# OntoSemStats: an ontology to express the use of 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.** For many users or automated agents, working with knowledge graphs may be a complicated task. Indeed, multiple tools using knowledge graphs rely on semantics to perform at their best. For example, in the context of data integration, some instance matching tools use semantic features such as functional and inverse functional properties or disjoint classes to discover instances that are the same (or not). Hence, in many cases, conducting an exploratory study is required to discover which semantic features are used or defined in a knowledge graph. In this paper, we propose an ontology and a large-scale ontology-based Web service that provides statistics about the use of OWL 2 and RDFS semantic features (e.g. functional properties or subclasses) in knowledge graphs. This will allow a human or automatic agent to choose the most appropriate tool or data for a given task. It also gives the data publishers a clear picture about the semantics they provide to data consumers. These statistics are represented in the form of an RDF graph (with different serialization possibilities), making them easy to use and share.

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

## 1 Introduction

For a given task, several types of approaches can be considered when it comes to using RDF-based knowledge graphs (KGs). Some approaches rely mainly on the semantics available in the graphs, others, on the contrary, make little or no use of it. Of course, in between these two extremes, approaches can take advantage of semantics, without relying entirely on it. For example, if the task is to interconnect several KGs, approaches may use a combination of techniques such as statistics, semantics, or data partitioning algorithms. Besides, approaches relying mainly on semantics can outperform other types of approaches if semantics is very present in the KG. However, if semantics is absent, the results may not be what the user expects. Therefore, it is often necessary to conduct a first exploratory study of KG to know which tool will be best suited for a given task or to choose between topic-related graphs. Such a study helps to understand what the data may have to offer. Unfortunately, this exploratory step is time-consuming, especially if the documentation accompanying KG is missing or not

very informative. Several vocabularies or ontologies have been proposed to provide the user with an overview of the data contained in the KG. For example, Dublin Core[1] [3], Creative Commons Rights Expression Language[2], Data Catalog Vocabulary[3], or VoID[4] [1] allow KGs to be described. However, they do not give the possibility to express which elements of OWL 2 or RDFS are used.

In this paper, we propose an ontology to express which OWL 2 and RDFS features (e.g. functional properties or subclasses) a KG uses and in what proportions. This ontology allows the necessary information to be brought directly to the data consumer to select, in full knowledge of the facts, the appropriate tool for the realization of her task. Besides, we provide a Web application 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 a KG uses OWL 2 and RDFS. The aggregation of statistics on all KG vocabularies or ontologies described with OWL 2 and RDFS makes it possible to achieve this goal. We also conducted a large-scale study of the current state of the Web of data regarding the usage of OWL 2. As the paper must remain succinct, a GitHub repository[5] presents the results of the study.

## 2 Ontology

The ontology we propose (available online [6]) aims to explain 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.

VoID [1] is a vocabulary that can be used to describe a KG. This description facilitates KG discovery and use. Besides, VoID offers elementary statistics such as the number of classes or triples. Our ontology extends this vocabulary by providing more detailed statistics on the use of OWL 2 and RDFS features. We represent a KG as an instance of the class *void:Dataset* that can have as many *:Stat*[7] instances as it uses OWL 2 and RDFS properties or classes. Each instance of *:Stat* has one and only one *:SemanticFeature* instance. The *:hasSemanticFeature* property (see Listing 1.1) allows an instance of *:Stat* to be linked to its *:SemanticFeature*. The different types of range of *:hasSemanticFeature* are disjointed two by two, thus making it possible to detect any error in the instantiation of this ontology.

```
:hasSemanticFeature rdf:type owl:ObjectProperty ,
    owl:FunctionalProperty , owl:AsymmetricProperty ,
    owl:IrreflexiveProperty ;  rdfs:domain :Stat ;
```

[1] http://www.dublincore.org/specifications/dublin-core/
[2] https://creativecommons.org/ns
[3] https://www.w3.org/TR/vocab-dcat/
[4] https://www.w3.org/TR/void/
[5] https://github.com/PHParis/sem_web_stats
[6] http://cedric.cnam.fr/isid/ontologies/OntoSemStats.owl
[7] *Classes and properties represented without a prefix belong to our ontology.*

```
rdfs:range :SemanticFeature ;
rdfs:comment "Specify which OWL 2 or RDFS semantic
feature is the target of the given stat."@en ;
rdfs:label "has semantic feature"@en .
```

**Listing 1.1.** Definition of the *hasSemanticFeature* property.

For each feature of OWL 2 and RDFS, we created its own interpretation for two reasons. First, if one has an OWL 1 KG and wants to integrate the stats, then to keep the OWL profile unchanged, we must represent the semantic features with our own IRI. For example, the triple ⟨ *:stat :hasSemanticFeature owl:FunctionalProperty* ⟩ would lead to OWL 1 Full and undecidability problems[8,9] since *owl:FunctionalProperty* is a class. Therefore, for every OWL 2 and RDFS feature, we created a subclass of *:SemanticFeature*. For example, *:OwlFunctionalProperty* represents the statistics of the functional properties. Second and more important, the different axioms of OWL 2 and RDFS can impact properties, classes, or instances. For this, we have chosen to ensure that the design of our ontology reflects these possibilities to keep a clean ontology. Depending on its purpose, an axiom will be "put" in a particular class. For example, Listing 1.2 shows the definition of *:PropertyType* (subclass of *:SemanticFeature*) used to represent the different types that a property can have (symmetrical, reflexive, etc.). Another example is the "PropertyRelation" class, which gathers, among others, statistics concerning *owl:propertyChainAxiom* or *owl:inverseOf*, which are axioms allowing the description of the nature of the relation between properties.

```
:PropertyType rdf:type owl:Class ; rdfs:subClassOf
    :PropertyAxiom ;
    owl:disjointUnionOf ( :OwlAsymmetricProperty
    :OwlFunctionalProperty :OwlInverseFunctionalProperty
    :OwlIrreflexiveProperty :OwlReflexiveProperty
    :OwlSymmetricProperty :OwlTransitiveProperty) .
```

**Listing 1.2.** Definition of the *Properties* class which represents the different types used to define a property.

To provide statistics for each feature of OWL 2, we have created two properties: *:definitionCount* and *:usageCount*. The first one is to state how many times the axiom is used in a definition (e.g. the number of functional properties) and the second one how many times the definitions using the axiom are used (e.g. how many triples use a functional property). Listing 1.3 shows the definition of the *:usageCount* property which allows us to declare, for example, that 3000 triples use a functional property.

```
:usageCount rdf:type owl:DatatypeProperty ,
  owl:FunctionalProperty ; rdfs:domain :Stat ;
  rdfs:range xsd:integer ;
  rdfs:comment "Number of usage of a semantic
```

[8] https://www.w3.org/2007/OWL/wiki/Profile_Explanations
[9] https://www.w3.org/TR/owl2-profiles/

```
feature."@en ; rdfs:label "usage count"@en .
```

**Listing 1.3.** Definition of the property allowing to specify how many times a feature is used.

## 3 Web application

Our application, OntoSemStatsWeb[10], is an open-source software (under the GPL open-source license) written in C# (using dotnetRDF[11]) and JavaScript (using Comunica[12] [2]). The application has three different forms: *(i)* a Web page that is our live demonstrator[13], *(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), to promote ease of use and adoption.

Depending on the used tool (i.e. Web page, API, or command-line), the graph is presented in various fashions. The Web page summarizes the results through a user-friendly table and a visual representation and provides a link to download the graph. On the other side, the Web API and the command-line applications allow the graph serialization in RDF/XML, Turtle, N-Triples, Notation3, and JSON-LD.

## 4 Conclusion

In this paper, we proposed an ontology that described the OWL 2 and RDFS features defined and used in a given KG. Moreover, we provided tools that automatically instantiate this ontology for a given SPARQL endpoint. A human agent can use these tools through a web page and command-line or an automated agent through Web API. By offering easy access to the statistics about semantic usages, we help data consumers in choosing the right tool or KG that best suited his or her objectives. Easy access may increase KG consumption and improve user experience. Finally, to show the usefulness of our application, we conducted a large-scale study that provides an up-to-date overview of the semantic usages in the LOD. In the future, we plan to add native support for HDT files.

## References

1. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets. In: Linked Data on the Web. CEUR Workshop Proceedings, vol. 538. CEUR-WS.org (2009)
2. Taelman, R., Herwegen, J.V., Sande, M.V., Verborgh, R.: Comunica: A modular SPARQL query engine for the web. In: International Semantic Web Conference (2). Lecture Notes in Computer Science, vol. 11137, pp. 239–255. Springer (2018)
3. Weibel, S., Kunze, J.A., Lagoze, C., Wolf, M.: Dublin core metadata for resource discovery. RFC **2413**, 1–8 (1998). https://doi.org/10.17487/RFC2413, `https://doi.org/10.17487/RFC2413`

[10] `https://github.com/PHParis/OntoSemStatsWeb`
[11] `https://github.com/dotnetrdf/dotnetrdf`
[12] `https://comunica.linkeddatafragments.org/`
[13] `https://ontosemstats.herokuapp.com/`