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

Christian Fuerber[1]

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

Abstract. The performance of business processes and the degree of automation strongly depends on the quality of data in information systems. 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. But the construction and maintenance of ontologies is a costly task. Thus, ontologies can only gain practical relevance for data quality management if we manage to provide certainty about its efficient usage. In my PhD research project, I aim at overcoming this bottleneck by developing a technical framework for ontology-based data quality management and a matching efficiency estimation model for ex ante cost benefit estimation.

Keywords: Ontologies, ontology-based data quality management, cost and benefits, data deficiencies

1 Problem Statement and Research Goals

Performing business processes based on poor data quality can 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 [1]. 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 [2], e.g. for improved data retrieval, data integration, or data cleaning. But due to its complexity ontology-based data quality management (OBDQM) activities will not always be efficient [3]. Hence, this research project focuses on three major research issues to be examined:

1. Understanding the technical and business impact of ontologies on data quality
2. Development of an efficient architecture to utilize ontologies in relational database management system (RDBMS) landscapes for OBDQM
3. Development of metrics and models for ex-ante cost-benefit estimation for OBDQM based on a minimal set of problem characteristics

2 Proposed Approach

Following questions represent the proposed guideline of my PhD research:

- What data quality problems can be improved by ontologies?
- How can ontologies be integrated in RDBMS to manage data quality?
- What kind of ontologies should be used?
- What has to be represented to identify and mitigate data deficiencies in information systems?
- How can we estimate cost for ontology construction, maintenance, population, and usage?
- How can we estimate poor data quality cost mitigated through ontology usage?

From the technical perspective high quality data are data that meet "conformance to specifications" [4] and are "free of defects" [5]. Based on this definition data quality problems occurring in data values and schema elements of single and multiple source scenarios are collected. Similar to [6] the identified data quality problems and possible improvement methodologies are documented within a task ontology complemented by domain ontologies for domain specific aspects. The aspired technical architecture aims to apply ontology-based techniques on RDBMS during the three main sections of the data lifecycle: data acquisition, data storage, and data usage [1]. Finally, appropriate metrics for ex ante estimation will be developed regarding cost for ontology construction, maintenance, population, and usage on the one hand. On the other hand metrics for benefit estimation will be developed to estimate the reduction in poor data quality cost referred to ontology usage. The estimation of poor data quality cost will also consider the fall in value of the deficient process. It is intended to validate the OBDQM architecture on real world data of RDBMS. The cost benefit estimation model will be evaluated by at least one case study, in which actual costs and benefits of OBDQM will be measured and compared with estimated results.

Aknowledgements. I would like to thank Martin Hepp for the precious feedback and his dedication to support me in my research.

3 References

1. Redman, T. C.: Data quality for the information age. Artech House, Boston (1996)
2. 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)
3. 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)
4. 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)
5. Redman, T. C.: Data quality: the field guide. Digital Press, Boston (2001)
6. Wang, X., Hamilton, H. J., Bither, Y.: An ontology-based approach to data cleaning. Regina: Dept. of Computer Science, University of Regina (2005)