LODeX: A tool for Visual Querying Linked Open Data

Fabio Benedetti, Sonia Bergamaschi, Laura Po

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

Abstract. Formulating a query on a Linked Open Data (LOD) source is not an easy task; a technical knowledge of the query language, and, the awareness of the structure of the dataset are essential to create a query. We present a revised version of LODeX that provides the user an easy way for building queries in a fast and interactive manner. When a user decides to explore a LOD source, he/she can take advantage of the Schema Summary produced by LODeX (i.e. a synthetic view of the dataset's structure) and he/she can pick graphical elements from it to create a visual query. The tool also supports the user in browsing the results and, eventually, in refining the query. The prototype has been evaluated on hundreds of public SPARQL endpoints (listed in Data Hub[1]) and it is available online at `http://dbgroup.unimo.it/lodex2`. A survey conducted on 27 users has demonstrated that our tool can effectively support both unskilled and skilled users in exploring and querying LOD datasets.

1 Introduction

Since 2008, SPARQL has been defined as the W3C standard to query LOD sources. Today, it is the major query language for the Semantic Web, thus also the unique way to access and explore the constantly growing LOD Cloud. Several tools have attempted to facilitate exploring and querying LOD datasets, but it still remains complex and restricted to skilled users.

For facilitating the end users in the creation of queries on a new LOD source, we argue that it is necessary to provide a comprehensive view of the source itself. In this way, the structure of the dataset is revealed and the composition of correct and consistent queries becomes an easy task.

LODeX implements this idea: on the basis of the Schema Summary provided by the first version of the tool [3][2], it now supports visual querying of a LOD source. The new functionalities are: (1) a new interface that allows to compose a visual query browsing the Schema Summary of a source; (2) a SPARQL compiler automatically produces and submits the corresponding SPARQL query to the SPARQL endpoint and the result is shown in a tabular view; (3) a refinement panel that allows to refine the visual query by adding or removing attributes, by defining some filter and ordering conditions.

In the literature, we found different visual querying tools that exploit different strategies. Some tools provides a visual query language (a graph that describes the SPARQL

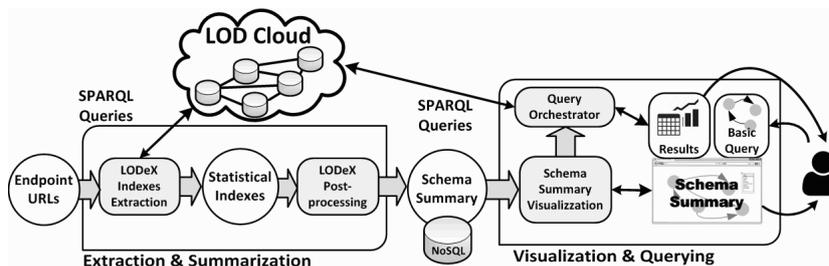
[1] `http://datahub.io`

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

query expression) like NITELIGHT, iSPARQL, RDF-GL. Even if they supply a visual composition, this is still strongly coupled to the syntax of the query language, thus, they remain unusable by unskilled users. Other approaches implement visual filter/flow graphs (series of nodes that implement operations, i.e. filter, sort, transform) like DE-RIPipes, MashQL, SparqlFilterFlow. Although these tools do not require any SPARQL skills, they do not provide any high level representation of the source's structure.

2   Backend

The architecture of LODeX relies on the composition of two main modules: *Extraction & Summarization* and *Visualization & Querying*, as shown in the Figure below.

The extraction of the statistical information and the production of the Schema Summary is implemented by a Python script and it uses a NoSQL database (MongoDB) to store these information. The process takes as input the URL of a SPARQL endpoint and it generates the queries needed to obtain structural and statistical information, i.e. *Statistical Indexes* [2], of the LOD source and then it combines these information to produce its Schema Summary.

We used different Javascript libraries to implement the Visualization & Querying module and to produce an interactive single page web application: Polymer, a new library based on Web Components, to manage the GUI; Data Driven Documents[3] in order to create the interactive Schema Summary; Sgvizler[4] to allow the querying of the remote endpoints. The crucial component of this module is the Query Orchestrator. It manages the interaction between the GUI and the user and it translates the visual query into a SPARQL query through a compiler. The compiler exploits an iterative algorithm that traverses the query to produce the SPARQL query, it supports a subset of SPARQL operators representing the core of the language: AND, OPTIONAL, FILTER, ORDER BY, LIMIT and OFFSET[5].

3   Frontend

Using the GUI of LODeX, the user is driven in the composition of a visual query by selecting classes, properties and attributes defined in the Schema Summary. A video tutorial is available at www.dbgroup.unimo.it/lodex2_video. The interface of LODeX

---

[3] http://d3js.org/

[4] http://dev.data2000.no/sgvizler/

[5] Major details about the LODeX model can be found at: http://goo.gl/FQIYLU

Fig. 1: An example of visual query and Schema Summary on the top ("Linked Clean Energy Data" dataset), and the refinement panel on the bottom.

is composed of three panels. The first panel lists all the datasets with some of their characteristics (number of triples, classes, properties etc.). After selecting a dataset, the browsable Schema Summary is displayed in the second panel and it is possible to start building a visual query (Fig. 1 A) by selecting the first node (i.e. a class) and by adding some of its attributes or properties (M: mandatory, O: optional). The color of the nodes, attributes and properties determine the belonging of the element to one of the vocabularies listed in the caption on the right of Fig 1. By positioning the mouse on a node, more information about the class are shown such as the class name, the number of instances (B), the properties (C) and the attributes (D). Each time a property is chosen, the connected class is added to the visual query. By repeating this process, the user might extend the query. By clicking on the "generate" button, the user is driven in the Refinement Panel where he/she can visualize the results and the corresponding SPARQL query. Here, the user may modify the query (F) adding/removing filter conditions, modifying the attribute list, sorting the records. After any change, the SPARQL query is compiled and the results updated.

Datasets: LODeX has been tested on 206 SPARQL endpoints compatible with SPARQL 1.1 listed in Datahub. We were able to make available online the Schema Summary of 110 endpoints for visual querying. A portability evaluation [4] has reduced the number of endpoints available online on LODeX to 110[6].

---

[6] Major details about this evaluation can be found at: `http://goo.gl/NhJHRf`

Usability evaluation: We designed a usability test[6] [4] to verify both if the graph visualization of the Schema Summary is a valid documentation for a dataset and if the process of building and refining a visual query is effective. We enrolled 27 anonymous users to fill an online survey; each functionality has been tested through micro-tasks followed by a SUS questionnaire. We obtained a median SUS [5] score of 82.5 that classifies LODeX as *Excellent*, according to [1].

4   Demonstration Scenario

The LODeX functionalities can be better emphasized by using datasets of different complexity, thus we will concentrate our attention on three datasets: the **Linked Clean Energy Data**, which is a comprehensive set of linked clean energy data; the **European Television Heritage** dataset, that contains information about the European television archives and acts as a domain aggregator for Europeana, Europes digital library; the **World War 1 as Linked Open Data** dataset, which contains information on events, actors and places related to the First World War. The demo is open to the active role of the audience, therefore any LOD data source available in LODeX can be used.

Three scenarios will be used to explore the benefits of the tool: **1 - Exploration of a LOD source**: we will show how a user can navigate through the Schema Summary of the three datasets to capture the underlying knowledge; **2 - Creation of a visual query**: we will start the composition of a simple query by selecting a class (and some attributes) and we invite the audience to try composing a visual query; **3 - Refinement of a query**: we will show how to refine a query by adding or removing some attributes, by filtering the query results with particular conditions and by ordering the results.

5   Conclusion

We described and demonstrated LODeX as a valuable tool to explore LOD sources and produce meaningful queries on them. Formulating a query based on the Schema Summary of a LOD source is easy and the user can avoid problems in naming classes, properties and he can be sure to generate consistent queries. The development and improvement of LODeX is still ongoing: we are working on clustering techniques to show an effective Schema Summary for large datasets; new query result formats (i.e. chart, plain text etc.), as well as adding more SPARQL operators (i.e. GROUP BY and UNION).

References

1. A. Bangor, P. Kortum, and J. Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123, 2009.
2. F. Benedetti, S. Bergamaschi, and L. Po. Online index extraction from linked open data sources. In *Proc.of the LD4IE Workshop 2014 co-located with the ISWC 2014*.
3. F. Benedetti, S. Bergamaschi, and L. Po. A visual summary for linked open data sources. In *ISWC (Posters & Demos)*, 2014.
4. F. Benedetti, S. Bergamaschi, and L. Po. Visual querying lod sources with lodex. To appear in Proceedings of the 8th international conference on Knowledge capture (K-CAP 2015), 2015.
5. J. Brooke. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.