Editorial: Special Issue on Quality Assessment of Knowledge Graphs Dedicated to the Memory of Amrapali Zaveri

ANISA RULA, University of Milano-Bicocca, Italy and University of Bonn, Germany
AMRAPALI ZAVERI, Maastricht University, The Netherlands
ELENA SIMPERL, King's College London, United Kingdom
ELENA DEMIDOVA, L3S Research Center, Leibniz Universität Hannover, Germany

This editorial summarizes the content of the Special Issue on Quality Assessment of Knowledge Graphs of the Journal of Data and Information Quality (JDIQ). We dedicate this special issue to the memory of our colleague and friend Amrapali Zaveri.

ACM Reference Format:
Anisa Rula, Amrapali Zaveri, Elena Simperl, and Elena Demidova. 2020. Editorial: Special Issue on Quality Assessment of Knowledge Graphs Dedicated to the Memory of Amrapali Zaveri. 1, 1 (September 2020), 3 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

We have recently witnessed a rise in the number and scale of knowledge graphs, including YAGO [7], DBpedia [6], Wikidata [3], EventKG [5] and the Google Knowledge Vault [1]. Knowledge graphs are essentially repositories of graph-based data [4]. They store millions of statements about entities of interest in a domain, for instance, people, places, organisations and events. They are extensively used in various AI contexts, from search and natural language processing to data integration, as a means to add context and depth to machine learning and generate human-readable explanations.

Building and using knowledge graphs come with several challenges: they draw content from multiple sources (variety), in large quantities (volume) and at speed (velocity). For these reasons, they may include outdated or inconsistent information (veracity), which affects the applications developed on top of them [5], [9], [4]. The quality of knowledge graphs is hence an ongoing concern, which so far has not received enough attention from the research community [8], [4]. By comparison, there is a substantial body of literature exploring, assessing and repairing the quality of other types of data sources, [2], [10], which consider aspects such as completeness, accuracy, timeliness, consistency, and absence of duplicates. These are not directly applicable to knowledge graphs because of their volume, variety, velocity and veracity characteristics. Knowledge-graph profiling [2], i.e. the extraction of metadata about knowledge graphs, is a step in the right direction, focusing and guiding quality assessment efforts towards particularly challenging types of data, including spatio-temporal information, events or information in several languages.

Authors' addresses: Anisa Rula, University of Milano-Bicocca, Italy, University of Bonn, Germany, anisa.rula@unimib.it; Amrapali Zaveri, Maastricht University, The Netherlands, amrapali.zaveri@maastrichtuniversity.nl; Elena Simperl, King's College London, United Kingdom, elena.simperl@kcl.ac.uk; Elena Demidova, L3S Research Center, Leibniz Universität Hannover, Germany, demidova@L3S.de.

XXXX-XXXX/2020/9-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

This special issue has sought contributions from the research and practitioners' community on novel approaches to assess, monitor, maintain and improve the quality of knowledge graphs, including methods that work directly on the graph data, but also profiling and user interaction frameworks and implementations. The aim is to provide an overview of the state of the art which knowledge graph developers can use to understand and fix quality issues.

2 ARTICLES INCLUDED IN THE SPECIAL ISSUE

For this special issue, we accepted three articles, which illustrate the complexity of the topic and provide complementary approaches to tackle it.

The article "Mining Expressive Rules in Knowledge Bases", by Naser Ahmadi, Vamsi Meduri, Stefano Ortona, and Paolo Papotti, presents RuDiK, a system for rule mining on RDF knowledge graphs. RuDiK can mine positive, conditional, and negative rules, to deal with issues such as inconsistent or incomplete knowledge. The article details the main elements of RuDiK, namely the negative example generation, the scoring function, and the strategy for rule search and selection. The extensive experimental evaluation shows that RuDiK outperforms the state-of-the-art approaches in terms of precision and runtime performance, which makes it a new state of the art in rule mining.

The article "What are Links in Linked Open Data? A Characterization of Links between Knowledge Graphs on the Web", by Armin Haller, Javier D. Fernandez, Maulik R. Kamdar, Axel Polleres, discusses the adoption of FAIR principles when publishing knowledge graphs as linked data. The authors first define the concept of a dataset in this context based on the notion of a namespace authority and distinguish between different link types, such as ontology and instance links. Furthermore, they conduct an extensive empirical analysis of the links in the current LOD cloud. To facilitate efficient analytics, the authors propose a wider adoption of the RDF Header-Dictionary-Triples (HDT) format and recommend the inclusion of dataset namespace authority and link statistics in the dataset profile.

The article "Content-based Union and Complement Metrics for Dataset Search over RDF Knowledge Graphs", by Michalis Mountantonakis and Yannis Tzitzikas, studies the problem of indexing and ranking RDF knowledge graphs (or datasets) to support content-based search. The work focuses on ranking sets of datasets that satisfy the search requirements of users in terms of coverage, complementarity, and uniqueness of datasets. The approach constructs inverted indexes that exploit entity and class semantics. Then, the approach leverages union- and complement-based metrics to measure the coverage, relative enrichment, and uniqueness of sets of RDF knowledge graphs. As the problem of computing and comparing combinations of dataset lattices is exponential regarding the number of indexed RDF knowledge graphs, the authors propose a heuristic for pruning and regrouping index entries to reduce the search space. The authors present theoretical results on the time and space complexity of the proposed approach as well as an empirical evaluation. The experiments conducted over 400 datasets confirm the effectiveness of the proposed solution.

We take this opportunity to sincerely thank the authors for their invaluable and inspiring contributions to the special issue. We are also grateful to the members of the International Editorial Review Board for reviewing the submissions and helping us publish an interesting special issue, as well as to Andrea Marrella for their advice and continuous support throughout the publication process.

Last but not least important, we would like to thank all reviewers for their valuable and careful review work that made the publication of this special issue a success. A special thank you goes to Paolo Missier, a Senior AE of JDIQ, as he helped a lot in the reviewing process.

3 REMEMBERING AMRAPALI ZAVERI

Amrapali Zaveri was a postdoctoral researcher at Maastricht University in the Institute of Data Science since January 2017. She officially received the Maastricht University Teaching Qualification (UTQ/BKO) certificate. She was about to join the University of Amsterdam as an Assistant Professor. Previously, she was a postdoctoral researcher at Stanford University for one and a half years. She received her PhD from the University of Leipzig, Germany in 2015. Her research interests included data quality, knowledge interlinking and fusion, biomedical and health care research. She conducted a comprehensive survey of the existing data quality assessment methodologies currently available to evaluate the quality of linked datasets. Additionally, she evaluated crowdsourcing methodologies for the assessment and improvement of linked data quality as well as biomedical metadata quality. She was working on finding the optimal balance between machine learning, crowdsourcing with non-experts and experts towards data quality assessment. Additionally, she has been a guest co-editor in the special issue on web data quality in the IJSWIS journal and guest co-editor of a special issue on quality management of semantic web assets (data, services and systems) in the Semantic Web Journal. She has served as a PC member for the 1st workshop on linked data quality and as co-organizer for the 2nd, 3rd and 4th Workshop on Linked Data Quality held at ESWC from 2015 to 2017. She has shaped the semantic web community in many ways, as a role model, and as a promoter of diversity in STEM. She was a great mentor to her students and to her peers, filling them with ideas and encouragement. She tackled her work with vigour and determination. She was true to herself and her friends and family. She travelled to eighteen countries for paper presentations, conferences, and managed to explore them despite tight schedules. The co-editors of this special issue are honoured to dedicate it to her memory.

REFERENCES

[1] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 601–610. ACM, 2014.

[2] Mohamed Ben Ellefi, Zohra Bellahsene, John G. Breslin, Elena Demidova, Stefan Dietze, Julian Szymanski, and Konstantin Todorov. RDF dataset profiling - a survey of features, methods, vocabularies and applications. *Semantic Web*, 9(5):677–705, 2018.

[3] Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandecic. Introducing wikidata to the linked data web. In *The 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, volume 8796 of *Lecture Notes in Computer Science*, pages 50–65. Springer, 2014.

[4] Michael Färber, Frederic Bartscherer, Carsten Menne, and Achim Rettinger. Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web*, 9(1):77–129, 2018.

[5] Simon Gottschalk and Elena Demidova. EventKG - the hub of event knowledge on the web - and biographical timeline generation. *Semantic Web*, 10(6):1039–1070, 2019.

[6] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195, 2015.

[7] Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. YAGO3: A knowledge base from multilingual Wikipedias. In *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings*. www.cidrdb.org, 2015.

[8] Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8(3): 489–508, 2017.

[9] Daniel Ringler and Heiko Paulheim. One knowledge graph to rule them all? analyzing the differences between dbpedia, yago, wikidata & co. In *KI 2017: Advances in Artificial Intelligence - 40th Annual German Conference on AI, Dortmund, Germany, September 25-29, 2017, Proceedings*, volume 10505 of *Lecture Notes in Computer Science*, pages 366–372. Springer, 2017.

[10] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93, 2016.