identification, measurement, and improvement. The task ontology will be connected to other domain ontologies for further support on the resolution of domain-specific data quality problems, such as business domain constraint violations. The final execution of the proposed data cleansing tasks will be performed by the user who receives a cleansing suggestion in his user interface based on the analysis of the RDB by the predefined ontologies. Before the analysis can be conducted, the concepts of the domain ontology need to be mapped to the schema elements of the RDB. This could be performed semi-automatically by applying schema matching techniques [18]. If multiple sources are mapped to the concepts of the domain ontology, then data retrieval and integration activities will possibly utilize the concepts of the ontology to identify corresponding schema elements supporting completeness of data and, thereby, mitigating redundancies.



How Can Ontologies Help?

As partly formalized, consensual conceptualizations of a domain of interest [2], ontologies provide helpful means for handling and organizing data. By providing machine-readable knowledge about real-world circumstances and, in particular, about characteristics of common data quality problems inconsistency, ambiguity, heterogeneity, and redundancy problems may be mitigated. In [2] possible technical effects of ontologies have been described, that may be used to explain how ontologies can mitigate data quality issues:

Excluding unwanted interpretations. Within ontologies formal and informal means can be used to clearly define the intension of a concept. In other words, ontologies can help to specify a consensual, unambiguous meaning of concepts. Mapped to relational databases (RDB) ontologies can, therefore, reduce ambiguity of RDB schema elements.

Spotting logical inconsistencies. Due to the clear definition of concepts and their relationships within ontologies, consistency of schema elements and their instances can possibly be enhanced and logical inconsistencies can be identified when applying mapped ontologies on RDB.

Identification of stable and reusable conceptual elements. Ontology modeling requires the concise definition of the concepts of a domain and their relationships establishing a consensual understanding of the domain of interest. This task is usually also performed when creating a local database schema. Since ontologies are not tied to a specific database, they provide a means to establish an independent conceptual model overcoming heterogeneity of multi-source scenarios, which suffer from locally defined schemata.

The Road Towards OBDQM

The OBDQM project focuses on providing answers to the following questions:

What data quality problems can be solved by ontologies? Based on existing work, we identified ontology-based techniques that can reduce data quality problems. Additionally, we have identified typical data quality problems of RDB scenarios. Upcoming work focuses on practical experiments with ontologies on real-world data of RDB. The experiments will include scenarios for ontology-based data integration, data retrieval, and data cleansing. Eventually, we will gain a detailed understanding about the range of application of ontologies to data quality problems.

How can we integrate ontologies to manage data quality of RDB? Each step in the data lifecycle requires a different solution approach. Hence, an architecture is currently being designed to enable ontology-usage for data retrieval, data integration, and data cleansing. Despite of the decentralized application the required ontologies should be centrally accessible to reduce maintenance efforts and avoid inconsistencies and redundancies.

What kinds of ontologies should be used? From the literature we know different types of ontologies, e.g. top-level ontologies, domain ontologies, task ontologies, and application ontologies [16]. One key challenge of applying ontologies to data quality management problems is the choice of suitable types of ontologies. Also the reuse of existing ontologies must be investigated, for example ontologies derived from WordNet for the identification of synonyms on the data instance and schema level, and DOLCE and PROTON.

What has to be represented within ontologies to identify and solve data quality problems in RDB? In some cases the representation of classes and properties of real-world concepts might already provide adequate means to solve data quality problems. Other cases might additionally require the representation of individuals and axioms. There exists a substantial body of work describing methodologies for the design of ontologies at different levels of expressiveness [19]. A key challenge in this part will be the identification of a suitable methodology for the construction of data quality management ontologies.

References

1. Leser, U., Naumann, F.: Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen, dpunkt-Verlag, Heidelberg (2007)
2. Hepp, M.: Ontologies: State of the Art, Business Potential, and Grand Challenges. In: Hepp, M., De Leenheer, P., de Moor, A., Sure, Y. (eds.) Ontology Management: Semantic Web, Semantic Web Services, and Business Applications, pp. 3--22. Springer, New York (2008)
3. Oliveira, P., Rodrigues, F., Henriques, P. R.: A Formal Definition of Data Quality Problems. In: International Conference on Information Quality (2005)
4. Kashyap, V., Sheth, A.P.: Semantic and Schematic Similarities Between Database Objects: A Context-Based Approach, Very Large Data Base Journal (5), 276--304 (1996)
5. Oliveira, P., Rodrigues, F., Henriques, P.R., and Galhardas, H.: A Taxonomy of Data Quality Problems, In: Proc. 2nd Int. Workshop on Data and Information Quality (in conjunction with CAiSE'05), Porto, Portugal (2005)
6. Rahm, E., Do, H.-H.: Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bulletin 23(4), 3--13 (2000)
7. Brüggemann, S.: Ontologiebasierte domänenspezifische Datenbereinigung in Data Warehouse Systemen. In: Grundlagen von Datenbanken. (2006)
8. Kedad, Z., Métais, E.: Ontology-Based Data Cleaning. In: Proceedings of the 6th International Conference on Applications of Natural Language to Information Systems-Revised Papers (2002)
9. Madnick, S., Zhu, H.: Improving data quality through effective use of data semantics. Data & Knowledge Engineering, 59(2), 460--475 (2006)
10. Niemi, T., Toivonen, S., Niinimaki, M., Nummenmaa, J.: Ontologies with Semantic Web/Grid in Data Integration for OLAP. International Journal on Semantic Web and Information Systems 3(4), 25--49 (2007)
11. Perez-Rey, D., Anguita, A., Crespo, J.: OntoDataClean: Ontology-Based Integration and Preprocessing of Distributed Data. In: Biological and Medical Data Analysis 4345/2006, pp. 262-272. Springer, Berlin / Heidelberg (2006)
12. Skoutas, D., Simitsis, A.: Ontology-Based Conceptual Design of ETL Processes for Both Structured and Semi-Structured Data. International Journal on Semantic Web & Information Systems 3(4), 1--24 (2007)
13. Brüggemann, S., Gruening, F.: Using Domain Knowledge Provided by Ontologies for Improving Data Quality Management. In: I-Know 2008 and I-Media 2008 International Conferences on Knowledge Management and New Media Technology. (2008)
14. Brüggemann, S.: Rule Mining for Automatic ontology-based Data Cleaning. In: 10th Asia-Pacific Web Conference, (2008)
15. Wang, X., Hamilton, H. J., Bither, Y.: An ontology-based approach to data cleaning. Regina: Dept. of Computer Science, University of Regina (2005)
16. Grimm, S., Hitzler, P., Abecker, A.: Knowledge representation and ontologies. In: Studer, R., Grimm, S., Abecker, A. (eds.) Semantic web services - concepts, technologies, and applications, pp. 51--105. Springer, Heidelberg (2007)
17. Koehler, A., Leopold, N.: 'Kurzkalkulationsverfahren - Übersicht und Einsatzmöglichkeiten', Industrie-Anzeiger 108, (82), 34--35 (1986)
18. Rahm, E., Bernstein, P.A.: 'A survey of approaches to automatic schema matching', VLDB J 10, (4), 334--350 (2001)
19. Gómez-Pérez, A., Fernández-López, M., Corcho, O.: Ontological engineering: with examples from the areas of knowledge management, e-commerce and the Semantic Web, Springer, London ; New York (2004)

der Bundeswehr Universität München

Universität der Bundeswehr München
Chair of General Management and E-Business
E-Business & Web Science Research Group
Prof. Dr. Martin Hepp
Werner-Heisenberg-Weg 39
85577 Neubiberg
Germany

http://www.unibw.de/ebusiness/