Loupe - An RDF Dataset Description Model for Expressing Vocabulary Usage Patterns

Nandana Mihindukulasooriya, María Poveda-Villalón, Raúl García-Castro, and Asunción Gómez-Pérez

Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
{nmihindu,mpoveda,rgarcia,asun}@fi.upm.es

Abstract. This position paper discusses the need for extending dataset descriptions, such as DCAT, in the case of RDF data to include comprehensive vocabulary usage and triple pattern information (for instance, as a DCAT profile for vocabulary usage and triple patterns in RDF data). As the basis of the discussion, the paper presents four use cases whose requirements can not be easily fulfilled by the current RDF dataset descriptions. To this end, we propose an extended RDF dataset description vocabulary, the Loupe model, which aims to capture an extensive set of vocabulary usage statistics and triple pattern information to satisfy such use cases.

Keywords: RDF, Vocabulary, Descriptions, Profiles, Statistics

1 Introduction

Dataset descriptions play an important role in providing useful metadata about datasets allowing data consumers to search and discover relevant datasets for a given use case. The ability find and understand suitable data is vital for fostering data resuse. On the one hand, DCAT[1] is one of the most commonly used vocabulary for dataset descriptions. It provides a wide range of useful information about features of a dataset such as abstract concepts, keywords, distributions, and access methods. However, DCAT is generic by design, *i.e.*, it is not specific to RDF datasets but suitable for any kind of dataset.

On the other hand, dataset descriptions that focus on RDF data, such as VoID[2], Dataset Description Vocabulary[3], and Aether VoID extension[4] provide information more specific to the RDF model. Moreover, there are other vocabularies that focus on specific aspects. Vocabularies focused on dataset statistics, such as LODStats[1] and RDFStats[2], provide several statistical metrics about RDF data. Dataset profile models, such as ProLOD++ or ExpLOD, provide a set of dataset characteristics that describes a given dataset in the best possible way and separates it maximally from other datasets.

[1] https://www.w3.org/TR/vocab-dcat/

[2] https://www.w3.org/TR/void/

[3] https://www.w3.org/2001/sw/hcls/notes/hcls-dataset/

[4] http://ldf.fi/void-ext

Nevertheless, the current RDF dataset descriptions fail to fulfill the requirements of some use cases that depend on extensive knowledge of the vocabularies used and the frequent patterns in datasets. In this paper, we discuss several motivating use cases that can not be easily performed using the current dataset descriptions and propose an extended RDF dataset description model to fulfill the requirements of such use cases.

The rest of the paper is organized as follows: Section 2 discusses four motivating use cases; Section 3 presents the proposed dataset description model; Section 3 describes tooling support for the model; and, Section 4 draws some conclusions.

2 Motivating Use Cases

– **Dataset discovery for a specific task** One preliminary task most data consumers have to perform is to find an appropriate dataset for a given task. For smaller tasks, such as training a classifier using machine learning exploiting data from the LOD cloud as training data, the data requirements can be expressed using SPARQL queries or SHACL[5] shapes. However, currently it is not easy to automatically discover the datasets in the LOD cloud that would contain data matching such query. At the moment, a lot of manual effort is required for finding a suitable dataset.
– **Understanding dataset content** Currently when a new RDF dataset is given to a consumer, though some information can be available such as SKOS themes in DCAT or vocabularies used in VoID, they contain enough metadata to the level that it enables her to write effective queries against the dataset or use it in an application. Typically a set of exploratory queries are needed to be performed to understand how the vocabularies are used, the structure of the data (shapes), and other characteristics. This requires consumers to spend a considerable time on such tasks before using the data.
– **Comparison of multiple dataset versions** When datasets evolve generating multiple versions over time, it is interesting for consumers to know how the dataset has evolved (*i.e.*, what has been added to the dataset or what has been removed). The current dataset descriptions allow to get some idea about the evolution using metrics such as the triple count, however a comprehensive analysis of which type and amount of content has been added, removed, and modified is not possible without doing a lot of manual inspection.
– **Vocabulary usage reports** How vocabularies are used by data producers is a useful information for ontology engineers and people who are involved in vocabulary management. An implementation report of an ontology, such as the PROV-O[6] one, is an example for this. Currently, it is not possible to generate them automatically and they require a lot of manual effort to collect the necessary information and process them.

[5] https://www.w3.org/TR/shacl/

[6] https://www.w3.org/TR/prov-implementations/

3 The Loupe Model

The Loupe model[3] is an extended dataset description model that is focused on the vocabularies used and triple patterns in the dataset. It extends information in the existing RDF dataset descriptions models such as VoID and LODStats and proposes metrics that capture further details on vocabulary usage and triple patterns (*e.g.*, which classes and properties are used in the dataset, and implicit domains, ranges, cardinalities of properties as seen in data), classes with common instances, frequent abstract triple patterns, or common RDF shapes (frequent subgraphs). The core of the Loupe model is illustrated in Figure 1 and it is described in the ontology documentation[7].

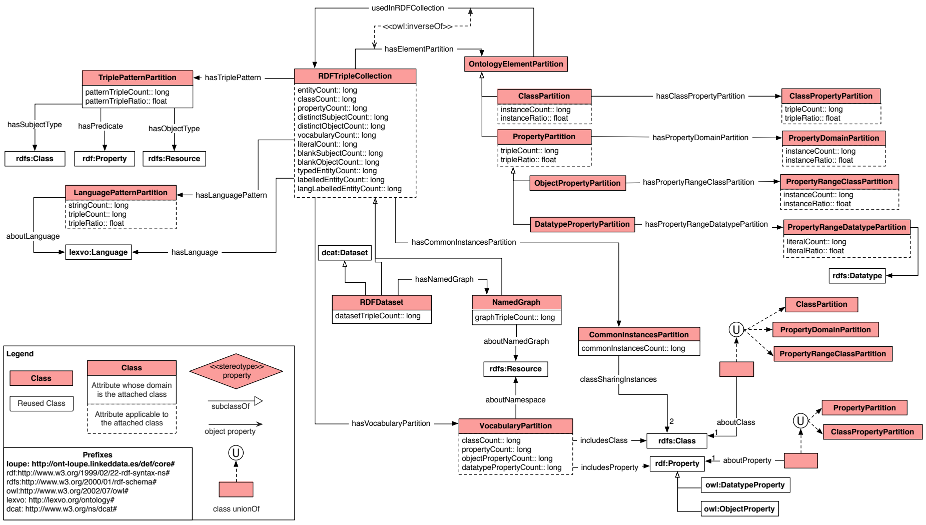

Fig. 1. The Loupe Model.

4 Implementation

The Loupe tools (Fig. 2) provide means for automatically extracting the relevant information using a set of parameterized SPARQL query templates and generating the dataset descriptions according to the Loupe model. The dataset can be either accessible as RDF data dumps or as public SPARQL endpoints. Generated descriptions can be used to build applications that cater the aforementioned use cases. In addition, Loupe web application[8] allows the exploitation of this information in a user friendly manner with visualization support.

[7] http://ont-loupe.linkeddata.es/def/docs/core/

[8] http://loupe.linkeddata.es

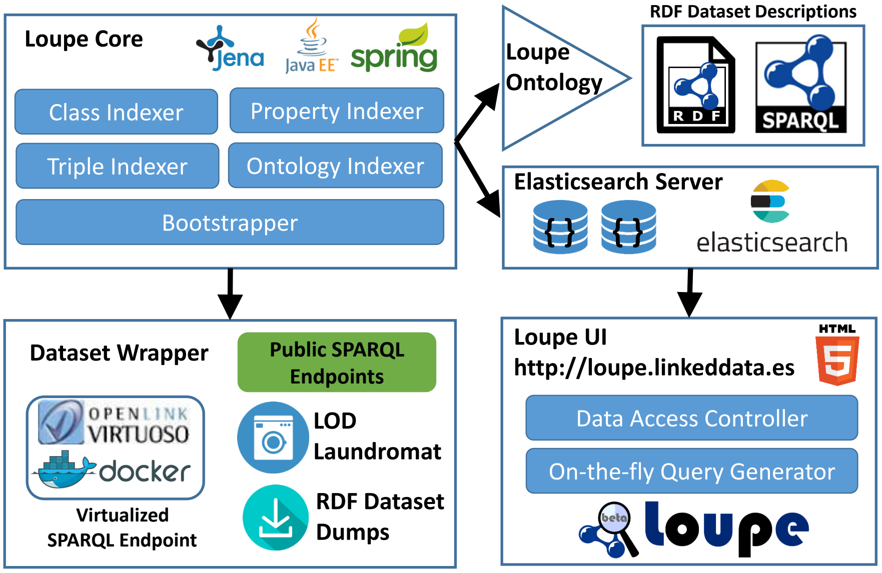

Fig. 2. Loupe tool chain.

5 Conclusions and Future Work

In this paper, we argue that the current RDF dataset descriptions are not capable of satisfying the requirements of some use cases that require in-depth information of the vocabulary usage and pattern information. To cater such use cases, we propose the Loupe model, that consists of a comprehensive set of metrics about vocabulary usage and triple patterns in an RDF dataset. In future, we plan to create these descriptions for all datasets in LOD Laundromat which includes large portion of the LOD cloud.

References

1. Demter, J., Auer, S., Martin, M., Lehmann, J.: LODStats-An Extensible Framework for High-performance Dataset Analytics. In: Proceedings of the EKAW 2012. Lecture Notes in Computer Science (LNCS) 7603, Springer (2012) 29% acceptance rate.
2. Langegger, A., Woss, W.: RDFStats An Extensible RDF Statistics Generator and Library. In: 20th International Workshop on Database and Expert Systems Application, IEEE (2009) 79–83
3. Mihindukulasooriya, N., Poveda-Villalón, M., García-Castro, R., Gómez-Pérez, A.: Loupe-An Online Tool for Inspecting Datasets in the Linked Data Cloud. In: Demo at the 14th International Semantic Web Conference, Bethlehem, USA (2015)