A Visual Summary for Linked Open Data sources

Fabio Benedetti, Sonia Bergamaschi, Laura Po

Università di Modena e Reggio Emilia - Dipartimento di Ingegneria "Enzo Ferrari" - Italy
firstname.lastname@unimore.it

Abstract. In this paper we propose LODeX, a tool that produces a representative summary of a Linked open Data (LOD) source starting from scratch, thus supporting users in exploring and understanding the contents of a dataset. The tool takes in input the URL of a SPARQL endpoint and launches a set of predefined SPARQL queries, from the results of the queries it generates a visual summary of the source. The summary reports statistical and structural information of the LOD dataset and it can be browsed to focus on particular classes or to explore their properties and their use. LODeX was tested on the 137 public SPARQL endpoints contained in Data Hub (formerly CKAN)[1], one of the main Open Data catalogues. The statistical and structural information extraction was successfully performed on 107 sources, among these the most significant ones are included in the online version of the tool[2].

1 Introduction

The RDF Data Model plays a key role in the birth and continuous expansion of the Web of data since it allows to represent structured and semi-structured data. However, while the LOD cloud is still growing, we assist to a lack of tools able to produce a meaningful, high level representation of these datasets.

Quite a lot of portals catalog datasets that are available as LOD on the Web and permit users to perform keyword search over their list of sources. Nevertheless, when a user starts exploring in details an unknown LOD dataset, several issues arise: (1) the difficulty in finding documentation and, in particular, a high level description of classes and properties of the dataset; (2) the complexity of understanding the schema of the source, since there are no fixed modeling rules; (3) the effort to explore a source with a high number of instances; (4) the impossibility, for non skilled users, to write specific SPARQL queries in order to explore the content of the dataset.

To overcome the above problems, we devise LODeX, a tool able to automatically provide a high level summarization of a LOD dataset, including its inferred schema. It is composed by several algorithms that discern between intensional and extensional knowledge. Moreover, it handles the problem of long running queries, that are subject to timeout failures, by generating a pool of low complexity queries able to return the same information.

This work has been accomplished in the framework of a PhD program organized by the Global Grant Spinner 2013, and funded by the European Social Fund and the Emilia Romagna Region.

[1] http://datahub.io

[2] http://dbgroup.unimo.it/lodex

As presented in [3], the majority of the tools for data visualization is not able to provide a synthetic view of the data (instances) contained in a single source. Payola[3] [4] and LOD Visualization[4] [2] are two recent tools that exploits analysis functionalities for guiding the process of visualization. However, these tools always need some querying parameters to start the analysis of a LOD dataset. Conversely, LODeX neither requires any a priori knowledge of the dataset, nor asks users to set any parameters; it focuses on extracting the schema from a LOD endpoint and producing a summarized view of the concepts contained in the dataset.

The paper is structured as follows. Section 2 describes the architecture of LODeX, while a use case and demonstration scenario is described in Section 3. Conclusions and some ideas for future work are described in Section 4.

2 LODeX - Overview

LODeX aims to be totally automatic in the production of the schema summary.

Figure 1 depicts the architecture of LODeX. The tool is composed by three main processes: *Index Extraction*, *Post-processing* and *Visualization*. The goal of the first two steps is to automatically extract from a SPARQL endpoint the information needed to produce its schema summary, while the third step aims to produce a navigable view of schema summary for the users. For an easy reuse, all the contents extracted and processed by LODeX are stored in a NoSQL document database, since it allows a flexible representation of the indexes.

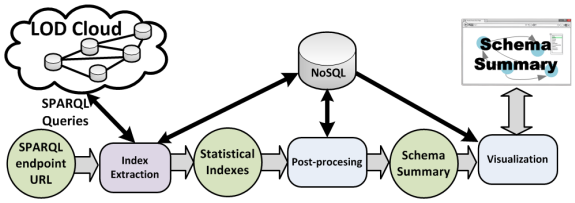

Fig. 1. LODeX Architecture

The **Index Extraction (IE)** takes as input the URL of a SPARQL endpoint and generates the queries needed to extract structural and statistical information about the source. Major details about the IE process can be found in [1]. The IE component has been designed in order to maximize the compatibility with LOD sources and minimize the costs in terms of time and computational complexity. The intensional and extensional knowledge are extracted and collected in a set of statistical indexes, stored in the NoSQL Database.

The **Post-processing (PP)** combines the information contained in the statistical indexes to produce the schema summary of a specific dataset. The summary is induced

[3] http://live.payola.cz/

[4] http://lodvisualization.appspot.com/

from the distribution of the instances in the dataset. The PP also collects synthetic information regarding the endpoint. Also the schema summary is stored in the NoSQL database.

The **Visualization** of the schema summary is performed through a web application written in Python that uses NoSQL database as backend. We used Data Driven Documents[5] to create a visual representation of the dataset with which the user can interact to navigate the schema and discover the information that he/she is looking for.

The tool has been tested on the entire set of sources described in *SPARQL Endpoint Status*(SPARQLES)[6], a specialized application that recursively monitors the availability of public SPARQL Endpoints contained in DataHub. At the time of our evaluation (May 2014), SPARQLES indicated that the 52% of SPARQL endpoints (244/469) were available and only the 13% of the endpoints presented a documentation, i.e. VoID and/or Service descriptions. LODeX was able to complete the extraction phase, thus building the visual summaries, for 107 LOD sources (78% of the 137 dataset that were compliant with the necessary SPARQL operators) that are now collected and shown in the online demo.

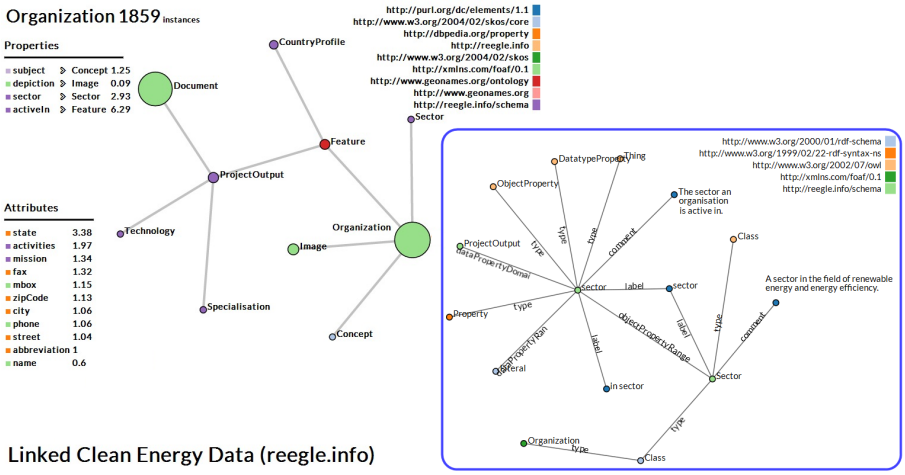

Fig. 2. Visual Summary of the Linked Clean Energy Data source and a particular of the "Sector" property.

3 Use Case and Demonstration Scenario

We refer to an hypothetical use-case involving a company in the clean energy sector. The company has its own products and services and attempts to discover new information on renewable energy and energy efficiency in the country where it is located. While

[5] http://d3js.org/

[6] http://sparqles.okfn.org/

searching the key datasets in the energy field, the company will likely find the Linked Clean Energy Data dataset[7]. This dataset, composed of 60140 triples, is described as a "Comprehensive set of linked clean energy data including: policy and regulatory country profiles, key stakeholders, project outcome documents and a thesaurus on renewable, energy efficiency and climate change for public re-use".

By using our application to explore this dataset (see Figure 2)[8], the user can, at a glance, have the intuition of all the instantiated classes (the nodes in the graph) and the connections among them (the arcs), besides the number of instances defined for each class (reflected in the dimension of the node). Focusing on the color of the nodes in the graph, a user can understand which classes are defined by the provider of the source and which others are taken from external vocabularies (in this case we can see that some of the class definitions are acquired from Foaf, Geonames.org and Skos). By positioning the mouse on a node, more information about the class are shown (as depicted in Figure 2 on the left). Since classes are linked to each others by some properties, it is possible to explore the property details. Thus, by clicking on a property another visual representation of the intensional knowledge is shown (see the right part of Figure 2).

4 Conclusions and Future Work

This paper has shown how LODeX is able to provide a visual and navigable summary of a LOD dataset including its inferred schema starting from the URL of a SPARQL Endpoint. The result gained by LODeX could also be useful to enrich LOD sources' documentation, since the schema summary can be easily translated with respect to a vocabulary and inserted into the LOD source. LODex is currently limited to display the contents of a source proposing a graph. However, new developments are being implemented in order to facilitate the query definition by exploiting the visual summary.

References

1. F. Benedetti, S. Bergamaschi, and L. Po. Online index extraction from linked open data sources. To appear in Linked Data for Information Extraction (LD4IE) Workshop held at International Semantic Web Conference, 2014.
2. J. M. Brunetti, S. Auer, and R. Garca. The linked data visualization model. In *International Semantic Web Conference (Posters & Demos)*, 2012.
3. A.-S. Dadzie and M. Rowe. Approaches to visualising linked data: A survey. *Semantic Web*, 2(2):89–124, 2011.
4. J. Klímek, J. Helmich, and M. Nečaskỳ. Payola: Collaborative linked data analysis and visualization framework. In *The Semantic Web: ESWC 2013 Satellite Events*, pages 147–151. Springer, 2013.

[7] http://data.reegle.info/

[8] The visual summary of this source is available at http://dbgroup.unimo.it/lodex/157