RESEARCH

Christian Fürber

Data Quality Management with Semantic Technologies

Springer Gabler

Christian Fürber

Data Quality Management with Semantic Technologies

Foreword by Prof. Dr. Martin Hepp

Springer Gabler

Christian Fürber
München, Germany

Dissertation Universität der Bundeswehr München, Neubiberg, 2015

OnlinePlus material to this book can be available on
http://www.springer-gabler.de/978-3-658-12224-9

ISBN 978-3-658-12224-9 ISBN 978-3-658-12225-6 (eBook)
DOI 10.1007/978-3-658-12225-6

Library of Congress Control Number: 2015959354

Springer Gabler

Printed on acid-free paper

Springer Gabler is a brand of Springer Fachmedien Wiesbaden
Springer Fachmedien Wiesbaden is part of Springer Science+Business Media
(www.springer.com)

Data Quality Management with Semantic Technologies

Für Tanja

Foreword

In contrast to physical machinery, computer-based information systems operate on the basis of *models of reality*. While traditional forms of automated systems directly handle the actual objects of a task, computers must rely on representations of the input objects of processing, and they return representations of the results when they are done. For the information to be processed, these representations are in the form of digital data, and for the details of the processing, they are computer programs, i.e. executable instructions.

By being models, both computer data and computer programs are purpose-bound abstractions of reality, and their appropriateness can only be judged in the light of the information processing task at hand.

Now, the overall reliability of an information system critically depends on how well the data represents the relevant subset of reality, and on how well the computer programs represent appropriate processing steps. This is valid for all computer-based information processing, from the most simplistic digital weather station up to the complex transaction support systems in entire value chains. This sounds like a triviality, but even if it was, it is an important one, because it helps understand the origin of many practical problems of computer information systems.

Reality shows that our ability to use computers for the automation of business processes is severely limited by our ability (1) to represent information and processing instructions properly in the form of data and computer programs, and (2) to keep these artifacts in alignment with the ever-changing reality. Our customers move from one address to another, while our customer database will typically contain at least some outdated addresses. Product designs change, but almost every Web shop will, every now and then, show outdated product images and product descriptions, and the picture of me on my university Web page does obviously not match with how I really look while writing this foreword. Data and programs are human-made artifacts, and they do not automatically align with changes in the environment they represent.

This problem at the interface between reality and the representations of reality in the components of computer information systems is one of the root causes whenever computers do not behave as we expect them to do: When they make wrong decisions, provide wrong information, or cancel business processes unexpectedly. If a customer database contains outdated address data, a shipment to that address will fail, if the weight of a product in data differs from the actual weight, incorrect shipping charges will be computed; and if a part number for consumables or spare parts is missing in a database of inventory, the automatic procurement of those items will fail. Relevant examples can be found in every major organization.

Since the 1990s, the systematic analysis of the quality of computer data has become an established field of research, known as "Data Quality Management" (DQM), and its broader notion "Information Quality Management" (IQM). One of the early works on this topic is the thesis by Mark David Hansen, entitled "Zero Defect Data"[1], published in 1991. In the following years, numerous theoretical concepts, technical solutions and practical implementations have emerged. In business practice, there is a wealth of products and services available that promise to systematically improve the quality of data or information in enterprises and value chains.

Sadly, though, data quality in many organizations is still insufficient. One reason for this is that the interface between reality and representations of reality in computer systems is itself not accessible for computer-based solutions. In essence, a computer program cannot determine whether its components properly represent reality, because it lacks a sufficient sensory apparatus. For instance, a Web application that supports declaring your income tax cannot validate whether its processing matches the latest state of the tax laws. Admittedly, computers can increasingly validate the consistency *within* those representations, e.g. spot outliers in data based on statistical approaches or compute logical contradictions within formally specified business models. Still, the interface between reality and the models of reality itself remains inaccessible to them.

Typical approaches in data quality management therefore focus either (1) on helping human actors to better collect and maintain data and process specifications, or (2) on spotting and correcting problems within the model world of a computer, as in the validation of data based on syntactic validation rules.

In computer science, the fundamental problem of the interface between reality on one hand and models of reality inside computers on the other has been studied for about 20 years under the term "ontologies". Ontologies are specifications of models of reality that aim at being consensual among many people and applicable to a broad range of

[1] Hansen, M. (1991): Zero Defect Data. MSc thesis, Sloan School of Management, Cambridge, Mass. (USA): MIT, http://hdl.handle.net/1721.1/13812.

scenarios. They typically include at least some formal axioms and the underlying modeling decisions are influenced by philosophical principles, e.g. regarding the identity and unity of objects. The formal axioms enable a computer to spot contradictions in the models, draw additional conclusions, and to automatically translate between multiple data models of the same subject area, at least to a certain degree. The philosophical grounding can increase the general validity of the model.

Ontologies are a promising attempt to improve the consistency and accuracy of models of reality. While they do not take away the fundamental barrier between reality and the model world of computers, because they are models themselves, they add a formally specified and philosophically grounded intermediate level, which can reduce the problem.

In 2001, Berners-Lee, Lassila and Hendler applied the idea of ontologies in computer science to the problem of information interchange on the World Wide Web and described the vision of a "Semantic Web", in which computers are increasingly able to process information at the level of meaning[2].

In this thesis, Christian Fürber analyzes the use of the ideas and technological components of the Semantic Web, in particular ontologies, for better data quality management. His approach is characterized by the following two main innovations.

(1) While traditional data quality management formulates requirements and metrics at the very low level of system-specific database schemas, he lifts these to a generic, business-level understanding of a domain of interest.

(2) He proposes the use of a Semantic-Web-powered Wiki for organizing the elicitation and management of validation rules and metrics, thus increasing the inclusion of domain experts into these processes.

In essence, this approach can increase the quality and reusability of data quality knowledge. It will be easier for domain experts to be involved, it will be less effort to validate the consistency of data quality rules and metrics, and the rules and metrics can be applied to a broad set of data sources, because they abstract from the implementation details of a particular database schema.

[2] Berners-Lee T., Hendler J., Lassila O. (2001): The Semantic Web. Scientific American. 284(5): 28-37.

The topic of this thesis is practically relevant to almost any organization, and the proposed solution is a very promising application of the Semantic Web technology stack to real-world problems. I sincerely recommend this work and am confident it can help improve both our understanding and the state of implementations of data quality management as a whole.

Dr. Martin Hepp

Professor of General Management and E-Business
Universität der Bundeswehr München

Preface

As this thesis is being published, we are in the middle of the digital age in which people utilize their mobile devices to permanently share and consume data, while society still struggles with data protection issues and credibility of information. Moreover, we are entering an age, in which the massive amount of data is being used to increase the degree of automation and to precisely predict future events. Data quality issues will more and more hinder these developments, unless suitable architectures will be provided that help to reduce them.

This dissertation, therefore, describes an innovative way on how to manage data quality by utilizing knowledge representation and processing technologies which have been brought up by the Semantic Web initiative of the World Wide Web Consortium (W3C) and the Semantic Web research community. Based on a literature analysis of typical data quality problems and typical activities of data quality management processes, I developed the Semantic Data Quality Management (SDQM) framework as a major contribution of this thesis. The SDQM framework consists of three major components:

(1) an ontology for the machine-readable representation of quality-relevant knowledge,
(2) a semantic wiki that is connected to the ontology to facilitate structured capturing of quality-relevant knowledge, and
(3) a Web-based reporting frontend for data quality monitoring and assessment based on the captured knowledge.

The framework has been evaluated in three different use cases based on real-world data. Moreover, we compared SDQM with conventional data quality software to identify strengths and weaknesses of the approach. Besides technical results, this thesis delivers four theoretical findings, namely

(1) a comprehensive typology of data quality problems of information systems and Semantic Web data,

(2) ten generic data requirement types,

(3) a requirement-centric data quality management process fitted to the needs of the SDQM framework, and

(4) an analysis of related work.

This dissertation would not have been possible without the support of my family, colleagues, and friends. Therefore, I would like to thank my supervisors, Prof. Dr. Martin Hepp and Prof. Dr. Michael Eßig, for the precious discussions, their guidance, and their dedication to support my thesis project.

Moreover, I would like to thank Andreas Radinger, Alex Stolz, Dr. Mouzhi Ge, Uwe Stoll, Dr. Bene Rodriguez-Castro, Leyla Jael García-Castro, Prof. Dr. Heiner Stuckenschmidt, Dr. Holger Knublauch, and everyone else from the Semantic Web community who supported me with valuable hints and discussions.

I would also like to thank my parents, Magrit and Claus-Dieter Fürber, for encouraging me to always follow my passion. But most of all, I have to thank my wife Tanja for her love and support over all these years and for giving me the freedom to spend so much time on this thesis.

Dr. Christian Fürber

List of Abbreviations

BIS	=	Business Information Systems
COIN	=	Context Interchange
CPU	=	Central Processing Unit
CRM	=	Customer Relationship Management
CSV	=	Comma-separated Value
DIKW	=	Data, Information, Knowledge, Wisdom
DQ	=	Data Quality
DQM	=	Data Quality Management
DSRM	=	Design Science Research Methodology
DSV	=	Delimiter-separated Values
DTD	=	Document Type Definition
ETL	=	Extraction, Transformation, and Loading
FDR	=	Functional Dependency Rule
FN	=	False Negative
FP	=	False Positive
FuncDepReferenceRule	=	Functional Dependency Reference Rule
HTTP	=	Hyper Text Transfer Protocol
IQ	=	Information Quality
IP	=	Information Product
IS	=	Information System
ISO	=	International Organization for Standardization
JSON	=	JavaScript Object Notation
KPI	=	Key Performance Indicator

LOD	=	Linked Open Data
MDM	=	Master Data Management
MIT	=	Massachusetts Institute of Technology
OS	=	Open Studio
OWL	=	Web Ontology Language
OXC	=	Ontology-based XML Cleaning
PHP	=	Hypertext Preprocessor
RDB	=	Relational Database
RDBMS	=	Relational Database Management System
RDF	=	Resource Description Framework
RDFS	=	RDF Vocabulary Description Language
RQ	=	Research Question
SCROL	=	Semantic Conflict Resolution Ontology
SDQM	=	Semantic Data Quality Management Framework
SDQMgr	=	Semantic Data Quality Manager
SMDM	=	Semantic Master Data Management
SMW	=	Semantic MediaWiki
SPARQL	=	SPARQL Protocol and RDF Query Language
SPIN	=	SPARQL Inferencing Notation
SQL	=	Structured Query Language
SSN	=	Social Security Number
SWRL	=	Semantic Web Rule Language
Talend OS for DQ	=	Talend Open Studio for Data Quality
TDQM	=	Total Data Quality Management

TIQM	=	Total Information Quality Management
TQDM	=	Total Quality Data Management
TP	=	True Positive
TSV	=	Tab-separated Values
UDF	=	User-defined Function (SPARQL)
URI	=	Uniform Resource Identifier
URL	=	Uniform Resource Locator
W3C	=	World Wide Web Consortium
WWW	=	World Wide Web
XML	=	Extensible Markup Language