LOD2 Tool for Validating RDF Data Cube Models

Valentina Janev, Vuk Mijović, Sanja Vraneš

Mihailo Pupin Institute, University of Belgrade, Belgrade, Serbia
{valentina.janev, vuk.mijovic, sanja.vranes@pupin.rs}

Abstract. The Open Government Data initiative aims at motivating governments and organizations to make information freely available and easily accessible online. This paper contributes to the understanding of the process of publishing public sector information as Linked Data, the use of the RDF Data Cube vocabulary and the LOD2 stack for publishing statistical data. In order to facilitate the publication process, a specialized component for the LOD2 Statistical workbench has been implemented – the RDF Data Cube Validation tool, (described in this paper). The tool can speed-up the processing and publishing Linked Data in RDF Data Cube format.

Keywords. Linked open data, government data, RDF Data Cube, tools, public sector, LOD2

1 Introduction

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. The term Linked Data here refers to a set of best practices for publishing and connecting structured data on the Web. These best practices have been adopted by an increasing number of data providers over the past three years, leading to the creation of a global data space that contains many billions of assertions - the Linked Open Data cloud (http://lod-cloud.net).

Although in the past governments have been protective over the data they collect, mentioning national security and citizen privacy as main reasons, global Open Government Data (OGD) initiatives, such as the Open Government Partnership (http://www.opengovpartnership.org/), have helped lower the barriers and open up governmental data for the public, by insisting on non-sensitive information, such as core public data on transport, education, infrastructure, health, environment, etc. Opening up data provides citizens with easier access to services, greater transparency and understanding of services, and improved communication through feedback loops, which immediately results in greater understanding and insight for the planning and delivery of community resources and societal support.

To make data truly open (for use and re-use), and increase transparency, it needs to be published in a non-proprietary, machine-readable format. The Government Linked Data (GLD, http://www.w3.org/2011/gld/) Working Group aims at providing stan-

dards and other information which help governments around the world publish their data as effective and usable Linked Data using Semantic Web technologies. Some of the vocabularies specified so far are RDF Data Cube vocabulary (http://www.w3.org/TR/vocab-data-cube/), Data Catalog Vocabulary (http://www.w3.org/TR/vocab-dcat/), an organization vocabulary (http://www.w3.org/TR/vocab-org/), registered organization vocabulary (http://www.w3.org/TR/vocab-regorg/), terms for describing people (http://www.w3.org/TR/vocab-people/), etc.

This article discusses the implementation of a tool for validating RDF Data Cube models. The tool is a part of the LOD2 Statistical WorkBench, the specialized version of the LOD2 Stack (http://stack.lod2.eu/) aimed at enabling corporations, government organizations and individuals to employ Linked Data technologies in the statistical domain. The LOD2 stack is an integrated collection of aligned state of the art software components delivered within the LOD2 project.

The article is organized as follows. First, Section 2 conveys some basic facts about the use of the RDF Data Cube model for representing multidimensional data. Then, Section 3 presents the RDF Data Cube Validation service using representative data published by the Statistical Office of the Republic of Serbia. Finally, Section 4 summarizes the main lessons learnt from the LOD approach and outlines the future work.

2 RDF Data Cube Model

2.1 Linked Data

Linked Data standards and technologies are a part of the Semantic Technologies that focus on meanings, connecting knowledge, and putting everything to work in ways that enable computers and people to cooperate better. Since the conception of the Sir Tim Berners-Lee's vision of the Web of Linked Data [1], the W3C Semantic Web Activity Group has accepted numerous Web technologies as standards or recommendations for building *Linked Data* applications. Tools that have been implemented so far around the *Linked Data* paradigm are named differently: graph databases, semantic annotation tools, named entity recognition and extraction tools, link discovery frameworks, linkage validation tools, indexing and search engines, rule-based engines, etc.

Linked Data approach enables datasets to be linked together through references to common concepts. A dataset is represented in the form of a graph, using the Resource Description Framework (RDF) as a general-purpose language. To ensure interoperability between applications that exchange machine-understandable information, RDF describes information in terms of objects ("resources") and the relations between them via the RDF Schema, which serves as a meta-language or vocabulary to define properties and classes of RDF resources. To avoid ambiguity, the RDF Schema uses uniform resource identifier (URI) references for naming. URI reference is a string that represents, for instance, name or address of an abstract or physical resource on the Web. HTTP URI identifiers are not a requirement, but rather a (recommended) design choice, which is employed in most statistical datasets.

Recently, several projects have been financed within the EU FP7 research program devoted to

- publishing data in Linked Data format, maintaining data catalogs, development and maintaining of open source toolkits that cover all stages of the Linked Data publication and consumption process e.g. projects LOD2 (http://lod2.eu), LATC (http://latc-project.eu);
- publishing and maintaining Linked geo-spatial data, e.g. TELEIOS (http://www.earthobservatory.eu/), PlanetData (http://planet-data.eu), GeoKnow (http://geoknow.eu);
- facilitating professional training for data practitioners, who aim to use Linked Data in their daily work e.g. project EUCLID (http://euclid-project.eu).

In this paper we will present some results achieved in the LOD2 framework.

2.2 Statistical Data as Linked Data

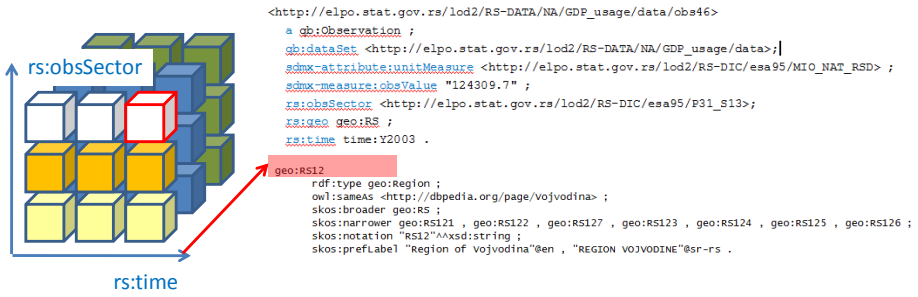

Figure 1. Statistical data as Linked Data

The Data Cube RDF vocabulary (http://www.w3.org/TR/vocab-data-cube/) is a model (namespace `http://purl.org/linked-data/cube#`, prefix `qb`) published by the Government Linked Data Working Group focused purely on the publication of multi-dimensional data on the Web. The model builds upon the core of the SDMX 2.0 Information Model [2]. The SDMX standards are now being widely adopted around the world for the collection, exchange, processing, and dissemination of aggregate statistics by official statistical organizations. As the cube model is very general, Data Cube can also be used for other data sets (e.g. survey data, spreadsheets and OLAP data cubes).

2.3 Example

A statistical data set (see Figure 1) comprises a collection of observations made at some points across some logical space. A resource representing the entire data set is created and typed as `qb:DataSet` and linked to the corresponding data structure definition via the `qb:structure` property.

```
@prefix rs: <http://elpo.stat.gov.rs/lod2/RS-DIC/rs/> .
@prefix accounts: <http://elpo.stat.gov.rs/lod2/RS-DATA/NA/dsd/>.
```

```
<http://elpo.stat.gov.rs/lod2/RS-DATA/NA/GDP_usage_Exports/data>
   a qb:DataSet ;
   rdfs:label "GDP usage - Exports"^^xsd:string ;
   rdfs:comment "Source: RZS (http://www.stat.gov.rs/)" ;
   qb:structure accounts:GDP_usage_Exports;
   dcterms:subject <http://purl.org/linked-data/sdmx/2009/subject/2.2>;
   dc:publisher "Stat. Office of the Republic of Serbia"^^xsd:string .
```

The collection can be characterized by a set of dimensions that define what the observation applies to (e.g. time, obsSector, country) along with metadata describing what has been measured (e.g. economic activity, prices), how it was measured and how the observations are expressed (e.g. units, multipliers, status).

```
@prefix time: <http://elpo.stat.gov.rs/lod2/RS-DIC/time/> .
@prefix geo: <http://elpo.stat.gov.rs/lod2/RS-DIC/geo/> .
<http://elpo.stat.gov.rs/lod2/RS-DATA/NA/GDP_usage/data/obs46>
  a qb:Observation ;
  qb:dataSet <http://elpo.stat.gov.rs/lod2/RS-DATA/NA/GDP_usage/data>;
  sdmx-attribute:unitMeasure <http://elpo.stat.gov.rs/lod2/RS-
  DIC/esa95/MIO_NAT_RSD> ;
  sdmx-measure:obsValue "124309.7" ;
  rs:obsSector <http://elpo.stat.gov.rs/lod2/RS-DIC/esa95/P31_S13>;
  rs:geo geo:RS ;
  rs:time time:Y2003 .
```

We can think of the statistical data set as a multi-dimensional space, or hyper-cube, indexed by those dimensions. This space is commonly referred to as a *cube* for short; though the name shouldn't be taken literally, it is not meant to imply that there are exactly three dimensions (there can be more or fewer) nor that all the dimensions are somehow similar in size.

3 Validation Service in Use

3.1 The LOD2 Statistical WorkBench

The LOD2 project[1] is a European FP7 initiative that aims to improve coherence and quality of data published on the Web, close the performance gap between relational and RDF data management, establish trust on the Linked Data Web and generally lower the entrance barrier for data publishers and users. 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. Therefore, a specialized version of the stack (Statistical WorkBench, see http://fraunhofer2.imp.bg.ac.rs/lod2demotest/stat) was developed to support the need of experts that work with statistical data (see Figure 2).

[1] LOD2 -"Creating knowledge out of interlinked data", http://lod2.eu.

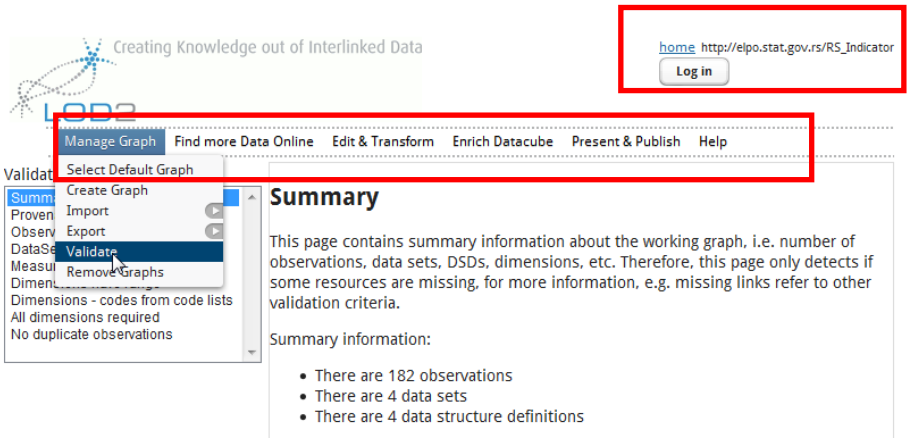

Figure 2. LOD2 Statistical WorkBench – main menu

3.2 Validating RDF Data cubes

Linked Data publication process refers to a set of activities related to extraction, transformation, validation, exploration and publication of RDF datasets originating from different sources (e.g., databases) on the Web. The ready for use RDF datasets can be either uploaded in the cloud [3] or registered by a portal (e.g. the Serbian CKAN [4]). The 2nd release of the LOD2 stack (announced in November 2012) offered the CSV2RDF tool for extracting RDF Data cubes from CSV data and the *CubeViz* tool for exploring and visualization RDF Data cubes.

Taking into consideration that future users could choose some other ways to prepare the RDF Data Cubes for publication (e.g. extract the data from XML/RDBMS using custom transformation), we came across the problem of integrity check validation for RDF Data Cubes prior to using the *CubeViz* component or publishing the datasets to a public portal. The use of the Validation service in this process is illustrated in Figure 3. The new component was developed by the PUPIN team, a partner in the LOD2 project, and contributed to the LOD2 stack.

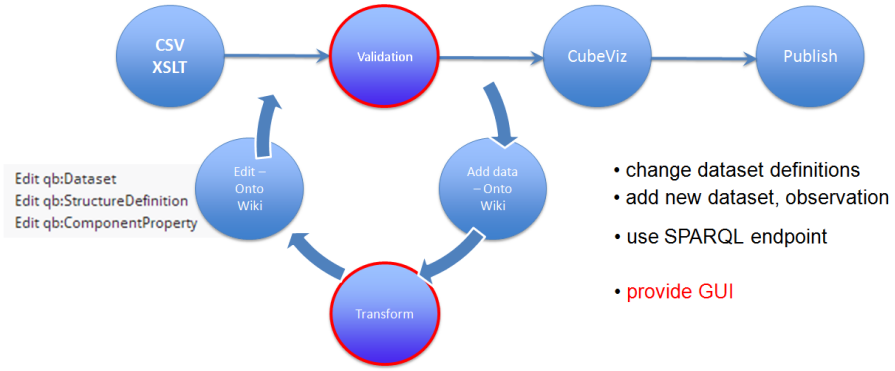

Figure 3. LOD2 publishing process

3.3 Tool GUI

Validation component checks if the supplied graph is valid according to the integrity constraints defined in the RDF Data Cube specification (http://www.w3.org/TR/vocab-data-cube/, version at the time of writing this paper is available at http://www.w3.org/TR/2013/WD-vocab-data-cube-20130312/). Each constraint in the document is expressed as narrative prose, and where possible, SPARQL ASK queries are provided. These queries return *true* if the graph contains one or more Data Cube instances which violate the corresponding constraint.

Our component shows a list of criteria corresponding to integrity constraints on the left side, while the details about the chosen criteria are shown on the right. Details about the chosen criteria include a list of resources which violate the constraint, an explanation about the problem, and if possible, a quick solution to the problem. Figure 4 shows a case where a user checks if observations are correctly linked to data sets. The list of observations violating this constraint is shown on the right along with properties/details of that observation and a proposed solution to the problem. The user can then choose to apply the proposed solution or to manually edit the resource in question.

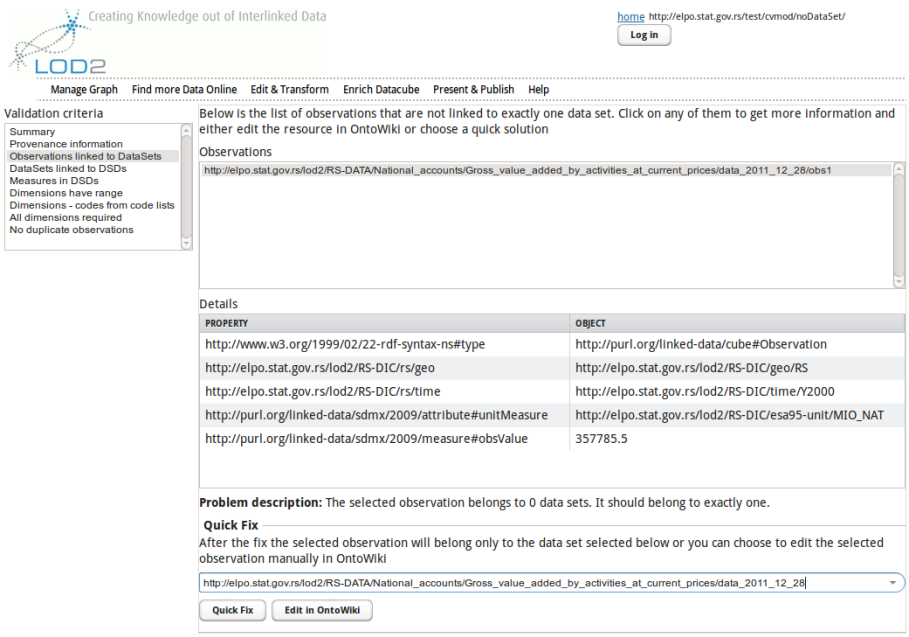

Figure 4 Validation component

In order to acquire the details that are shown on the right side, ASK queries defined in the RDF Data Cube specification were slightly modified and turned into

SELECT queries that extract which resources are problematic along with some additional information if needed. All graphs handled by the Statistical WorkBench are stored in open-source edition of Virtuoso universal server (http://www.openlinksw.com/wiki/main/), which includes a scalable high-performance RDF Quad Store, while SPARQL queries were performed by using the Sesame framework (http://www.openrdf.org/doc/sesame2/system/). The user interface is implemented in Vaadin (https://vaadin.com/home), a Java web application framework for Rich Internet Application and the application is deployed on Apache Tomcat. Finally, Statistical WorkBench, as well as all other components in the LOD2 stack is available as a Debian package for the Ubuntu Linux distribution.

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 [4] for more details).

4 Conclusion and Outlook

This paper contributes to the understanding of the RDF Data Cube vocabulary, the specialized LOD2 RDF Data Cube Validation service and the process of publishing Linked Data. 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. The missing link (needed to connect the transformation and visualization steps in the statistical data processing) has been resolved by an implementation of a specialized component for the LOD2 Statistical workbench.
- For publishers who currently offer only static files, Linked Data offers a flexible, non-proprietary, machine readable means of publication.

We conclude that LOD2 tools and technologies yield to establishment of interoperable Open Government Data ecosystem. The benefits of Open Government Data are economic, through the identification of new business opportunities, and social, through increased transparency, participation and accountability.

Work in progress includes further enhancement of the Validation tool toward a Reasoning service that will enable automatic repair of observed/identified errors that will speed-up the processing and publishing Linked Data in RDF Data Cube format.

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).

References

1. Berners-Lee, T., Hendler, J., & Lassila, O. The Semantic Web. Scientific American, May **2001**. http://*www.sciam.com/article.cfm?id=the-semantic-web.*
2. SDMX Information Model: UML Conceptual Design (Version 2.0), November 2005, Statistical Data and Metadata Exchange Initiative. URL: http://sdmx.org/docs/2_0/SDMX_2_0%20SECTION_02_InformationModel.pdf
3. Janev, V., Milošević, U., Spasić, M., Vraneš, S., Milojković, J., Jireček, B. "Integrating Serbian Public Data into the LOD Cloud", In Z. Budimac, M. Ivanović, M. Radovanović (Eds.) 5th Balkan Conference in Informatics (BCI'12, September 16–20, 2012, Novi Sad, Serbia), pp.94-99. New York: ACM International Conference Proceeding Series VOL. 641.
4. Vraneš, S., Janev, V., Spasić, M., Milošević, U.: Establishment of the Serbian CKAN. LOD2 Deliverable 9.5.1, Institute Mihajlo Pupin (2012). Retrieved from http://static.lod2.eu/Deliverables/LOD2_D9.5.1_Serbian_CKAN.pdf