Editorial: Special Issue on Web Data Quality

We are delighted to present this special issue of the *Journal of Data and Information Quality* on web data quality. This issue includes four innovative research articles covering the areas of web data profiling, web data quality assessment, and web data cleansing.

Over the last few years, the volume and variety of data that is available on the Web has risen sharply. In addition to traditional data sources and formats such as CSV files, HTML tables, and deep web query interfaces, new techniques such as microdata, RDFa, microformats, and linked data have found wide adoption. In parallel, techniques for extracting structured data from web text and semistructured web content have matured resulting in the creation of large-scale knowledge bases such as NELL, YAGO, DBpedia, and the Knowledge Vault. Independent of the specific data source or format or information extraction methodology, data quality challenges persist in the context of the web. Applications are confronted with heterogeneous data from a large number of independent data sources while metadata is sparse and of mixed quality. Before one can utilize the data, a potential user must first overcome the challenges of handling a wide range of quality issues in the available data and metadata.

In this setting, this special issue of the *Journal of Data and Information Quality (JDIQ)* was organized with the goal to present the latest research on web data profiling, web data quality assessment, and web data cleansing. We received 11 submissions for the special issue. The papers went through a rigorous procedure of review involving at least three expert reviewers for each paper. After two rounds of review, we selected four high-quality research papers that made contributions to both research and practice.

Automated Quality Assessment of Metadata across Open Data Portals, by Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres, investigates the quality of metadata that describes datasets, provided by open-license dataset catalogues. For this task, the authors first map the metadata models of the catalogs to the W3C Data Catalog Vocabulary (DCAT) schema. Based on this shared model, they define a set of relevant quality dimensions and report comprehensive empirical findings concerning the quality of the provided metadata. The article presents the first systematic investigation into the quality of metadata on open data portals and thus contributes to laying the foundation for developing better data search and data integration methods for these portals.

Towards More Accurate Profiling of Deployed schema.org Microdata, by Robert Meusel, Dominique Ritze, and Heiko Paulheim, is based on the insight that the microdata that has been used extensively by major search engines to improve the user experience, can simultaneously be used to develop a corpus profile. However, while a corpus such as the Web Data Commons (WDC) supports a common schema and vocabulary (i.e., schema.org), there are many data quality challenges around duplicate detection and schema errors that must be addressed. The authors consider a number of metrics such as the count of providers for a class, the frequency of duplicate entities for each provider, and so on. They also apply semantics-based identity resolution to

find references to the same real-world object. This research is able to demonstrate the significant improvement of accuracy of the subsequent corpus profile.

Luzzu—A Methodology and Framework for Linked Data Quality Assessment, by Jeremy Debattista, Sören Auer, and Christoph Lange, recognizes that assessing the quality of data in today's complex and rich landscape usually requires quality measures across multiple dimensions. In response, they present the Luzzu modular open-source framework and methodology. They attempt ambitious goals of being scalable, extensible, interoperable, and customisable. Luzzu is accompanied by a set of ontologies for capturing quality-related information. It allows users to customize and prioritize quality measures.

BayesWipe: A Scalable Probabilistic Framework for Improving Data Quality, by Sushovan De, Yuheng Hu, Vamsi Meduri, Yi Chen, and Subbarao Kambhampati, addresses a data cleaning solution that can scale to big data without reliance on the availability of clean data for training or an expert to generate training examples. BayesWipe assumes that a statistical process underlies the generation as well as the corruption of data (errors). Consequently, noisy data can be used to learn the parameters of both the generative and error models, eliminating the dependence on the existence of clean master data. The authors also address how to efficiently perform consistent query answering over a dirty database.

We wish to thank Editor-in-Chief Louiqa Raschid and Senior Associate Editor Paolo Missier for their support and guidance throughout this process, and Ms. Janet Cavanagh, administrative assistant, for her help in coordinating this special issue. We are grateful to the authors who have contributed the four articles included in this issue. Finally, this issue would not have achieved its high level of quality without the diligence and insights of the over twenty five reviewers who provided high quality reviews.

Maribel Acosta
Charu Aggarwal
John Breslin
Volha Bryl
Anup Chalamalla
Xu Chu
Oscar Corcho
Luis Galárraga
Jing Gao
Anna Lisa Gentile
Oktie Hassanzadeh
Melanie Herschel
Dimitris Kontokostas
Magnus Knuth
Christoph Lange
Xian Li
Paolo Missier
Carlos Ordonez
Mourad Ouzzani
Heiko Paulheim
Axel Polleres
Theo Rekatsinas
Anisa Rula

Shazia Sadiq
Ina Schieferdecker
Nan Tang
Herbert Van de Sompel

Enjoy reading the special issue!

Christian Bizer
University of Mannheim, Germany

Luna Dong
Amazon, USA

Ihab Ilyas
University of Waterloo, Canada

Maria-Esther Vidal
Universidad Simon Bolivar, Venezuela, and Fraunhofer IAIS, Germany

Guest Editors