Semantic Web 14 (2023) 3–4
DOI 10.3233/SW-223261
IOS Press

Editorial of knowledge graphs validation and quality

Jose Emilio Labra Gayo [a,*], Anastasia Dimou [b], Katherine Thornton [c] and Anisa Rula [d]

[a] *Dept. Computer Science, University of Oviedo, Spain*
E-mail: labra@uniovi.es
[b] *KU Leuven, Belgium*
E-mail: anastasia.dimou@kuleuven.be
[c] *Yale University Library, United States*
E-mail: katherine.thornton@yale.edu
[d] *Università Degli Studi Di Brescia, Italy*
E-mail: anisa.rula@unibs.it

Editors: Pascal Hitzler, Kansas State University, USA; Krzystof Janowicz, University of California, Santa Barbara, USA

1. Preface

Knowledge Graphs are one of the main recent advances that embody the Semantic Web view [1]. They store millions of statements about entities of interest in a domain, like people, places, organisations and events. They are extensively used in various Artificial Intelligence 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.

Although building and using knowledge graphs is important, they give rise to quality concerns which may be a limitation to their usage. Independently of the (kinds of) source(s) from which a knowledge graph is created, data extracted for the initial knowledge graph will usually be incomplete, and will contain duplicate, contradictory or even incorrect statements – especially when taken from multiple sources. Assessing the quality of the resulting knowledge graphs is a crucial step. By quality, we here refer to fitness for purpose. Quality assessment then helps to ascertain for which purposes a knowledge graph can be reliably used.

To guarantee and to ensure that one knowledge graph is of a certain quality it is necessary to assess quality at both the instance and schema level. The semantic interoperability of Knowledge Graphs is attained through the employment of linked data principles and RDF. In most use cases, the producers of data in knowledge graphs employ an implicit schema for the different kinds of data that they are representing. There have been several proposals for RDF description and validation languages which enable the explicit definition of those schemas. In 2014, Shape Expressions (SHEx) was proposed as a concise language for this purpose [3]. In that year W3C launched the Data Shapes working group which would produce SHACL [2], which was adopted as a W3C recommendation in 2017. Those languages define a shapes graph or schema which can be used to check which nodes in an RDF graph conform to it.

There are several approaches to describing and validating knowledge graph data as well as for checking data constraints, which have been proposed and that create opportunities for new practical applications. Apart from that,

*Corresponding author. E-mail: labra@uniovi.es.

there are approaches that can leverage explicit knowledge graph schemas for new tasks, like curation, data analysis, entity linking or knowledge graph summarization.

This special issue of the Semantic Web Journal has been addressed to the members of the community interested in new methods and techniques for assessing and validating the quality of knowledge graphs as well as applications related with those approaches.

Overall, we received 10 submissions of which the following 4 papers were accepted:

- In *Using the W3C Generating RDF from Tabular Data on the Web Recommendation to manage small Wikidata datasets*, the authors deal with the problem of editing or entering data into Wikidata [4]. They propose a simple system that is used for transforming tabular snapshots stored in CSV into RDF thus generating Wikidata subgraphs which will ensure human readability and track changes over time. The use of tabular data as an input to software has become a topic of research in both academia and industry. It is particularly interesting in the case of documenting and driving data contributions to Wikidata, considering the contributors from different domains, who may not have expertise in linked data or formats like RDF. In terms of data quality the contribution of this work covers the problem of statements' timeliness and validity.
- *An Assertion and Alignment Correction Framework for Large Scale Knowledge Bases*. This work studies the problem of correcting assertions and alignments, which limit the usefulness and usability of knowledge bases due to quality issues. The authors present a general correction framework which corrects assertions whose objects are either erroneous entities or literals in an individual knowledge base and erroneous alignments in a heterogeneous knowledge base as a result of aligning multiple knowledge bases. The framework exploits related entity estimation, assertion prediction (using both semantic embeddings and observed features), and constraint-based validation.
- In *Learning SHACL shapes from Knowledge graphs*, the authors tackle the problem of how to obtain shapes from existing knowledge graphs. The paper presents an approach called SHACLearner that extracts rules from a Knowledge Graph in the form of paths (sequences of properties) which they call Inverse Open Paths and are later converted to SHACL shapes. The approach is based on embedding-based open path rule learning. The validity of those rules is established by some quality metrics: support, head coverage, and standard confidence. The paper also establishes a benchmark which may be useful for future research in this domain.
- In *Instance Level Analysis on Linked Open Data Connectivity for Cultural Heritage Entity Linking and Data Integration*, reviewing eleven widely-used resources, Go Sugimoto provides an assessment of the links provided by each resource for a sample of one hundred entities. The purpose is to use the sample entities to estimate whether researchers working in the domain of cultural heritage are able to discover additional information from these providers of linked open data (LOD). The author argues that reducing the manual analysis required before reusing existing LOD would benefit practitioners seeking to build upon these datasets in their research. Researchers in the domain of cultural heritage may benefit from consulting the link traversal data presented in this paper while planning their future research projects involving LOD in order to identify the resources most likely to contain relevant data.

References

[1] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G.D. Melo, C. Gutierrez, S. Kirrane, J.E.L. Gayo, R. Navigli, S. Neumaier, A.-C.N. Ngomo, A. Polleres, S.M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab and A. Zimmermann, Knowledge graphs, *ACM Computing Surveys* **54**(4) (2022), 1–37. doi:10.1145/3447772.

[2] H. Knublauch and D. Kontokostas, Shapes Constraint Language (SHACL), W3C Recommendation 20 July 2017, W3C Recommendation, World Wide Web Consortium, 2017, https://www.w3.org/TR/2017/REC-shacl-20170720/.

[3] E. Prud'hommeaux, J.E. Labra Gayo and H. Solbrig, Shape expressions: An RDf validation and transformation language, in: *Proceedings of the 10th International Conference on Semantic Systems – SEM'14*, ACM Press, 2014. doi:10.1145/2660517.2660523.

[4] D. Vrandečić and M. Krötzsch, Wikidata: A free collaborative knowledge base, *Communications of the ACM* **57**(10) (2014), 78–85. doi:10.1145/2629489.