Proceedings of the I-SEMANTICS 2012 Posters & Demonstrations Track, pp. 26-30, 2012.

Linked Open Data Infrastructure for Public Sector Information: Example from Serbia

Valentina Janev[1], Uroš Milošević[1], Mirko Spasić[1], Jelena Milojković[2], Sanja Vraneš[1]

[1]Mihailo Pupin Institute, University of Belgrade, Belgrade, Serbia
{valentina.janev, uros.milosevic, mirko.spasic,
sanja.vranes@pupin.rs}
[2]Statistical Office of the Republic of Serbia, Belgrade, Serbia
{jelena.milojkovic@stat.gov.rs}

Abstract. To improve transparency and public service delivery, national, regional and local governmental bodies need to consider new strategies to openning up their data. We approach the problem of creating a more scalable and interoperable Open Government Data ecosystem by considering the latest advances in Linked Open Data. More precisely, we showcase how an integrated and coherent collection of aligned state of the art software tools, the LOD2 Stack, can be used to deliver trusted, open and rich collections of interlinked datasets to the public. The usage of the Tool Stack is demonstrated on the case of one of the largest data providers in the Republic of Serbia – its Statistical Office.

Keywords. linked open data, open government data, infrastructure, tools, public sector, Serbia

1 Introduction

In order to improve efficiency in the provision of public services, increase transparency and interaction with citizens and society as a whole, but also create new businesses and job opportunities, both local and national governments need to find better strategies for delivering large amounts of trusted data to the public. The fact that the European Commission is investing considerable amounts of finances to overcome this problem is a strong indicator of its significance. As a direct example, consider the ISA (Interoperability Solutions for European Public Administrations) program for the period from 2010-2015 that has been assigned a budget of 164,1 million euros[1]. The program enables "the delivery of electronic public services and ensures the availability, interoperability, re-use and sharing of common solutions"[2]. To make government data truly open (for use and re-use), and increase transparency, it needs to be published in a non-proprietary, machine-readable format (e.g. RDF, http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210).

[1] European Commission ISA Webpage, http://ec.europa.eu/isa/
[2] European Commission ISA Webpage, http://ec.europa.eu/isa/faq/faq_en.htm

In this paper, we will show why Linked Data is considered a promising approach to the above problem, and how the LOD2 Stack, a powerful set of software tools and components, can be used to lower the cost of addressing the challenges of publishing and integrating Open Government Data (OGD). The evaluation of the tools used in the *National Statistical Office* use case workflow (see Fig. 1) will be given in section 2. Section 3 discusses the achieved results in the process of integration of Serbian public data in the LOD cloud, with a special attention to the case of one of the largest data providers in the Republic of Serbia – its Statistical Office (SORS).

1.1 LOD2: The Project and the OGD Use Case

In the last few years the Linked Data paradigm has evolved as a powerful enabler for the transition of the current document-oriented Web into a Web of interlinked Data and, ultimately, into the Semantic Web. Aimed at speeding up this process, the LOD2 project ("Creating knowledge out of interlinked data", http://lod2.eu) partners have delivered the LOD2 Stack, "an integrated collection of aligned state of the art software components that enable corporations, organizations and individuals to employ Linked Data technologies with minimal initial investments" [1].

One of the LOD2 objectives is to showcase the wide applicability of the LOD2 Stack for building public services for ordinary citizens of the European Union. As partners of the LOD2 project, the *Mihailo Pupin Institute*'s team established the Serbian CKAN,[3] the first catalogue of this kind in the West Balkan countries, with a goal of becoming an essential tool for enforcing business ventures based on open data in this region. The RDF datasets cataloged with the Serbian CKAN (rs.ckan.net) are periodically harvested and synchronized at an international level with the PublicData.eu portal[4] and integrated into the LOD cloud.

2 Evaluation of LOD Tools and Technologies

The LOD2 Stack was evaluated for allowing governments and governmental agencies to publish their data based on open standards. Requirements identified for the National Statistical Office scenario [2] were grouped into the following types: *Data extraction and transformation*, *Domain-specific modeling*, *Data enrichment and interlinking*, *Data storage*, *Exploration and analysis*, and *Data and Service administration*. Table 1 shows how the LOD2 Stack responds to these requirements.

Vocabularies suitable for modeling statistical data in RDF format are the Data Cube vocabulary [3] which is fully compatible with the cube model that underlines SDMX[5], and VoID (Vocabulary of Interlinked Datasets, http://www.w3.org/TR/void/), an RDF based schema used to describe linked datasets.

[3] CKAN is a data catalogue system used by various institutions and communities to manage open data.

[4] PublicData.eu has been developed as a part of the LOD2 project.

[5] SDMX (Statistical Data and Metadata eXchange), http://code.google.com/p/publishing-statistical-data/wiki/Documentation.

Table 1. Overview of LOD2 Stack capabilities

Data Extraction and Transformation
In a case where direct central database access is enabled, the *D2R server* and D2RQ mapping language can be used to represent the content in RDF format (e.g. using the SPARQL endpoint). Otherwise, for data provided in Excel or XML format, *OntoWiki*'s stat2RDF extension or the LOD2 *XSLT* processor can be used.
Domain-specific Modeling
The *PoolParty Thesaurus Manager* (PPT, http://lod2.poolparty.biz) tool for enterprise metadata management and linked data publishing is based on standard SKOS vocabulary and can be combined with text mining and linked data technologies. Additionally, knowledge models developed with PoolParty can be edited and enhanced with *OntoWiki* (http://ontowiki.net/) authoring tool.
Data Enrichment and Interlinking
These features are very important as a pre-processing step in integration and analysis of statistical data from multiple sources. The LOD2 tools such as *SILK* (http://www4.wiwiss.fu-berlin.de/bizer/silk) and *Limes* (http://aksw.org/Projects/LIMES) facilitate mapping between knowledge bases, while *GRefine* can be used to enrich the data with descriptions from DBpedia or reconcile with other information in the LOD cloud.
Data Storage
The LOD Cloud Cluster knowledge store for the LOD2 Project (http://lod.openlinksw.com) hosting 50 billion plus triples, consists of a *Virtuoso* clustered instance hosted on 8 server nodes at the *Sindice* Data Centre at DERI (NUIG)[4].
Exploration and Analysis
The LOD2 Stack offers tools such as SparQLed, *Sindice's assisted SPARQL editor* (http://sindicetech.com/sindice-suite/sparqled/) and the RDF Data Cube visualization component *CubeViz* (https://github.com/AKSW/cubeviz.ontowiki), that are of special importance for statistical data analysis and visualization.

3 Linked Open Data Example from Serbia

In an attempt to adopt the LOD2 Stack for the Statistical Office of the Republic of Serbia, over 100 datasets were extracted from the central statistics database (http://webrzs.stat.gov.rs/WebSite/public/ReportView.aspx), transformed into RDF, stored as RDF dump files on a local server (http://elpo.stat.gov.rs/lod2/) and registered with the Serbian CKAN. The data includes statistics from the *Prices*, *National accounts*, *Usage of Information and Communication Technologies,* and *Science*, *Technology and Innovation* domains (see [2] for more details). Performed activities can be summarized as follows.

Metadata Management. The statistics published by National Statistical Offices or Eurostat are organized by theme, presented in aggregate form by using a wide range of standard metadata (code lists). In the SORS Use case, a knowledge model was built where standard code lists were modeled using the SKOS vocabulary [2]. The model

(http://lod2.poolparty.biz/) currently incorporates 12 concept schemas including the *NACE (revision 1 and revision 2)*, *COICOP*, and *SITC (revision 4)*, as well as other schemas used in SORS statistical publications, such as *geographical*, *time* and *statistical areas* code lists. In order to formalize the conceptualization of the *National accounts* domain, for instance, the ESA 95 (European system of accounts ESA, http://circa.europa.eu/irc/dsis/nfaccount/info/data/ESA95/en/titelen.htm) was used. In governmental organizations, the metadata management activity is carried out by users with administration permissions (depicted in Fig.1). Using Silk and LODGefine (http://code.zemanta.com/sparkica/) some of the code lists were interlinked with DBpedia and Eurostat code lists.

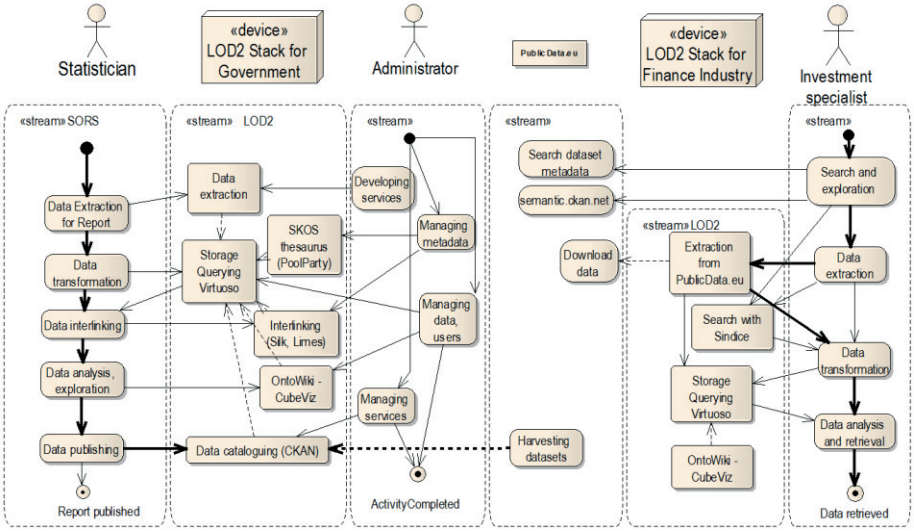

Fig. 1. Using LOD2 tools for publishing and consuming statistical data

The Serbian CKAN. The Serbian CKAN portal is deployed on a server with the following characteristics: Intel® Xeon® CPU 5140, dual core @ 2.33GHz 8GB RAM, Ubuntu 11.04, with kernel version: 2.6.38-12. The CKAN software was fully translated to Serbian, enabling support for two character sets (Latin and Cyrillic). Furthermore, a large number of dataset relationships have been defined, making the CKAN browsing and navigation experiences more comfortable. The Serbian CKAN is currently maintained by the *Mihailo Pupin Institute*'s team.

The SORS LOD Cloud. The SORS statistical data in XML form was passed as input to the XSLT processor and transformed into RDF using the aforementioned vocabularies (RDF Data Cube, SDMX-RDF, SKOS, Dublin Core Terms, VoID) and developed concept schemes. The VoID definition of the SORS LOD dataset is given in Fig.2. The SORS dataset (87.968 triples, see http://stats.lod2.eu/serbia) was also uploaded to the LOD Cloud Cluster knowledge store under the graph name http://elpo.stat.gov.rs/lod2/.

```
rzs:LOD
      rdf:type void:Dataset ;
      rdfs:label "Linked Open Data published by Statistical Office of the Republic of Serbia"@en ;
      dcterms:description "Linked Open Data published by Statistical Office of the Republic of Serbia" ;
      dcterms:modified "2012-04-24"^^xsd:date ;
      dcterms:source <http://webrzs.stat.gov.rs/WebSite/Public/PageView.aspx> ;
      dcterms:subject <http://purl.org/linked-data/sdmx/2009/subject> , <http://dbpedia.org/resource/Statistics> ;
      dcterms:title "SORS Linked Open Data" ;
      void:subset rzs:Prices , rzs:National_accounts , rzs:ICT_Usage , rzs:Science_technology_innovations .
```

Fig. 2. VoID description of the SORS LOD

4 Conclusion and Outlook

This paper contributes to the understanding of the LOD2 tools and technologies and discusses their use for publishing and consuming public sector information through the SORS Use case. The main lessons learnt from this study are:

- The Data Cube RDF vocabulary is mature enough to be used for publishing statistical data as it improves interoperability and allows comparison of data from different statistical sources.
- The LOD2 Stack provides a wide range of data transformation, enrichment and exploitation tools. However, advanced tools for analysis and visualization of statistical data are still under development.
- For publishers who currently only offer static files, Linked Data offers a flexible, non-proprietary, machine-readable means of publication that supports an out-of-the-box web API for programmatic access.
- The Serbian CKAN increases the visibility and accessibility of Serbian public sector data

We conclude that adoption of LOD2 tools and technologies leads to establishment of an interoperable Open Government Data ecosystem. Future work will include an analysis of the LOD2 Stack components for building custom applications for different LOD stakeholders.

Acknowledgements. The research presented in this paper is partly financed by the European Union (FP7 LOD2 project, Pr. No: 257943), and partly by the Ministry of Science and Technological Development of Republic of Serbia (SOFIA project, Pr. No: TR-32010). The Linked Open Data example was realized through close cooperation with the Statistical Office of the Republic of Serbia.

References

1. Auer, S., Martin, M., Frischmuth, P., Deblieck, B.: Facilitation the publication of Open Governmental Data with the LOD2 Stack. Share-PSI workshop, Brussels. Retrieved from http://share-psi.eu/papers/LOD2.pdf (2011)
2. Vraneš, S., Janev, V., Spasić, M., Milošević, U.: Establishment of the Serbian CKAN. LOD2 Deliverable 9.5.1, Institute Mihajlo Pupin (2012).
3. Cyganiak R., Reynolds D., Tennison J.: The RDF Data Cube vocabulary (July 14. 2010).
4. Williams, H., Boncz, P., Tummarello, G., Auer, S.: 50 Billion plus Triple LOD Cloud Hosted on the LOD2 Knowledge Store Cluster. LOD2 Deliverable 2.1.3 (2012).