Lost in Semantics? Ballooning the Web of Data

ARTICLE · JANUARY 2014

3 AUTHORS:

Florian Stegmaier
ONE LOGIC GmbH
34 PUBLICATIONS **44** CITATIONS
SEE PROFILE

Kai Schlegel
Universität Passau
13 PUBLICATIONS **7** CITATIONS
SEE PROFILE

Michael Granitzer
Universität Passau
202 PUBLICATIONS **699** CITATIONS
SEE PROFILE

words, observation parameters and measuring units (bottom part of the picture), it is then possible to populate widgets with content that is tailored to the user's profile (e.g., datasets and services applicable to the user's research area). Also, recourse to a gazetteer for expressing geographic locations simplifies discovery by reducing the need for a map.

On the back-end side, the architecture comprises collaboration tools for supporting activities that are not directly related to the discovery of resources but that, nevertheless, constitute essential phases in data production (e.g., the management of field work). Once users start enriching their profile data (even by simply using the infrastructure), the user experience should slowly but steadily converge to the user's expectations.

Acknowledgement
The activities described in this paper have been funded by the Italian Flagship Project RITMARE.

Links:
RITMARE Flagship Project: http://www.ritmare.it
Data Catalog Vocabulary (DCAT): http://www.w3.org/TR/vocab-dcat/
BODC webservices: http://seadatanet.maris2.nl/v_bodc_vocab/welcome.aspx

References:
[1] P. Carrara et al.: "An interoperable infrastructure for the Italian Marine Research", IMDIS 2013
[2] Commissione di Coordinamento SPC: "Linee guida per l'interoperabilità semantica attraverso i Linked Open Data", 2013

Please contact:
Cristiano Fugazza, CNR-IREA, Italy
E-mail: fugazza.c@irea.cnr.it

Lost in Semantics? Ballooning the Web of Data

by Florian Stegmaier, Kai Schlegel and Michael Granitzer

Although Linked Open Data has increased enormously in volume over recent years, there is still no single point of access for querying the over 200 SPARQL repositories. The Balloon project aims to create a Meta Web of Data focusing on structural information by crawling co-reference relationships in all registered and reachable Linked Data SPARQL endpoints. The current Linked Open Data cloud, although huge in size, offers poor service quality and is inadequately maintained, thus complicating access via SPARQL endpoints. This issue needs to be resolved before the Linked Open Data cloud can achieve its full potential.

Today's vision of a common Web of Data is largely attributable to the Linked Open Data movement. The first wave of the movement transformed silo-based portions of data into a plethora of open accessible and interlinked data sets. The community itself provided guidelines (e.g., 5 stars Open Data) as well as open source tools to foster interactions with the Web of data. Harmonization between those data sets has been established at the modelling level, with unified description schemes characterizing a formal syntax and common data semantic.

Without doubt, Linked Open Data is the de-facto standard to publish and interlink distributed datasets within the Web commonly exposed in SPARQL endpoints. However, a request considering the globally described data set is only possible with strong limitations:

- The distributed nature of the Linked Open Data cloud in combination with the large amount of reachable endpoints hinders novice users from interacting with the data.
- Following the Linked Data principle, specific URIs are in use to describe specific entities in the endpoints and are further resolvable to get further information on the given entity. The problem arises since each endpoint uses its own URI to describe the single semantic entities leading to semantic ambiguities.

One outcome of the EU FP7 CODE project is the Balloon framework. It tackles exactly this situation and aims to create a Meta Web of Data focusing on structural information. The basement for this is a crawled subset of the Linked Data cloud, resulting in a co-reference index as well as structural information. The main idea behind this index is to resolve the aforementioned semantic ambiguities by creating sets of semantically equivalent URIs to ease consumption of Linked Open Data. This is enabled by crawling information expressing the links between the endpoints. For this purpose, we consider a specific set of predicates, e.g., sameAs or exactMatch, to be relevant. The complete crawling process relies on SPARQL queries and considers each LOD endpoint registered at the CKAN platform. Here, RDF dumps are explicitly excluded. During the crawling, a clustering approach creates the co-reference clusters leading to a bi-directional view on the co-reference relationships and is the result of a continuous indexing process of SPARQL endpoints. In addition to properties defining the equality of URIs, the indexing service also takes into account properties that enable structural analysis on the data corpus, e.g., rdfs:subclass. On the basis of this data corpus, interesting modules and application scenarios can be defined. For instance, on-going research is focusing on the creation of the following two modules as starting point:

- Intelligent and on the fly query rewriting by utilizing co-reference clusters and SPARQL 1.1 Federated Query.
- Data analysis, e.g., retrieving common properties or super types for a given set of entities.

These modules are integrated in the overall Balloon platform and serve as a starting point for further applications. To foster community uptake and to increase available modules in the platform, the Balloon project along with the data corpus will soon be made available as open source project.

The idea of leveraging co-reference information is nothing new: The Silk

framework [1], SchemEX [2] and the well-known sameAs.org project proposed similar techniques. Nevertheless, the Balloon co-reference approach further considers consistent data provenance chains and the possibilities of cluster manipulations to enhance the overall quality and correctness. Further, the explicit limitation to LOD endpoints sets a clear focus on the data that is (in principle) retrievable, in contrast to RDF dumps that are not searchable out of the box.

While creating the co-reference index, we encountered several issues in the current Linked Open Data cloud. Missing maintenance of endpoints over years as well as a lack of quality of service hinders the Linked Open Data cloud from reaching its potential. Our findings gathered during the crawling process are in keeping with the current statistics provided by the LOD2 project of the Linked Open Data cloud: From a total of 700 official data sets, only approximately 210 are enclosed in a SPARQL endpoint and registered at the CKAN platform. Further, more than half of the available endpoints had to be excluded due to insufficient support of SPARQL as well as unattainability. Finally, only 112 endpoints have been actively crawled for co-reference information leading to a total of 22.4M distinct URIs (approx. 8.4M synonym groups). During the crawling phase we also encountered the need for a SPARQL feature lookup service. The main intention is to describe the actually supported retrieval abilities of an endpoint in a standardized way. Currently, discussions on this topic are observable at community mailing lists.

Links:
Code Project: http://code-research.eu/
Overview of Balloon: http://schlegel.github.io/balloon/
Crawled data: ftp://moldau.dimis.fim.uni-passau.de/data/ (on-going research, frequently/live updated)
5 stars Open Data: http://5stardata.info/
CKAN platform: http://ckan.org/
LOD2 project: http://stats.lod2.eu/
Community mailing lists: http://lists.w3.org/Archives/Public/public-lod/2013Oct/0077.html

References:
[1] J. Volz et al: "Silk–a link discovery framework for the web of data," in proc. of the 2nd Linked Data on the Web Workshop, 2009, pp. 559–572, http://events.linkeddata.org/ldow2009/papers/ldow2009_paper13.pdf

[2] M. Konrath, et al: "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, vol. 16, no. 5, 2012, http://www.websemanticsjournal.org/index.php/ps/article/view/296/297

Please contact:
Florian Stegmaier, Kai Schlegel, Michael Granitzer
University of Passau, Germany
Tel: +49 851 509 3063
E-mail: <forename.surname>@uni-passau.de

Publishing Greek Census Data as Linked Open Data

by Irene Petrou and George Papastefanatos

Linked Open Data technology is an emerging way of making structured data available on the Web. This project aims to develop a generic methodology for publishing statistical datasets, mainly stored in tabular formats (e.g., csv and excel files) and relational databases, as LOD. We build statistical vocabularies and LOD storage technologies on top of existing publishing tools to ease the process of publishing these data. Our efforts focus on census data collected during Greece's 2011 Census Survey and provided by the Hellenic Statistical Authority. We develop a platform through which the Greek Census Data are converted, interlinked and published.

Statistical or fact-based data about observations of socioeconomic indicators are maintained by statistical agencies and organizations, and are harvested via surveys or aggregated from other sources. Census data include demographic, economic, housing and household information, as well as a set of indices concerning the population over time, such as mortality, dependency rate, total fertility rate, life expectancy at birth, etc.

The main objective in publishing socio-demographic data, such as census data, as LOD is to make these data available in an easier-to-process format (they can be crawled or queried via SPARQL), to be identifiable at the record level through their assignment with URIs and finally to be citable, ie, make it possible for other sources to link and connect with them. Being available in LOD format will make them easier to access and use by third parties, facilitating data exploration and the development of novel applications. Furthermore, publishing Greek census data as LOD will facilitate their comparison and linkage with datasets derived from other administrative resources (e.g. public bodies, Eurostat, etc.), and deliver consistency and uniformity between current and future census results.

Best practices for publishing Linked Data encourage the reuse of vocabularies for describing common concepts in a specific domain. In this way, interoperability and interlinking between published datasets is achieved. In the statistics field, a number of statistical vocabularies and interoperability standards have been proposed, such as the SDMX (Statistical data and metadata exchange) standard, the Data Cube Vocabulary, and SCOVO. In our approach, we employ the Data Cube Vocabulary for representing census results. The Data Cube Vocabulary relies on the multidimensional (or cube) model. Its main components are the