A study about the use of OWL 2 semantics in RDF-based knowledge graphs

Pierre-Henri Paris[1], Fayçal Hamdi[1], and Samira Si-said Cherfi[1]

Conservatoire National des Arts et Métiers, Paris, France
pierre-henri.paris@upmc.fr faycal.hamdi@cnam.fr samira.cherfi@cnam.fr

Abstract. RDF-based knowledge graphs have been attracting increasing attention since Google popularized the term in 2012. However, historically, knowledge graphs are based on Semantic Web technologies. Many years ago, several works pointed out the lack of semantics in some RDF graph. So the question is whether semantics is there somewhere. Hence, we conducted an up-to-date large-scale study of the current state of the Web of data regarding the OWL 2 semantics to confirm or deny older results. Moreover, we propose an ontology to capture which OWL 2 features are defined or used in a given RDF-based knowledge graph and the tools to instantiate such an ontology.

Keywords: knowledge graph, statistics, semantics, OWL, ontology

1 Introduction

One of the key points using RDF-based knowledge graphs (KGs) is the possibility to reason on data thanks to OWL 2 and description logic. For example, users can check the consistency of the KG or infer new data. Furthermore, many tools rely on semantics to perform at their best for a given task. However, when dealing with a KG, human or automated agents might deal with the lack of necessary OWL 2 features.

A decade ago, several works focused on the study of OWL semantics in KGs and found that data was often devoid of semantics. Hence, in this paper, we propose a large-scale study of the current state of the Web of data from the OWL 2 semantics perspective. Moreover, we built an ontology to express, for a given KG, which OWL 2 and RDFS features (e.g., functional properties or subclasses) are used and in what proportions. This ontology allows the necessary information to be brought directly to the data consumer to select the appropriate tool for the realization of his or her task. Besides, we provide applications to instantiate the ontology for a given KG thanks to its SPARQL endpoint. The objective is to enable data consumers to know precisely how and to what extent OWL 2 and RDFS are used in the KG.

2 Related work

In this section, we present some works that focus on the study of the use of semantics in knowledge graphs of Linked Open Data. In [3], the authors analyzed 25500 knowledge graphs in terms of expressivity. Although compelling, this study is old and deals with a tiny number of knowledge graphs. [8] denounces the lack of expressiveness of knowledge graphs, i.e., that many knowledge graphs do not use all the different features of OWL 2, far from it. In [6], the authors emphasize that some data publishers focus solely on publishing data (i.e., triples) without annotating them with shared ontologies. They conclude that, apart from the *owl:sameAs* property, the features of OWL 2 are little used. However, this study is more of an empirical finding than a systematic study. [7] covers 12.5 million triples and aims to raise the various issues facing the Semantic Web. However, the small sample size and the age of the study this study does not provide answers to our questions. Moreover, the study lacks relevant metrics on the use of semantics. In [5], the authors proposed the biggest and deepest evaluation of OWL 2 usage so far. They evaluated more than 2 billion triples and found a wide disparity in usage among the features of OWL 2. Our study covers more recent and more numerous data (more than 30 billion triples). [4] proposes to investigate the quality of some of the best-known knowledge graphs. The authors provide basic statistics on DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. Although not a large-scale study of the use of semantics, some statistics are interesting (number of triples, number of classes, number of relations, etc.), but do not sufficiently address the semantics expressed by ontologies based on OWL 2. In [1], the authors proposed a study of the modeling style in Linked Open Data. Hence, they mostly focus on the hierarchies of the classes.

None of the cited works proposes a complete study on the use of OWL 2 semantics in RDF knowledge graphs with precise figures and at such a scale.

3 Current state of the LOD

In this paper, we only present some OWL 2 results, but RDFS and the rest of OWL 2 results are also available on our GitHub repository[1].

3.1 Data sources

We chose *LOD Laundromat*[2] that gives access to about 650 thousand KGs in HDT format. Some of these graphs refer to different versions of the same dataset, e.g., DBpedia-en, DBpedia-fr, or DBpedia 3.8. Because our demonstrator works only with SPARQL endpoints, we used Jena Fuseki[2] to query those HDT files.

Thanks to *LOD Laundromat*, 647,858 KGs have been analyzed (an HDT file represents a graph). We consider an RDF KG as a serialization of a graph expressed using the RDF graph model, i.e., composed of subject-predicate-value triples. It contains data (A-Box) and ontology (T-Box).

[1] `https://github.com/PHParis/sem_web_stats`

[2] `http://www.rdfhdt.org/manual-of-hdt-integration-with-jena/`

3.2 Results

The first view of these results is presented in Figure 1. Each of the three box plots describes a subset of knowledge graphs with their number of subjects, i.e., the graphs have been ranked by their number of subjects. It is the easiest way to expose the global shapes of KGs through their quartiles. The first box plot describes all 650K knowledge graphs. As we can see, there are a large number of very small graphs. The vast majority of the KGs contains barely 1000 subjects. However, several KGs are above the millions of subjects. Only a very small portion of the KGs (1.53%, ∼10K KGs) uses at least one OWL 2 feature. This is really astonishing, since we were expecting a small portion, according to the previous studies, but not that small. The statistics of this small portion, i.e., KGs with semantics, can be read on the second box plot. As we can see, KGs with semantics are a little bit larger in terms of the number of subjects. Finally, the last box plot represents the 100 largest KGs in terms of the number of triples. Large KGs have almost all more than 1M of subjects. Surprisingly, only 34% of top 100 KGs use at least one OWL 2 feature. It is largely more than when considering all KGs, but it is still a very low percentage if we consider they are composed of millions of triples and subjects.

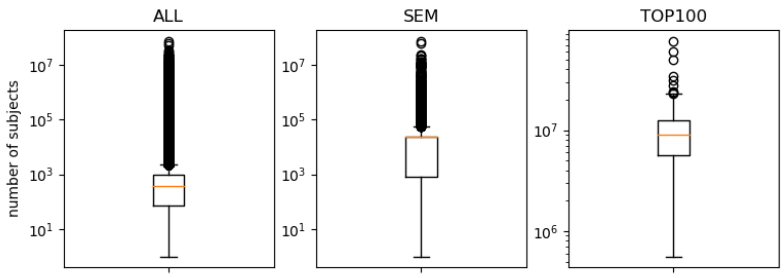

Fig. 1. Box plots of the number of subjects by selectors. ALL = all KGs, SEM = KGs with at least one OWL 2 feature, TOP100 = top 100 KGs w.r.t. their number of triples.

Table 1 concerns the types of properties (for example a property that would be defined as functional). The second column shows the number of graphs using a property of the considered type, and the third column their weighted average regarding the number of triples. The last two columns are similar, but for subjects and predicates. For example, inverse functional properties are found in 310 graphs. Among these 310 graphs, we can expect to find an average of 2.54 definitions of such properties that are used in 22.7 triples with 20.6 different subjects. As we can see, some predicates are used very little, such as the *owl:ReflexiveProperty* which is only used in 16 graphs. In these 16 graphs, very few reflexive properties are defined (1.28) and used.

Table 1. Analysis by type of property.

Type	# of graphs	weighted mean of triples	weighted mean of subjects	weighted mean of predicates
FunctionalProperty	434	9	5.76	3.06
InverseFunctionalProperty	310	22.7	20.6	2.54
TransitiveProperty	396	2.84	2.63	2.4
SymmetricProperty	320	7	4.77	2.87
AsymmetricProperty	15	4.7	4.66	4.66
IrreflexiveProperty	21	1.66	1.65	1.65
ReflexiveProperty	16	1.32	1.32	1.32

Because of space limitation, we present many other results on our GitHub repository, e.g., an analyze by topics of graphs (life science, cross domain, etc.), or class restrictions and domain/range axiom statistics.

3.3 Discussion

The main objective of our study is to verify the old results on more recent and more important data. Our observations do not defer from those of previous work. Indeed, despite being a W3C recommendation since 2009, many OWL 2 features have not been adopted by ontologist or data publishers. The state of Linked Data is the same as it was for the last large study in 2012 [5]. The most surprising results are the very low number of KGs using semantics. Even when considering the largest KGs, a great number of them still do not use OWL 2 features. Moreover, there is a great disparity between the usage of the different features. While several are heavily used, most features are barely present in studied KGs. There is a need to understand why such inertia. Is OWL 2 too powerful regarding the needs of data modelers? Or too hard to be used? Even if more complex OWL 2 features were used in KGs, will users need them? Maybe a specification like SHACL [9] will encounter a greater success and could be considered as a viable alternative or a complement in some cases.

4 Ontology

We propose an ontology[3] to explicit the use of classes and properties defined with OWL 2 and RDFS features in a KG. For instance, an objective for a user could be to know the number of properties that are transitive and their number of uses in the graph. We extended the VoID[4] vocabulary with properties to explicit *(i)* how many properties and classes are defined with a given OWL 2 feature, or *(ii)* the number of use of a given OWL 2 feature. The OWL 2 features

[3] `http://cedric.cnam.fr/isid/ontologies/OntoSemStats.owl`

[4] `https://www.w3.org/TR/void/`

are organized depending on their utility, e.g., *owl:sameAs* and *owl:differentFrom* have the same superclass because they both are related to identity.

To instantiate the ontology for a given SPARQL endpoint, we propose OntoSemStatsWeb[5], an open-source software (under the GPL open-source license) written in C# (using dotnetRDF[6]). The application has three different forms: *(i)* a Web page that is our live demonstrator, *(ii)* a Web API to operate seamlessly with an automated agent, and *(iii)* a command-line application. All the tools that we developed are available as Docker images (one for the command-line application and one for the Web application and the Web API), in order to promote ease of use and adoption.

5 Conclusion

In this paper, we conducted a large-scale study that provides an up-to-date overview of the semantic usages in the LOD. This study confirmed older papers results: only a small portion of KGs uses OWL 2 semantics, and those KGs use only some features of OWL 2 heavily. Moreover, we proposed an ontology to capture the present semantics in a KG. The ontology *(i)* facilitate knowledge discovery for users and *(ii)* may increase the visibility of data publishers' KG.

References

1. Asprino, L., Beek, W., Ciancarini, P., van Harmelen, F., Presutti, V.: Observing LOD using equivalent set graphs: It is mostly flat and sparsely linked. In: ISWC (1). Lecture Notes in Computer Science, vol. 11778, pp. 57–74. Springer (2019)
2. Beek, W., Rietveld, L., Bazoobandi, H.R., Wielemaker, J., Schlobach, S.: LOD laundromat: A uniform way of publishing other people's dirty data. In: International Semantic Web Conference (1). Lecture Notes in Computer Science, vol. 8796, pp. 213–228. Springer (2014)
3. d'Aquin, M., Baldassarre, C., Gridinoc, L., Angeletou, S., Sabou, M., Motta, E.: Characterizing knowledge on the semantic web with watson (2007)
4. Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked Data quality of DBpedia, Freebase, OpenCyc, Wikidata, and Yago. Semantic Web (Preprint), 1–53 (2016)
5. Glimm, B., Hogan, A., Krötzsch, M., Polleres, A.: OWL: yet to arrive on the web of data? In: LDOW. CEUR Workshop Proceedings, vol. 937. CEUR-WS.org (2012)
6. Hitzler, P., van Harmelen, F.: A reasonable semantic web. Semantic Web **1**, 39–44 (2010)
7. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the pedantic web. In: LDOW. CEUR Workshop Proceedings, vol. 628. CEUR-WS.org (2010)
8. Jain, P., Hitzler, P., Yeh, P.Z., Verma, K., Sheth, A.P.: Linked data is merely more data. In: AAAI Spring Symposium: Linked Data Meets Artificial Intelligence (2010)
9. Knublauch, H., Kontokostas, D.: Shapes constraint language (shacl). W3C Candidate Recommendation **11**(8) (2017)

[5] `https://github.com/PHParis/OntoSemStatsWeb`
[6] `https://github.com/dotnetrdf/dotnetrdf`