Editorial: Special Issue on Improving the Veracity and Value of Big Data

We are delighted to present this special issue of the *Journal of Data and Information Quality on Improving the Veracity and Value of Big Data*. This issue includes four innovative research articles from areas that include linked data, context-aware data quality, knowledge base construction, and sequence clustering.

Big data poses serious challenges with regard to data quality. Indeed, data quality approaches need to be revisited to deal with large quantities of data (volume), to accommodate the dynamic aspects of data (velocity), to be able to uniformly deal with heterogeneous data originating from different sources (variety), to assess and improve the accuracy of data (veracity), and to provide an indication of the impact of data quality for both decision making and monetary aspects (value).

In this special issue of *JDIQ*, we provide a representative collection of insightful articles at the *intersection of data quality and big data*, from theory to practice, with a focus on improvements in veracity and value.

The accepted contributions cover a range and diversity of angles within this broader scope: a theoretical perspective on an Ontological Multidimensional Data Model to enable quality assessment of multidimensional data (Bertossi and Milani), algorithmic methods for assessing the connectivity (Mountantonakis and Tzitzikas) and the veracity (Esteves, Rula, Reddy and Lehmann) of RDF datasets, and, finally, an experimental evaluation of scalability versus accuracy tradeoffs specifically for biological databases (Chen, Wan, Yang, Zobel, and Verspoor).

The articles went through a rigorous procedure of review involving at least three expert reviewers for each article. After two rounds of review, we selected four articles that made contributions to both research and practice.

“Scalable Methods for Measuring the Connectivity and Quality of Large Numbers of Linked Datasets,” by Michalis Mountantonakis and Yannis Tzitzikas, concerns the measuring of connectivity of linked data (LOD). Such connectivity information is crucial for obtaining complete information about a set of entities, thereby adding value to these huge data collections. Furthermore, it provides crucial information for cleaning linked data. To handle large, dynamic, and heterogeneous sets of entities, indexes are designed to speed up the measurement of connectivity. Moreover, the presented indexes can be constructed on top of parallel computation platforms such as MapReduce. An extensive experimental evaluation is presented in which, among other things, the connectivity of the LOD cloud, consisting of billions of triples, is measured.

“Ontological Multidimensional Data Models and Contextual Data Quality,” by Leopoldo Bertossi and Mostafa Milani, provides a theoretical framework for modelling data quality assessment and cleaning as context-dependent activities. By mapping dirty data into a contextual ontology (expressed in terms of $\text{datalog}^{+/-}$), a whole range of possible quality versions of the data are obtained. At query time, one wants to obtain query results that hold in all of these clean versions. That is, the certain answers should be returned. A generalized multidimensional data model underlies the entire framework, allowing categorical attributes to be linked to dimensional categories, accommodating incomplete data, dimensional navigation, and a logical interpretation of

dimensional data and metadata. Computational properties of the framework are derived with regards to certain query answering. The framework is illustrated by various challenging case scenarios.

“Toward Veracity Assessment in RDF Knowledge Bases: An Exploratory Analysis,” by Diego Esteves, Anisa Rula, Aniketh Reddy, and Jens Lehmann, empirically evaluates techniques for estimating the veracity of knowledge bases. The research takes place in the context of the DeFacto fact validation system, which verbalizes claims from a knowledge base and searches the web for evidence that supports or refutes the claims. The article includes studies, using benchmark datasets, into a variety of techniques used within fact validation, including verbalization, negation, named entity recognition, and sameAs annotations.

“Comparative Analysis of Sequence Clustering Methods for De-Duplication of Biological Databases,” by Qingyu Chen, Yu Wan, Lei Yang, Justin Zobel, and Karin Verspoor, empirically evaluates two prominent clustering algorithms for biological sequences. Such algorithms are used to support both search and curation of large collections of sequences, and a detailed review of clustering for sequences can potentially provide insights for the use of clustering in other domains. The evaluation, which involves a recent release of the Swiss-Prot protein sequence database, applied both domain-independent metrics, such as cohesion and separation, and domain-specific metrics, such as the consistency of functional annotations. Insights are drawn that can inform the selection of sequence clustering algorithms, for example, relating to the tradeoff between efficiency and accuracy and the impact of the role of the clustering on algorithm selection and configuration.

We thank the previous Editor-in-Chief Louiqa Raschid for her guidance throughout this process and Ms. Janet Cavanagh, administrative assistant, for her help in coordinating this special issue. Finally, we express our thanks to the authors and reviewers, without whose input the issue would not have been possible.

Floris Geerts
University of Antwerp, Belgium

Paolo Missier
Newcastle University, UK

Norman Paton
University of Manchester, UK

Guest Editors