Exploring type-specific topic profiles of datasets: a demo for educational linked data

Davide Taibi[1], Stefan Dietze[2], Besnik Fetahu[2], Giovanni Fulantelli[1]

[1]Istituto per le Tecnologie Didattiche, Consiglio Nazionale delle Ricerche, Palermo, Italy
{davide.taibi, giovanni.fulantelli}@itd.cnr.it

[2]L3S Research Center, Hannover, Germany
{dietze, fetahu}@l3s.de

Abstract. This demo presents the dataset profile explorer which provides a resource type-specific view on categories associated with available datasets in the Linked Data cloud, in particular the ones of educational relevance. Our work utilises type mappings with dataset topic profiles to provide a type-specific view on datasets and their categorisation with respect to topics, i.e. DBpedia categories. Categories associated with each dataset are shown in an interactive graph, generated for the specific profiles only, allowing for more representative and meaningful classification and exploration of datasets.

Keywords: Dataset profile, Linked Data for Education, Linked Data Explorer

1 Motivation

The diversity of datasets in the Linked Data (LD) cloud has increased in the last few years, and identifying a dataset containing resources related to a specific topic is, at present, a challenging activity. Moreover, the lack of up-to-date and precise descriptive information has exacerbated this challenge. The mere keywords-based classification derived from the description of the dataset owner is not sufficient, and for this reason, it is necessary to find new methods that exploit the characteristics of the resources within the datasets to provide useful hints about topics covered by datasets and their subsequent classification.

In this direction, authors in [1] proposed an approach to create structured metadata to describe a dataset by means of topics, where a weighted graph of topics constitutes a dataset profile. Profiles are created by means of a processing pipeline[1] that combines techniques for datasets resource sampling, topic extraction and topic ranking. Topics have been associated to dataset by using named entity recognition (NER) techniques and a score, representing the relevance of a topic for a dataset, has been calculated using algorithms to evaluate node relevance in network such as PageRank, K-Step Markov, and HITS.

[1] http://data-observatory.org/lod-profiles/profiling.htm

The limitations of such an approach are related mainly to the following aspects. First, the meaning of individual topics assigned to a dataset can be extremely dependent on the type of resources they are attached to. Also, the entire topic profile of a dataset is hard to interpret if categories from different types are considered at the same time. As an example of the first issue, the same category (e.g. "Technology") might be associated to resources of very different types such as "video" (e.g. in the Yovisto Datset[2]) or "research institution"(e.g. in the CNR dataset[3]). Concerning the second issue, the single topic profile attached for instance to bibliographic datasets (such as: the LAK dataset[4] or Semantic Web Dog Food[5]) - in which people ("authors"), organisations ("affiliations") and documents ("papers") are represented – is characterized by the diversity of its categories (e.g. DBpedia categories: Scientific_disciplines, Data_management Information_science but also Universities_by_country, Universities_and_colleges). Indeed, classification of datasets in the LD Cloud is highly specific to the resource types one is looking at. While one might be interested in the classification of "persons" listed in one dataset (for instance, to learn more about the origin countries of authors in DBLP), another one might be interested in the classification of topics covered by the documents (for instance disciplines of scientific publications) in the very same dataset.

The approach we propose in this demo to overcome the limitations described above relies on filtering the topic profiles defined in [1] according to the types of the resources. This results in a type-specific categorisation of datasets, which considers both the categories associated with one dataset and the resource types these are associated with.

However, the schemas adopted by the datasets of the LD cloud are heterogeneous, thus making difficult to compare the topic profiles across datasets. While there are many overlapping type definitions representing the same or similar real world entities, such as "documents", "people", "organization", type-specific profiling relies on type mappings to improve the comparability and interpretation of types and consequently, profiles. For this aim the explicit mappings and relations declared within specific schemas (as an example *foaf:Agent* has as subclasses: *foaf:Group*, *foaf:Person*, *foaf:Organization*) as well as across schemas (for instance through *owl:equivalentClass* or *rdfs:subClassOf* properties) are crucial.

While relying on explicit type mappings we have based our demo on a set of datasets where explicit schema mappings are available from earlier work [2]. This includes education-related datasets identified by the LinkedUp Catalog[6] in combination with the dataset profiles generated by the Linked Data Observatory[7]. While the latter provides topic profiles for all selected datasets, the LinkedUp Catalog contains explicit schema mappings which were manually created for the most frequent types in the dataset. Specifically, the profile explorer proposed in this demo aims at providing a resource type-specific view on categories associated with the datasets in the LinkedUp Catalog. In

[2] http://www.yovisto.com/
[3] http://data.cnr.it/
[4] http://lak.linkededucation.org
[5] http://data.semanticweb.org
[6] http://data.linkededucation.org/linkedup/catalog/
[7] http://data-observatory.org/lod-profiles

this initial stage a selection of 23 dataset of the catalog have been considered, as representative of datasets including different resource types related to several topics. Type mappings across all involved datasets link "documents" of all sorts to the common *foaf:Document* class, "persons" and "organisations" to the common *foaf:Agent* class, and course and module to the *aiiso:KnowledgeGrouping*[8] class. Categories associated with each dataset are shown in an interactive graph, generated for the specific types only, allowing for more representative and meaningful classification and exploration of datasets (Figure 1).

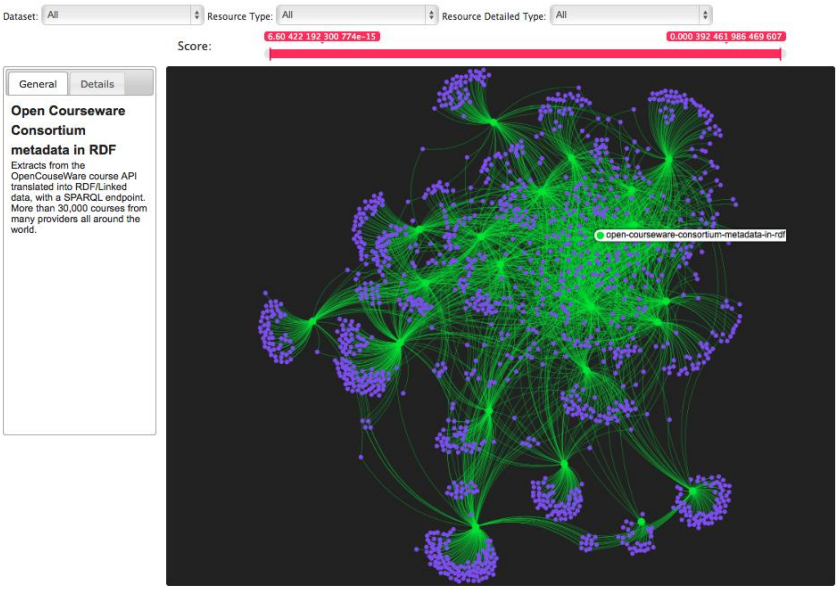

Fig. 1. A screenshot of the demo

2 The Dataset Profile Explorer

The dataset explorer is available at: http://data-observatory.org/led-explorer/. The explorer is composed of three panels: the panel at the center of the screen shows the network of datasets and categories, the panel on the left shows general and detailed descriptions about datasets and categories, and at the top of these panels the selection panel allows users to apply specific filters on the network. In the central panel, green nodes represent datasets while blue nodes represent categories. An edge connects a dataset to a category if the category belongs to the dataset topic profile. In order to draw the network, the *sigmajs*[9] library has been used and the nodes of the network have been displayed using the ForceAtlas2 layout. By clicking on a node (dataset or category), general and detailed descriptions are shown on the left panel. In the case of a dataset,

[8] http://purl.org/vocab/aiiso/schema#KnowledgeGrouping
[9] http://sigmajs.org

the general description reports the description of the dataset retrieved from the Datahub repository[10]. In the detailed description, the list of the top ten categories (and the related score) associated to the dataset is reported. In the case of a category, the description panel reports the list of datasets which have that category in their profile. The datasets including the category in their top ten list are highlighted in bold.

The selection panel at the top allows users to filter the results by means of three combo boxes, respectively related to: dataset, resource type, and resource *sub-type*. The list of dataset is composed by the dataset of the LinkedUp catalog. Regarding the resource type, the explorer is focused on three classes: *foaf:Document*, *foaf:Agent* and *aiiso:KnowledegeGrouping*. The *foaf:Document* is related to learning material such as: research papers, books, and so on; the *foaf:Agent* resource type has been included to take into account elements such as persons and organizations. The *aiiso:Knowledege-Grouping* is a type representing resources related to courses and modules. This initial set of resource type can be easily enlarged by means of configuration settings. The resource sub-type has been included with the aim of refining the results already filtered by resource type. Another filter that has been included into the explorer is related to the score of the relationships between datasets and categories. A slider bar allows users to filter results based on a specific range of the scores. As stated before, the scores have been calculated by the linked dataset profiling pipeline. The filters on datasets, resource types and resource sub-types can be combined and, as a result, only the portion of the network consistent with the filter selections is highlighted

3 Conclusion

In order to foster an effective use of the resources in the LD cloud, it is important to make explicit the topics covered by the datasets even in relation to the types of resources in the datasets. To this aim, we have developed a dataset profile explorer focused on the domain of educational related datasets. In this domain, topic coverage and the type of the resources assume a key role in supporting the search for content suitable for a specific learning course. The explorer allows users to navigate topic profiles associated with datasets with respect to the type of the resource in the dataset.

The explorer can be configured to be used with different datasets provided that the dataset topic profile is available, thus extending the application of the proposed approaches to several fields.

4 References

1. Fetahu, B., Dietze, S., Nunes, B. P., Taibi, D., Casanova, M. A., Generating structured Profiles of Linked Data Graphs, 12th International Semantic Web Conference (ISWC2013), Sydney, Australia, (2013).
2. D'Aquin, M., Adamou, A., Dietze, S., Assessing the Educational Linked Data Landscape, ACM Web Science 2013 (WebSci2013), Paris, France, May 2013.

[10] http://datahub.io