ELLIS: Interactive Exploration of Linked Data on the Level of Induced Schema Patterns

Thomas Gottron[1], Malte Knauf[2], Ansgar Scherp[3], Johann Schaible[4]

[1] Innovation Lab, SCHUFA Holding AG, Wiesbaden, Germany
Thomas.Gottron@schufa.de
[2] Institute for Web Science and Technologies, University of Koblenz-Landau, Germany
mknauf@uni-koblenz.de
[3] ZBW – Leibniz Information Center for Economics, Kiel University, Kiel, Germany
asc@informatik.uni-kiel.de
[4] GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany
johann.schaible@gesis.org

Abstract. We present ELLIS, a demo to browse the Linked Data cloud on the level of induced schema patterns. To this end, we define *schema-level patterns* of RDF types and properties to identify how entities described by *type sets* are connected by *property sets*. We show that schema-level patterns can be aggregated and extracted from large Linked Data sets using efficient algorithms for mining frequent item sets. A subsequent visualisation of such patterns enables users to quickly understand which type of information is modelled on the Linked Data cloud and how this information is interconnected.

1 Introduction

The Linked Open Data (LOD) cloud does not have a fixed or pre-defined schema. However, the use of RDF types and properties to describe the data provides an emerging schema. This implicit schema can be induced from data observations on the Web and, thereby, can be made explicit. A subsequent visualisation of the induced schema information enables users to investigate the structure of Linked Data in an interactive and exploratory way. The insights and understanding of the data gained in this way are beneficial for several applications. It can help users in finding relevant vocabulary terms when modelling data as LOD [10] or in programming a Linked Data application that requires to obtain data of specific type and with specific properties [11]. Furthermore, it allows users to understand what type of information is available on the LOD cloud and how this information is interconnected on the Web of Data. In this paper, we present ELLIS, a graph-based approach for visualising and exploring induced schema information for Linked Data on the basis of *schema-level patterns*.

2 Schema-level Patterns

There are various approaches of different granularity for inferring schema information from observations made on the Linked Data cloud. For the purpose of providing a consistent and browsable view of schema-level information, we need to describe (at least)

two aspects: an aggregated representation of the entities modelled in the Linked Data graph as well as a notion of the relations connecting them. The entities can be grouped together on the basis of the sets of RDF types associated with them. Likewise, the sets of RDF properties interlinking the entities can serve to describe the relations between groups of entities of the same type. Hence, we model *schema-level patterns* (SLP) as a combination of subject type sets *sts* and object type sets *ots* (i. e., sets of RDF types T of entities modelled on the Linked Data cloud) which are connected by property sets *ps* (i. e., sets of predicates P). Formally, an SLP is defined as a triple

$$(\mathit{sts}, \mathit{ps}, \mathit{ots}) \in \mathcal{P}(T) \times \mathcal{P}(P) \times \mathcal{P}(T) \tag{1}$$

This schema-level representation of Linked Data lends itself for a graph-based interpretation and visualisation. As the subject and object type sets follow the same formal definition, they can be seen as nodes connected by edges consisting of property sets.

When computing SLPs for a (potentially distributed) segment R of the RDF data graph on the LOD cloud, we consider all URIs appearing in the subject position and object position of RDF triples (s, p, o), extract their RDF types and the unified set of all predicates used to model a relation between them. Formally, we define the set of observed SLPs over an RDF data set R:

$$\begin{aligned} \mathit{SLP}(R) = \{(\mathit{sts}, \mathit{ps}, \mathit{ots}) \mid \exists s, o : (\forall t_s \in \mathit{sts} : (s, \mathsf{rdf{:}type}, t_s) \in R) \\ \wedge (\forall p \in \mathit{ps} : (s, p, o) \in R) \wedge (\forall t_o \in \mathit{ots} : (o, \mathsf{rdf{:}type}, t_o) \in R)\} \end{aligned} \tag{2}$$

The set $\mathit{SLP}(R)$ can be computed with relatively little overhead from large data sets using the Apriori algorithm for frequent item set mining. As a result, we obtain the above mentioned graph structure over induced schema-level patterns.

3 ELLIS

Based on the definition of SLPs, we implemented the ELLIS prototype for visualising and navigating the LOD cloud on a schema level[5]. The system provides four essential functionalities: (a) a visualisation of SLPs as a graph, (b) browsable rendering of the graph nodes together with annotations of the relevant schema information, (c) a history trace to keep track of previous steps in the exploration path, and (d) a search functionality to find relevant entry points for browsing the SLP graph.

The graph visualisation represents the type set information as well as the property set information as nodes in a graph as shown in Figure 1. The edges connect the nodes in a directed way to indicate the order of the triple in an SLP starting from the subject type set over the connecting property set to the object type set. Representing all relevant information as nodes in a browsable graph has two advantages. First, it condenses information on a high level. This enables users to quickly grasp the structure of the data. When needed and requested, additional information can be revealed and displayed. In ELLIS we use hover info boxes and an additional info field in the menu to indicate the

[5] A screencast of ELLIS is publicly available at `https://www.youtube.com/watch?v=q47YFKyf32I&feature=youtu.be`.

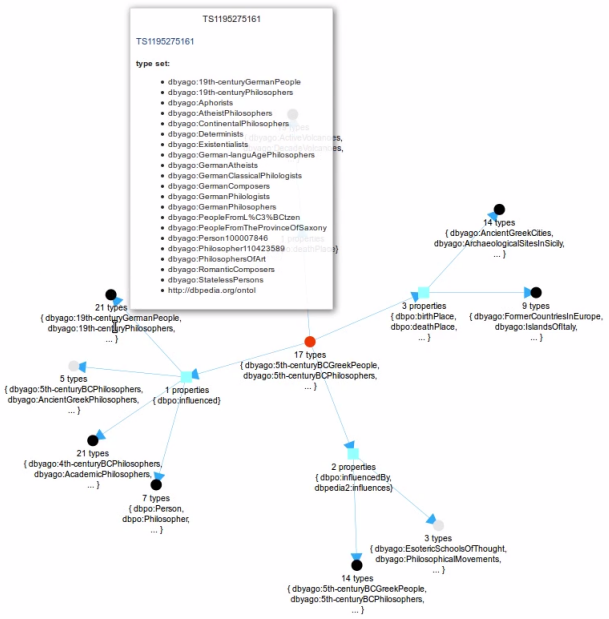

Fig. 1. Visualisation of an initial query over induced schema-level patterns in ELLIS.

type and property sets associated with nodes of the SLP graph. Second, the graph can easily be navigated by selecting any of the displayed type set nodes. Upon selection of a node, the visualisation interface updates the graph by retrieving all connected property sets and type sets as given by the SLPs. A history trace [1] allows the users to identify the path they took in the exploration of the LOD cloud on a schema level. SLPs in the history trace older than the last three steps are removed from the visualisation. This provides orientation and context without overloading the interface with all previously visited schema-level patterns. Finally, a search functionality permits the users to search for specific RDF types. Subsequently, ELLIS lists all type sets containing these types. In this way, it is possible to flexibly chose an entry point type set and the embedding SLPs for starting to browse the schema graph.

ELLIS is designed following a classical three-tier architecture. The Web front end visualises the graph constructed from SLPs, displays additional information, and provides interaction functionality. Figure 1 illustrates the graph visualisation in ELLIS. The middle tier encapsulates functions for search and navigation. In particular, it allows to resolve for a given type set node all relevant SLPs containing this type set as subject type set and object type set. The backend tier consists of a database containing all SLPs obtained from a Linked Data set. In our ELLIS demo, we constructed the SLPs from the BTC 2012 dataset, containing approximately 1.4 billion triples.

Figure 1 shows the result of an initial query about Greek philosophers to ELLIS. The best matching type set of the query is marked in red and shown in the middle of the graph. The related sets of RDF resources with a similar set of properties and types

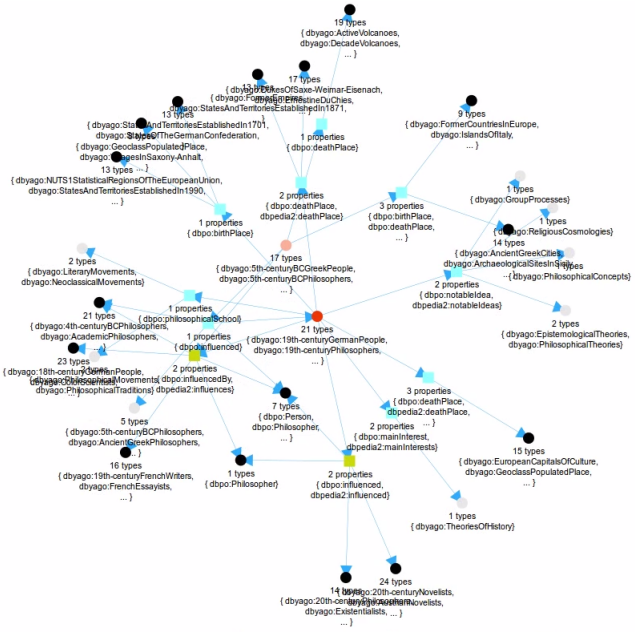

Fig. 2. Navigation of SLPs and history trace.

are connected via relations. In the example shown in Figure 1, these are properties like dbpo:influencedBy and dbpo:influenced. The user hovers with the mouse over a type set TS1195275161. It mainly contains German philosophers that are dbpo:influenced by the Greek philosophers. Subsequently, the user clicks on this type set of German philosophers in order to further navigate through the induced SLPs in ELLIS. The result is shown in Figure 2. The clicked type set is now indicated in red and moved to the center of the graph visualization. Further properties, such as the birthplace and place of death of the philosophers, of this node are shown and can be explored further.

4 Related Work

There are numerous approaches for inducing schema information from Linked Open Data. The applications vary from statistical schema inferencing [12] over cardinality estimation for query result sets [9] and analytics of the dynamics of Linked Data sources [4] to schema-level indices [6,8]. Most similar to the presented SLPs are the equivalence classes in SchemEX [8] or the Node-Collection Layer from the RDF graph summary [2], which capture even more fine grained schema information. Regarding the visualisation of Linked Data, most approaches address visualisation on an instance level [3]. In contrast, Katifori et al. [7] present a survey of different approaches for visualising ontologies, i. e, schema-level information (so-called T-Box). A more recent visualisation approach involving schematic information on the LOD cloud is LODSight [5]. It uses a dataset summarization algorithm which induces the schema from

a dataset via SPARQL queries. Such SPARQL queries can get quite complicated. ELLIS induces the schema via SLPs which are computed in a less complicated manner by using the Apriori algorithm for mining frequent item sets.

5 Conclusion

With schema-level patterns, we have defined a structure which is suitable for inducing and aggregating schema-level information from Linked Data. The ELLIS demo visualises schema-level patterns as a graph structure and allows for an interactive exploration and browsing of the schema information induced from the Linked Data cloud.

As future work, we plan to integrate the visualisation technique with a novel tool for modelling data as LOD [10]. It will allow data engineers to not only conduct textual queries to find relevant vocabulary terms for reuse but also enable them to visually explore terms that are related with the model they are working on.

References

1. Campbell, I.: Interactive evaluation of the ostensive model using a new test collection of images with multiple relevance assessments. Information Retrieval 2(1), 89–114 (2000)
2. Campinas, S., Perry, T.E., Ceccarelli, D., Delbru, R., Tummarello, G.: Introducing RDF graph summary with application to assisted SPARQL formulation. In: 23rd International Workshop on Database and Expert Systems Applications. pp. 261–266. IEEE (2012)
3. Dadzie, A.S., Rowe, M.: Approaches to visualising linked data: A survey. Semant. web 2(2), 89–124 (Apr 2011), http://dx.doi.org/10.3233/SW-2011-0037
4. Dividino, R., Gottron, T., Scherp, A.: Strategies for efficiently keeping local linked open data caches up-to-date. In: The Semantic Web-ISWC 2015, pp. 356–373. Springer (2015)
5. Dudáš, M., Svátek, V., Mynarz, J.: Dataset summary visualization with lodsight. In: The Semantic Web: ESWC 2015 Satellite Events, pp. 36–40. Springer (2015)
6. Gottron, T., Gottron, C.: Perplexity of Index Models over Evolving Linked Data. In: ESWC'14: Proceedings of the Extended Semantic Web Conference. pp. 161–175 (2014)
7. Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., Giannopoulou, E.: Ontology visualization methods—a survey. ACM Comput. Surv. 39(4) (Nov 2007)
8. Konrath, M., Gottron, T., Staab, S., Scherp, A.: SchemEX—Efficient Construction of a Data Catalogue by Stream-based Indexing of Linked Data. Web Semantics: Science, Services and Agents on the World Wide Web 16(5), 52 – 58 (2012), the Semantic Web Challenge 2011
9. Neumann, T., Moerkotte, G.: Characteristic sets: Accurate cardinality estimation for rdf queries with multiple joins. In: Proceedings of the 27th International Conference on Data Engineering, ICDE 2011. pp. 984–994. IEEE Computer Society (2011)
10. Schaible, J., Gottron, T., Scheglmann, S., Scherp, A.: Lover: support for modeling data using linked open vocabularies. In: Joint 2013 EDBT/ICDT Conferences, EDBT/ICDT '13, Genoa, Italy, March 22, 2013, Workshop Proceedings. pp. 89–92. ACM (2013)
11. Scheglmann, S., Leinberger, M., Gottron, T., Staab, S., Lämmel, R.: Sepal: Schema enhanced programming for linked data. KI-Künstliche Intelligenz pp. 1–4 (2015)
12. Völker, J., Niepert, M.: Statistical schema induction. In: The Semantic Web: Research and Applications - 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29-June 2, 2011, Proceedings, Part I. Lecture Notes in Computer Science, vol. 6643, pp. 124–138. Springer (2011)