Datavore: A Vocabulary Recommender Tool Assisting Linked Data Modeling

Mohamed Ben Ellefi, Zohra Bellahsene, Konstantin Todorov
{firstname.lastname@lirmm.fr}

LIRMM / University of Montpellier, France

Abstract. In this paper, we introduce the vocabulary recommendation system *Datavore* (*Da*ta *vo*cabulary *re*commender). The tool is oriented towards metadata designers providing ranked lists of vocabulary terms to reuse in the web of data modeling process, together with additional metadata and cross-terms relations. *Datavore* relies on the Linked Open Vocabulary ecosystem for acquiring vocabularies and metadata. The system is able to deal with noisy and multilingual input data.

Keywords: Linked Data, Vocabulary Recommendation, Modeling Web Data.

1 Introduction and Motivation

Following the fourth principle of linked data, a large number of datasets from different domains and using different vocabularies have been published and interlinked on the web. To guide data discovery and reuse, catalogs of linked data, such as the Data Hub[1], have been created.

With the increasing use of Linked Open Data (*LOD*), it becomes more and more important for data providers not only to publish their data but also to model and describe them following the LOD best practices[2]. Here, we draw the reader's attention to the recommendation of building on, instead of replicating, existing *RDF schema* and vocabularies, in an effort to improve interoperability [1]. Thus, an important step towards the web data modeling task is the discovery of all relevant vocabularies to reuse.

In this paper, we introduce *Datavore*, a vocabulary recommender system, which uses the Linked Open Vocabularies[3] (LOV) as a vocabulary search engine. In addition to a list of ranked recommended concepts/properties, the tool provides important additional metadata, as well as cross-terms relations in the form of a set of triples combinations, handling noisy and multilingual input. To our knowledge, there is only one comparable tool *Karma* [2] – which is a semi-automatic tool that proposes a mapping between semantic types (an OWL class or the range of a data property) from the introduced ontology to the data source

[1] `http://datahub.io/`
[2] `http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook`
[3] `http://lov.okfn.org/dataset/lov/`

columns– the main contribution of *Datavore* is the fact of handling the multi-lingual data as well as the benefit of the whole LOV up-to-date vocabularies in service of the linked data modeling task.

2 Overview of Datavore

We proceed to describe the workflow of the system, shown in Figure 1.

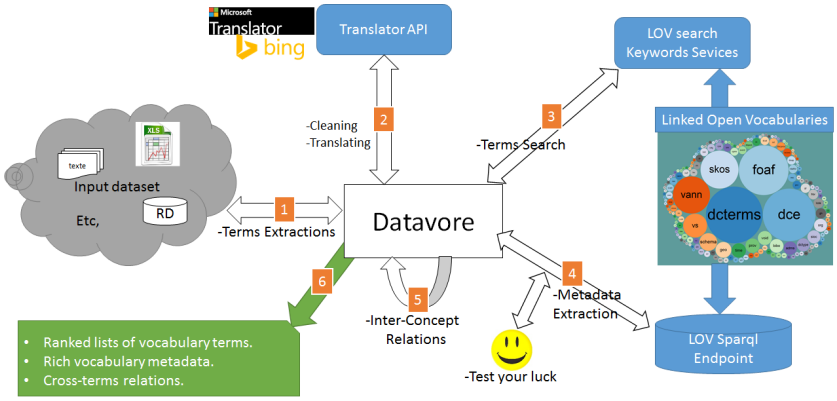

Fig. 1. Datavore Workflow

Source Terms Extraction (1). The input of *Datavore* is a list of terms extracted from the data source. In the current version, we parse a loaded CSV file as input and extract the list of column names[4]. The use of other kinds of structured or semi-structured input data is envisageable.

Cleaning and Translating (2). In most cases, the extracted string of characters needs to be cleaned-up by removing or modifying the unwanted characters. We use the *Microsoft Translator* java API[5] in order to clean up the initial string and render it in a linguistically correct form. For example, translate(CatÃ©gorie, fr, fr) returns Catégorie. In case there are no sufficient or satisfactory results by using the source language, the system uses the same service to translate the source item into English, the most common language on the LOD.

Terms Search (3). We opted for the LOV as a vocabulary search engine, which, to the best of our knowledge, is the only purpose-built vocabulary search engine available on the Web with an up-to-date index. As a design decision, *Datavore*

[4] We note that considering the actual values from the CSV misled the recommendation in most cases that is why we limited ourselves to the column names.
[5] `https://code.google.com/p/microsoft-translator-java-api/`

queries the LOV search service with the extracted cleaned or/and translated terms. The result is a list of concepts for each source term ranked by the LOV metric, which is based on the vocabulary terms popularity in the LOD datasets and in the LOV ecosystem.

Metadata Extraction (4). Metadata designers are recommended to select popular vocabularies found in the search phase but it is not straightforward to judge which vocabulary is better suited to the application. *Datavore* queries the LOV endpoint (/dump file) to extract the needed metadata to help designers to choose the appropriate vocabularies. As a result, for each concept c, extract: (i) the set of object properties having c as domain that includes labels and hierarchical relations, (ii) the set of datatype properties that can be used with c as domain, and (iii) a link to the vocabulary web site. In addition, we provide a "*test-your-luck*" option, which recommends to the user only one, the top ranked, datatype property. This "lucky" property has the highest Levenshtein string similarity [3] with the source term.

Inter-Concept Relations (5). From the extracted lists of recommended concepts, *Datavore* queries the LOV endpoint (/dump file) to retrieve cross-lists relations, i.e., relations between concepts from different lists. These metadata are crucial for selecting the best combination of predicate names to reuse for the input dataset.

3 Example Scenario

Imagine a designer who wants to model the data in Table 1 using an ontology editor. *Datavore* will guide him/her to find vocabularies to reuse, returning a sorted list of concepts for each column name. For the column "City Name" *Datavore* queries the LOV using the keyword "City" and returns the sorted list of concepts {"akt:City", "place:City", "lgdo:City", etc.}. When the designer selects the concept "akt:City", *Datavore* presents to him/her the following metadata: (i) literals (like rdfs:label, rdfs:comment, etc.) and the hierarchical relations of "akt:City", (ii) a set of datatype properties like "foaf:name" that have "akt:City" as *rdf:domain* to represent the column "City Name". After the concepts extraction, *Datavore* queries the LOV again to extract inter-columns triples and recommends to the user a set of relations between column names. In our example, the recommended relation between the two columns "Person Name" and "PostalAdress" is the triple: $< foaf : Person >< akt : hasAddress >< akt : PostalAddress >$.

4 Technical Notes

For a proper use of *Datavore*, we take note of the following.

— *Datavore* is meant to be used in complementarity with ontology development tools for dataset modeling.

id	Person Name	Profession	Lab	City Name	Postal Adress	Country
1	M. Ben Ellefi	PhD Student	LIRMM	Montpellier	34090	France
2	K. Todorov	Assoc. Pr.	LIRMM	Montpellier	34000	France
3	Z. Bellahsene	Pr.	LIRMM	Montpellier	34000	France

Table 1. LIRMM Open Data Team Example.

— Modeling Linked Data generally requires an ontology engineer and a domain expert [4].

— *Datavore* uses the LOV search service, for which, as for any search engine, the choice of input keywords is crucial. For that reason, *Datavore* enables the user to update the source terms from the interface.

— The complexity of the *inter-concept relations* step is of order $O(N^2 M_1 M_2)$, where, N is the cardinality of the source terms, M_1 and M_2 are respectively the cardinalities of the two compared lists of recommended concepts.

— *Datavore* has been implemented in Java and it is available as a *GUI* desktop application[6] together with a demonstration video[7].

5 Conclusion and Future Work

We introduced *Datavore*, a vocabulary recommender system based on LOV assisting linked data modeling. Among the original features of the tool is the fact that it provides metadata of the recommended predicates and that it automatically retrieves existent relations between the predicates to guide metadata design. LOV is a trustworthy search engine but the actual version contains less than 480 vocabularies. In the future, we intend to extend *Datavore* via the Swoogle search engine, which contains over $10,000$ ontologies.

Acknowledgements. *This research is funded under the Datalyse project*[8].

References

1. B. Hyland, B. Terrazas, and S. Capadisli, "Cookbook for open government linked data," *W3C, W3C Task Force-Government Linked Data Group*, 2011.
2. C. A. Knoblock, P. Szekely, J. L. Ambite, A. Goel, S. Gupta, K. Lerman, M. Muslea, M. Taheriyan, and P. Mallick, "Semi-automatically mapping structured sources into the semantic web," in *ESWC*, pp. 375–390, Springer, 2012.
3. V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, pp. 707–710, 1966.
4. J. Schaible, T. Gottron, S. Scheglmann, and A. Scherp, "Lover: support for modeling data using linked open vocabularies," in *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, pp. 89–92, ACM, 2013.

[6] Download and unzip this file: `http://www.lirmm.fr/benellefi/Datavore_ExeFile`
[7] `http://www.lirmm.fr/benellefi/Datavore_VideoDemo`
[8] `http://www.datalyse.fr/` - FSN-AAP Big Data $n3$.