

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: first@somewhere.com

^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; Krzysztof 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 [2]. In that year W3C launched the Data Shapes working group which would produce SHACL [3], 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: first@somewhere.com.

1 there are approaches that can leverage explicit knowledge graph schemas for new tasks, like curation, data analysis, 1
2 entity linking or knowledge graph summarization. 2

3 This special issue of the Semantic Web Journal has been addressed to the members of the community interested 3
4 in new methods and techniques for assessing and validating the quality of knowledge graphs as well as applications 4
5 related with those approaches. 5

6 Overall, we received 10 submissions of which the following 4 papers were accepted: 6

- 7 – In *Using the W3C Generating RDF from Tabular Data on the Web Recommendation to manage small Wikidata* 7
8 *datasets*, the authors deal with the problem of editing or entering data into Wikidata [4]. They propose a simple 8
9 system that is used for transforming tabular snapshots stored in CSV into RDF thus generating Wikidata sub- 9
10 graphs which will ensure human readability and track changes over time. The use of tabular data as an input 10
11 to software has become a topic of research in both academia and industry. It is particularly interesting in the 11
12 case of documenting and driving data contributions to Wikidata, considering the contributors from different do- 12
13 mains, who may not have expertise in linked data or formats like RDF. In terms of data quality the contribution 13
14 of this work covers the problem of statements' timeliness and validity. 14
- 15 – *An Assertion and Alignment Correction Framework for Large Scale Knowledge Bases*. This work studies the 15
16 problem of correcting assertions and alignments, which limit the usefulness and usability of knowledge bases 16
17 due to quality issues. The authors present a general correction framework which corrects assertions whose 17
18 objects are either erroneous entities or literals in an individual knowledge base and erroneous alignments in 18
19 a heterogeneous knowledge base as a result of aligning multiple knowledge bases. The framework exploits 19
20 related entity estimation, assertion prediction (using both semantic embeddings and observed features), and 20
21 constraint-based validation. 21
- 22 – In *Learning SHACL shapes from Knowledge graphs*, the authors tackle the problem of how to obtain shapes 22
23 from existing knowledge graphs. The paper presents an approach called SHACLearner that extracts rules from 23
24 a Knowledge Graph in the form of paths (sequences of properties) which they call Inverse Open Paths and are 24
25 later converted to SHACL shapes. The approach is based on embedding-based open path rule learning. The 25
26 validity of those rules is established by some quality metrics: support, head coverage, and standard confidence. 26
27 The paper also establishes a benchmark which may be useful for future research in this domain. 27
- 28 – In *Instance Level Analysis on Linked Open Data Connectivity for Cultural Heritage Entity Linking and Data* 28
29 *Integration*, reviewing eleven widely-used resources, Go Sugimoto provides an assessment of the links pro- 29
30 vided by each resource for a sample of one hundred entities. The purpose is to use the sample entities to esti- 30
31 mate whether researchers working in the domain of cultural heritage are able to discover additional information 31
32 from these providers of linked open data (LOD). The author argues that reducing the manual analysis required 32
33 before reusing existing LOD would benefit practitioners seeking to build upon these datasets in their research. 33
34 Researchers in the domain of cultural heritage may benefit from consulting the link traversal data presented in 34
35 this paper while planning their future research projects involving LOD in order to identify the resources most 35
36 likely to contain relevant data. 36

37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 References

- 41 [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.- 41
42 C.N. Ngomo, A. Polleres, S.M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab and A. Zimmermann, Knowledge Graphs, *ACM* 42
43 *Computing Surveys* **54**(4) (2022), 1–37. doi:10.1145/3447772. 43
- 44 [2] E. Prud'hommeaux, J.E. Labra Gayo and H. Solbrig, Shape expressions: an RDF validation and Transformation Language, in: *Proceedings* 44
45 *of the 10th International Conference on Semantic Systems - SEM '14*, ACM Press, 2014. doi:10.1145/2660517.2660523. 45
- 46 [3] H. Knublauch and D. Kontokostas, Shapes Constraint Language (SHACL), W3C Recommendation 20 July 2017, W3C Recommendation, 46
47 World Wide Web Consortium, 2017. <https://www.w3.org/TR/2017/REC-shacl-20170720/>. 47
- 48 [4] D. Vrandečić and M. Krötzsch, Wikidata: A Free Collaborative Knowledge base, *Communications of the ACM* **57**(10) (2014), 78–85. 48