Exploring RDF Usage and Interlinking in the Linked Open Data Cloud using ExpLOD

Shahan Khatchadourian
University of Toronto
shahan@cs.toronto.edu

Mariano P. Consens
University of Toronto
consens@cs.toronto.edu

ABSTRACT

Publishing interlinked RDF datasets as links between data items identified using dereferenceable URIs on the web brings forward a number of issues. A key challenge is to understand the data, the schema, and the interlinks that are actually used both within and across linked datasets. Understanding actual *RDF usage* is critical in the increasingly common situations where terms from different vocabularies are mixed.

In this demonstration we present a tool, ExpLOD, that supports exploring summaries of RDF usage and interlinking among datasets from the Linked Open Data cloud. ExpLOD's summaries are based on a novel mechanism that combines text labels and bisimulation contractions. The labels assigned to resources in RDF graphs are hierarchical, enabling summarization at different granularities. The bisimulation contractions are applied to subgraphs defined via queries, providing for summarization of arbitrary large or small graph neighbourhoods. Our tool also generates SPARQL queries from summaries.

1. OVERVIEW

A key challenge of describing Linked Data [2] datasets is to understand the data, schema, and interlinks that are actually used both within and across linked datasets. Understanding how resources are used (or *RDF usages*, as in [4]) is critical in the increasingly common situations where terms from different vocabularies are mixed. This is because developers need to know the concepts mentioned in the dataset in order to contribute interlinks between datasets, or for novel mashups (enabled by the abundance of linked data, see [5]) involving resources described across several datasets.

Existing solutions to the challenge currently involves using an RDF browsers (such as [3] and [1]) to explore linked data which can be a tedious and time-consuming exercise for large datasets. A trial-and-error approach on a schema-conformant dataset means testing for structures permitted by the schema and becomes complex if the dataset does not use the full schema or uses multiple schemas due to the combinatorial ways of mixing terms. In the LOD cloud this is exacerbated since there is no schema describing interlinking *between* datasets, so another approach is needed.

In this demonstration we present a tool, ExpLOD, that supports exploring summaries of RDF usage and interlinking among datasets from the Linked Open Data cloud. An RDF summary can be used to describe a dataset by grouping equivalent RDF resources. ExpLOD's summaries are based on a novel mechanism that combines text labels and bisimulation contractions (see [6] for details).

LDOW2010, April 27, 2010, Raleigh, USA.

The summaries produced by ExpLOD can be created, viewed, and explored in an interactive graphical environment (and they can also be exported in a variety of formats, including RDF). ExpLOD is a Java application developed within the Eclipse environment with support for plug-ins. Custom code was developed using the Jena toolkit (`jena.sourceforge.net`). ExpLOD can also invoke the Virtuoso RDF store (`www.openlinksw.com/virtuoso`).

2. EXPLOD DEMONSTRATION

Summaries of RDF data are computed using bisimulation labels (BL) based on RDF usage. RDF usage, discussed and enumerated in [4], is a way describe the semantics of RDF resources based on how they are used, such as whether a resource is used as part of the data or schema. The demonstration describes the flexibility of modifying the labeling scheme to obtain a coarser or more detailed summary for datasets in the LOD cloud (Section 2.1). We also demonstrate how a larger RDF usage neighbourhood can be employed to describe interlinked datasets (Section 2.2).

2.1 Class and Predicate Usage

Four RDF usages that describe the interaction of data and meta-data are: (i) *class instantiation*, the number of instances that are typed as a particular class; (ii) *predicate instantiation*, the number of times a predicate is used to describe all instances; (iii) *class usage*, the *sets* of classes to which instance belongs; and (iv) *predicate usage*, the *sets* of predicates used to describe an instance. A class and predicate RDF usage summary is used to describe DBTune's Jamendo, a dataset from the LOD cloud containing information about music artists and their productions. Our goal is to understand how tracks and records are described

Figure 1 shows the RDF usage summary of class and predicate instantiation, and class and predicate usage, for records and tracks in Jamendo. Class instantiation for each class is reported in its block extent size (between parentheses). So 5,786 instances are typed as records, as can be seen in class block 5 whose BL begins with 'C/'. Even though class instantiation cannot distinguish instances that belong to more than one class, it is possible to do so with class usage. The class usage of instance block 1430 is the singleton *mo:Track* since there is only class block reachable from the instance block. Predicate instantiation is captured in the extent size of predicate blocks (whose BL has prefix 'P/'). For example, the predicate *mo:license* has been used to describe instances 45,634 times. The predicate usage of instance block 1430 (containing tracks) is the set of predicates {*mo:license*,

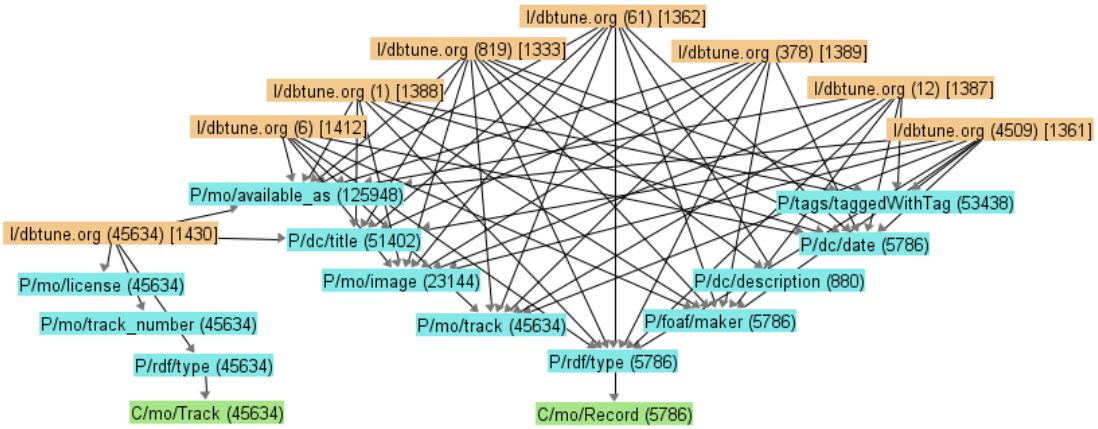

Figure 1: Jamendo: RDF usage summary of class and predicate instantiation, and class and predicate usage

mo:track_number, *mo:available_as*, *dc:title*, *rdf:type*}, visible by the edges from the instance block to each of those predicate blocks. Notice that grouping instances having the same class usage *and* predicate usage into the same instance block (the extent size reports the number of instances) results in 7 record instance blocks, each of which has a unique predicate usage. Amongst records, there is variation in the extent size of predicate usages - the extent size of instance block 1361 is 4,509 compared to the extent size of instance block 1388 that contains only 1 record instance.

To reduce the number of blocks in an RDF usage summary, using a reduced portion of the BL hierarchy can sometimes produce a summary with fewer blocks, and as many blocks as before (in the worst case). For example, excluding the local part of each predicate's BL groups predicates by their namespace. Applying this modification to the BL to the summary in Figure 1 produces the summary shown in Figure 2. The change in BL reduced the number of instance blocks from 8 to 3.

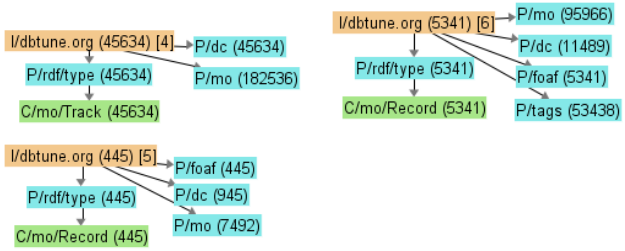

Figure 2: Jamendo: predicate usage summary of records and tracks grouped by namespace

2.2 Interlinking

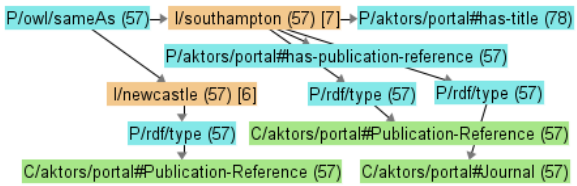

Figure 3: Southampton and Newcastle: Interlink usage summary

Two interlinked RDF datasets in the LOD cloud may contain information about the same real-world entity, but it is possible that each dataset uses its own unique URI to represent it. A triple with an *owl:sameAs* predicate captures the information that the subject and object URIs refer to the same real-world entity, referred to as *URI-equivalence*; such statements can be found readily within many datasets in the LOD cloud. A resource's description in one dataset may not match the description of its URI-equivalent resource in another dataset.

An *interlink usage* summary is created from the bisimulation contractions of URI-equivalent resources is used to understand each dataset's contribution to a real-world entity's description. An interlink usage neighbourhood is the subgraph that includes the nodes and edges on the path linking the subject of a triple whose predicate is *owl:sameAs* to the object node. Additionally, the incoming edge of each *owl:sameAs* predicate node is reversed, both in the labeled graph and the ExpLOD graph, so that it points to the subject. Since we are interested in the description of instances, we consider only statements in which the subject and object are instances. Creating a summary based on interlink usage neighbourhoods as well as class and predicate usage neighbourhoods provides additional information about URI-equivalent resource descriptions.

An interlink usage summary of URI-equivalent instances from two datasets in the RKB Explorer collection is shown in Figure 3. It displays a block containing 57 instances that have the same class and predicate usage in Southampton, and the block of URI-equivalent instances in Newcastle that have the same class usage. This asymmetry is chosen as a way to demonstrate the flexibility of the neighbourhoods that are summarized.

3. REFERENCES

[1] S. Bernhard. Tripcel: Exploring RDF Graphs using the Spreadsheet Metaphor. In *ISWC*, 2009.
[2] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee. Linked data on the web. In *LDOW*, pages 1265–1266, 2008.
[3] S. F. C. de Araujo and D. Schwabe. Explorator: a tool for exploring RDF data through direct manipulation. *LDOW*, 2009.
[4] L. Ding and T. Finin. Characterizing the Semantic Web on the Web. In *ISWC*, pages 242–257, 2006.
[5] M. Hausenblas, W. Halb, Y. Raimond, and T. Heath. What is the Size of the Semantic Web? In *I-SEMANTICS*, pages 9–16, 2008.
[6] S. Khatchadourian and M. P. Consens. Exploring RDF Usage and Interlinking in the Linked Open Data Cloud using ExpLOD. *To appear in ESWC 2010.*