Vocab-express - Exploring the Vocabulary used in a Dataset

Boris Villazón-Terrazas[1] and Michael Hausenblas[2]

[1] OEG-DIA, FI, Universidad Politécnica de Madrid, Spain
bvillazon@fi.upm.es
[2] Digital Enterprise Research Centre, NUI Galway, Ireland
michael.hausenblas@deri.ie

Abstract. In the context of the Linked Open Data, it is important to promote the reuse and linkage of datasets, and to this end, it is necessary to know the structure of datasets. In order to have a clear picture of the structure, we need to explore the vocabulary used in the datasets. In this paper we present a brief description of vocab-express, a tool that aims at exploring the vocabulary used in a dataset.

Key words: Linked Data, RDF Dataset, Vocabulary

1 Introduction and Motivation

So far, Linked Data principles and practices are being adopted by an increasing number of data providers, getting as result a global data space on the Web containing hundreds of LOD datasets [1]. In this context it is important to promote the reuse and linkage of datasets, and to this end, it is necessary to know the structure of datasets. One step forward for knowing in depth the structure of a given dataset is to explore the vocabulary used in the dataset, and how the dataset is actually using such vocabulary.

There are available works such as (1) *LODStats*[3] that provides the information related to the vocabulary used in a dataset, and (2) *make-void* [4] that computes statistics about RDF files. However, LODStats is thought for the whole set of LOD datasets registered in The Data Hub [5], and it is based on declarative descriptions of those datasets; and *make-void* is thought for RDF files but not for RDF datasets.

In this paper we present vocab-express[6], a simple tool for exploring the vocabulary used in a given dataset. The tool provides all the related information of the vocabulary: (1) the list of all classes, (2) the list of all the properties, (3) the number of instances of each class, (4) the number instances of each property, (5) the language of labels and comments, of the vocabulary elements.

[3] http://stats.lod2.eu/
[4] https://github.com/cygri/make-void
[5] http://thedatahub.com
[6] http://vocab-express.nodester.com/

2 Vocab-express overview

Vocab-express is being implemented in *node.js*[7], which is a platform built on V8[8] for easily building fast, scalable network applications. Figure 2 depicts the workflow of vocab-express. Next we describe the steps of the workflow

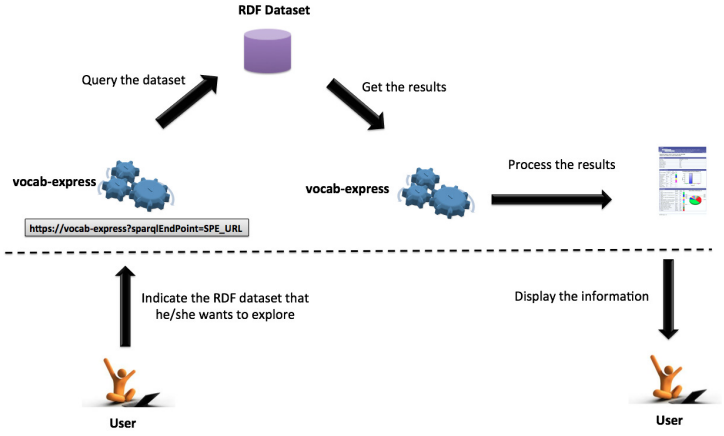

Fig. 1. vocab-express workflow

1. The user indicates the RDF dataset he/she wants to explore by providing its SPARQL endpoint URL
2. vocab-express receives and validates the SPARQL endpoint URL
3. vocab-express queries the SPARQL endpoint and gets results
4. vocab-express processes the results and displays the information to the user

The workflow is similar if the user provides the URL of the VoID file instead of the SPARQL endpoint URL.

3 Conclusions

In this paper we have presented our ongoing work for exploring the vocabulary used in an RDF dataset. The tool provides a clear picture of the vocabulary and how it is actually used in the dataset. As future work we want to evaluate the performance of the tool, comparing it against approaches not implemented in *node.js*. Moreover, we will improve how the information is displayed to the user.

References

1. T. Heath and C. Bizer. Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web Theory and Technology*, 1(1):1–136, 2011.

[7] `http://nodejs.org/`
[8] V8 is Google's open source JavaScript engine.