Editorial: Special Issue on Metadata Discovery for Assessing Data Quality

It is a pleasure for us to introduce this special issue of the *Journal of Data and Information Quality* on *Metadata Discovery for Assessing Data Quality*. This issue includes five novel research articles on guidelines, methodologies, and tools to process metadata from several types of data sources, aiming to use them for assessing data quality.

Modern big data sources contain several hidden metadata that can be exploited to assess the quality of data. Among them, value patterns and their distributions, functional and inclusion dependencies, and key and foreign key candidates are well-known examples, and can be used to detect several data quality violations. However, other than many different types of metadata, in modern big data settings it might be necessary to extract metadata not only from well-structured relational data sources but also from graphs and several other types of sources, including social networks and IoT ecosystems. This makes the extraction and organization of metadata a particularly complex process, requiring the development of novel heuristics and distributed computational paradigms to make it computationally tractable.

In this special issue, we provide a representative collection of insightful articles, including novel metadata extraction techniques from several types of data sources, and application papers providing useful feedbacks and lessons learned in relevant big data application domains. In particular, the accepted contributions include one technique for extracting metadata concerning database schemas (Beneventano, Bergamaschi, Gagliardelli, and Simonini) aiming to detect schema alignment properties, and one technique for extracting relaxed functional dependencies from data instances in an incremental way (Cirillo and Caruccio), opening the way to the development of distributed computing paradigms for extracting such complex metadata, approaches dealing with metadata associated to graph structures: a data profiling technique to extract metadata from property graphs (Maiolo, Etcheverry, and Marotta) and a set of logical rules aiming to improve the quality of data processed within public knowledge graphs (Ahmadi, Truong, Dao, Ortona, and Papotti), and finally, a report on the management of several types of metadata related to open data sources used in smart city contexts (Lämmel, Dittwald, Bruns, Tcholtchev, Glikman, Cuni, Flügge, and Schieferdecker).

All of the articles went through a rigorous procedure of review involving three expert reviewers for each article. After two rounds of review, we selected five articles that made relevant contributions to both research and practice.

“BLAST2: An Efficient Technique for Loose Schema Information Extraction from Heterogeneous Big Data Sources,” by Domenico Beneventano, Sonia Bergamaschi, Luca Gagliardelli, and Giovanni Simonini, presents the technique BLAST2 for extracting metadata concerning the schemas of datasets, aiming to improve the scalability of schema alignment processes, reduce their complexity, and increase the quality of results. BLAST2 extends the previous version BLAST by

ACM Reference format:

Giuseppe Polese, Vincenzo Deufemia, and Shaoxu Song. 2020. Editorial: Special Issue on Metadata Discovery for Assessing Data Quality. *J. Data and Information Quality* 12, 4, Article 17 (October 2020), 2 pages. <https://doi.org/10.1145/3423321>

© 2020 Copyright held by the owner/author(s).

1936-1955/2020/10-ART17

<https://doi.org/10.1145/3423321>

increasing speed and improving precision. The authors experimentally evaluated BLAST2 on publicly available real-world datasets, including 7 datasets about movies and 10 datasets about books, showing that it outperforms current state-of-the-art (meta-)blocking techniques.

“[Incremental Discovery of Imprecise Functional Dependencies](#),” by Stefano Cirillo and Loredana Caruccio, faces the problem of incremental discovery of metadata, focusing on relaxed functional dependencies, which are particularly useful metadata for assessing the quality of data. An incremental discovery strategy is necessary in the big data context so as to avoid to restart the whole discovery process whenever the data instance evolves. To this end, the authors propose the BIRD (Bit-vector-based Incremental RFD_e Discoverer) algorithm, capable to incrementally update the set of holding dependencies upon the insertion of new tuples in a dataset, checking whether they invalidate previously holding dependencies, and possibly generating new candidates. BIRD relies on novel data structures and a parallel execution strategy, and it is experimentally validated on several publicly available real-world datasets.

“[Data Profiling in Property Graph Databases](#),” by Sofia Maiolo, Lorena Etcheverry, and Adriana Marotta, proposes a set of data profiling tasks for property graph databases, including a taxonomy to classify them. In addition, the authors provide a formal specification method for the proposed profiling tasks based on the Property Graph model, and a pipeline to guide the application of the proposed tasks. Some of these tasks have been implemented in the query language Cypher and executed on a dataset derived from the case study presented in the article.

“[RuleHub: A Public Corpus of Rules for Knowledge Graphs](#),” by Naser Ahmadi, Thi-Thuy-Duyen Truong, Le-Hong-Mai Dao, Stefano Ortona, and Paolo Papotti, introduces RuleHub, a system embedding more than 7,000 rules for managing knowledge graphs. Such rules are openly available through a Web portal that allows users to retrieve those needed for their tasks, add new ones, or refine/annotate some of the available rules. Since metadata play a fundamental role in this context, the authors provide a new confidence measure enabling the evaluation of the quality of negative rules. Finally, the authors perform a comparative evaluation against human computed confidences, showing that the proposed measure performs suitable quality estimations.

“[Metadata Harvesting and Quality Assurance within Open Urban Platforms](#),” by Philipp Lämmel, Benjamin Dittwald, Lina Bruns, Nikolay Tcholtchev, Yuri Glikman, Silke Cuni, Mathias Flügge, and Ina Schieferdecker, reports the lesson learned from the development of open data platforms in the smart city domain. In particular, the authors focus on the quality of metadata extracted from large public datasets, also including data streams originating from IoT devices, verifying how they fulfill the following three main properties: findability, comprehensibility, and further processability of the data. The reported experience concerns the automatic extraction and quality assurance of metadata from a public dataset of the German government, also providing a set of guidelines for public servants and officers toward the quality provisioning of metadata.

We thank the Editor-in-Chief Tiziana Catarci for her guidance throughout this process and Dr. Andrea Marrella, administrative assistant, for his 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.

Giuseppe Polese
University of Salerno, Italy

Vincenzo Deufemia
University of Salerno, Italy

Shaoxu Song
Tsinghua University, China

Guest Editors