Ontology-Based Data Quality Management: Methodology, Cost, and Benefits

Christian Fürber[1]

[1]E-Business & Web Science Research Group, Werner-Heisenberg-Weg 39,
85577 Neubiberg, Germany
{c.fuerber}@unibw.de

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. In my PhD research work, I aim at overcoming this bottleneck by developing a reference process for ontology-based data quality management and a matching efficiency estimation model for ontology-usage in data quality management scenarios.

Keywords: Ontologies, Ontology-based Data Quality Management, Cost and Benefits, Data Deficiencies

1 Problem

In today's business environment, data is the indispensable source for almost every transaction or decision. Performing business processes based on poor data quality can, therefore, directly account for expensive errors. The impact of poor data quality on the enterprise's business thereby ranges from dissatisfaction of customers and employees to unnecessary costs and missed revenues [3].

Recently, ontologies, i.e. partly formalized, consensual conceptualizations of a domain of interest, have been suggested as a promising means to assure data quality at a high level. They are expected to provide machine-readable access to knowledge about business rules and data handling [5]. Hence, through ontologies business transactions may be better equipped with relevant business knowledge to assure correct, automated processing and minimize data errors. But since the construction, maintenance, and population of ontologies is a difficult and costly task, ontology-based data quality improvement activities will not always be efficient [6]. Without adequate metrics and practices, decision makers cannot assess and guide the application of ontology-based techniques for a given business challenge. As of today, there is substantial uncertainty about efficient application scenarios for ontology-based data quality improvement techniques. An efficiency estimation model which allows cost and benefits estimation before applying ontology-based techniques for data quality improvement purposes could provide a better ground for informed decision making, potentially increase the adoption of ontology technology in industry, and reduce the risk of failed projects.

2 Related Work

In the past some research efforts have been addressed to the problem of applying ontologies for data quality improvement (DQI). But only a few findings have been achieved at estimating cost and benefits of ontology application.

Ontology-based DQI approaches can be separated into ontology-based data integration, ontology-based data retrieval, and ontology-based data cleaning. Ontology-based data integration and retrieval approaches mainly focus on reducing 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 of the retrieved data [14],[15],[17],[18],[19],[20]. An excellent ontology-based data retrieval approach, called COntext INterchange (COIN), has been developed by the Massachusetts Institute of Technology (MIT). COIN uses sub-queries for data retrieval from disparate sources which are combined by a context mediator for holistic data presentation. The context mediator is able to identify and reconcile semantic differences by accessing domain knowledge about the underlying sources, which is mainly represented in a shared ontology and context definitions [17]. Ontology-based data cleaning approaches aspire to detect and remove data deficiencies, such as inconsistencies, data duplicates, or violation of domain constraints, directly in the source [12],[13],[16]. Therefore, they utilize knowledge represented within ontologies. In [16] a so-called task ontology is used besides the domain ontology to describe tasks and methods of DQI for automatic identification of the underlying data deficiency type. After its identification the task ontology suggests appropriate improvement methods. The approach shown in [13] proposes more automation by counting the choices made for data correction. After passing a user-defined threshold, the most popular improvement algorithm is automatically applied for the accordant data deficiency. Hence, the manual effort decreases over usage time. Even though the ontology-based DQI approaches promise significant improvement of data quality, they lack the integration into a generic management process, e.g. total data quality management as proposed in [22], to enable pragmatic use for businesses purposes. On top, they only address a few data quality problems and do not offer any certainty about efficiency. Due to the novelty of this application area, there is currently no theoretical examination about the exact contribution of ontologies for data quality, i.e. about the potential effects of ontologies on data quality.

To the best of our knowledge, ONTOCOM [7] is currently the only model for cost estimation in ontology engineering. It is based on cost estimation models known from the software engineering area, and predicts costs related to activities performed during the ontological lifecycle. Some certainty about the costs related to the use of ontologies can be given by ONTOCOM, but it is tied to a sequential lifecycle of ontologies, and, as usual for cost estimation models, it does not regard any benefits or even provide an efficiency calculation model [7]. Menzies has elaborated some research findings also considering benefits of ontologies [21]. But his work does not supply a cost-benefit estimation model as well. In the special case of applying ontologies for data quality improvement, the benefits can be expressed best by the reduction in costs for poor data quality due to the usage of ontologies. Some general findings on data quality costs have been achieved by the data quality community [4],[8],[24],[25]. Additional approaches not mentioned in this section may be discovered with forthcoming research.

3 Proposed Approach and Methodology

The proposed research work aims at providing a framework for materializing the theoretically positive effects of ontologies for data quality improvement scenarios in real business cases based on a better understanding of their efficient usage. On the road towards this goal, three major research issues need to be examined:

- Understanding the technical and business impact of ontologies on data quality
- Developing a generic ontology-based data quality management (OBDQM) process
- Metrics and models for ex-ante cost-benefit estimation with the OBDQM-process based on a minimal set of problem characteristics

My PhD research work accounts for two perspectives on data quality. Seen from the data consumer data are of high quality if they are “fit for use” [1]. From the technical perspective high quality data are data that meet “conformance to specifications” [10] and are “free of defects” [11]. To examine the possible impact of ontologies on data quality, we first use the technical perspective and develop a typology of data deficiencies, i.e. data that contains defects or does not conform to specifications. By summarizing existing typologies of [9],[26], and [27], and comparison with possible effects of ontologies identified in [6] we analyze in detail the theoretical impact of ontologies to solve data quality problems. For example a domain ontology containing knowledge about the relevant business domain will likely to be helpful for mitigation of business domain constraint violations. As a result the application range of ontologies in the area of data quality improvement will be identified. In addition to our theoretical model trade-offs between data deficiencies need to be considered and analyzed, and connections need to be drawn between data deficiencies and the dimensions of data quality as perceived by data consumers [1].

Adjacent, a generic OBDQM framework will be developed focused on data deficiencies improvable by ontologies. Data quality management is not a one-time performed process. Instead, it has to support perpetually almost every business process. In our approach ontologies are the core part of the OBDQM process to maintain knowledge about business rules, data handling, possible data deficiencies, and improvement methodologies for continuous data quality assurance in the information flow of an enterprise. Hence, existing DQM Methodologies from data quality research [2],[4],[22], need to be extended by the use of ontologies and ontology-based DQI-methodologies. The new OBDQM-process shall cover all activities related to the ontology lifecycle and the data lifecycle including possible reuse of ontologies for other scenarios. Simultaneously, OBDQM shall leave room for customization according to the data quality strategy. Furthermore, the contribution of lightweight upper ontologies and the contribution of ontology-learning from relational structures will be evaluated regarding its use for data quality management purposes. As a result, it should be possible to identify cost objects and develop metrics for cost estimation.

Based on the OBDQM-process, a cost-benefit estimation model and its parameters required for ex ante estimation will be developed. The costs will include all cost categories related to OBDQM-activities including costs for ontology construction, maintenance, and population. Findings of ONTOCOM [7] will be helpful for this part of the research, but have to be extended substantially. The assessment of the benefits will built upon the identification of costs related to data quality shortcomings which can be avoided or mitigated through OBDQM-activities, and on additional benefits through ontology-usage outside of data quality management, e.g. benefits through documentation of business knowledge. The identification of ontologies' application range in improving data quality will help to identify the benefits of ontologies for data quality improvement. Findings from data quality research will support the estimation model concerning poor data quality costs [4],[8],[24],[25]. The estimation model will take into account the customization of the process and will build on such parameters that can realistically be determined in practical business settings. As a result, this research work should provide instructions on how to use ontologies efficiently for data quality management in typical real-world settings. The research findings will be evaluated by at least one case study applying OBDQM and comparing estimated values with actual costs and benefits. If successful, the PhD thesis will provide certainty and guidelines for practitioners about the efficient usage of ontologies in data quality management scenarios.

4 Conclusions, Results, and Future Work

Machine-readable knowledge is the key for effective management of data quality. Only with machine-readable knowledge about business processes and data handling the syntactic and semantic quality of the massive amount of data produced every day can be assured. Ontologies can provide such machine-readable knowledge and have already been applied to improve data quality. Existing approaches only address a small range of possible data deficiencies. Neither do they provide a management methodology, nor models for ex ante estimation of cost and benefits of the proposed techniques. Hence, my PhD thesis aims at providing a generic framework for ontology-based data quality management and a matching cost benefit estimation model.

The research project has just started in September 2008. Based on existing research work [3] a data lifecycle has been developed to gain a better understanding about the possible data deficiencies that might occur during data's life. Currently a holistic typology of data deficiencies is being developed covering extensional quality, i.e. the quality of data values [15], and intensional quality, i.e. the quality of the conceptual environment [15], to identify the toeholds for ontology-based DQI methodologies. By the time of the doctoral consortium I will be able to present a first draft of the theoretical model about ontology's potential impact on data quality and the application range of ontology-based DQI-methodologies.

Future work should be focused on the development of new ontology-based DQI-approaches to cover a broader set of data deficiencies in addition to the proposed methodology. One of our approaches intends to use a linguistic ontology, e.g. WordNet [23], to analyze and improve conciseness and consistency of database schema labels raising the intensional quality of data models.

5 References

1. Wang, R. Y., Strong, D. M.: Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems, 12(4), 5--33 (1996)
2 English, L. P.: Improving data warehouse and business information quality: methods for reducing costs and increasing profits. Wiley, New York (1999)
3. Redman, T. C.: Data quality for the information age. Artech House, Boston (1996)
4. Batini, C., Scannapieco, M.: Data quality: concepts, methodologies and techniques. Springer, Berlin (2006)
5. 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)
6. 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)
7. Simperl, E., Sure, Y.: The Business View: Ontology Engineering Costs. In: Hepp, M., De Leenheer, P., de Moor, A., Sure, Y. (eds.) Ontology Management: Semantic Web, Semantic Web Services, and Business Applications, pp. 207--225. Springer, New York (2008)
8. Eppler, M., Helfert, M.: A Framework for the Classification of Data Quality Costs and an Analysis of their Progression. In: 9th International Conference on Information Quality (IQ 2004), pp. 311--325, MIT, Cambridge (2004)
9. Hoxmeier, J. A.: Typology of database quality factors. Software Quality Control, 7(3-4), 179-193 (1998)
10. Kahn, B. K., Strong, D. M., Wang, R. Y.: Information quality benchmarks: product and service performance. Communications of the ACM 45(4), 184--192 (2002)
11. Redman, T. C.: Data quality: the field guide. Digital Press, Boston (2001)
12. 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)
13. Brüggemann, S.: Rule Mining for Automatic ontology-based Data Cleaning. In: 10th Asia-Pacific Web Conference, (2008)

14.Brüggemann, S.: Ontologiebasierte domänenspezifische Datenbereinigung in Data Warehouse Systemen. In: Grundlagen von Datenbanken. (2006)
15.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)
16.Wang, X., Hamilton, H. J., Bither, Y.: An ontology-based approach to data cleaning. Regina: Dept. of Computer Science, University of Regina (2005)
17.Madnick, S., Zhu, H.: Improving data quality through effective use of data semantics. Data & Knowledge Engineering, 59(2), 460--475 (2006)
18.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)
19.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)
20.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)
21.Menzies, T.: Cost benefits of ontologies. Intelligence 10(3), pp. 26--32 (1999)
22.Wang, R. Y.: A product perspective on total data quality management. Commun. ACM 41(2), 58--65 (1998)
23.Miller, G.: WordNet: An on-line lexical database. International Journal of Lexicography (3), 235--244 (1990)
24.Raeburn, V. P.: Understanding the Cost-Benefit Quality Curve. DM Review 18(12), p. 37. (2008).
25.Raneses, A., Mielke, M.: Information Quality Cost Curves: Empirical Evidence from PTR / SCR Costing. In: International Conference on Information Quality (2005)
26.Oliveira, P., Rodrigues, F., Henriques, P. R.: A Formal Definition of Data Quality Problems. In: International Conference on Information Quality (2005)
27.Rahm, E., Do, H.-H.: Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bulletin 23(4), 3-13 (2000)