Web Semantics: Science, Services and Agents on the World Wide Web 41 (2016) 1–8

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

# Exploratory spatio-temporal analysis of linked statistical data

Vuk Mijović [a], Valentina Janev [b,*], Dejan Paunović [b], Sanja Vraneš [b]

[a] School of Electrical Engineering, University of Belgrade, Institute Mihajlo Pupin, Belgrade, Serbia
[b] University of Belgrade, Institute Mihajlo Pupin, Belgrade, Serbia

ARTICLE INFO

*Article history:*
Received 11 April 2016
Received in revised form
11 August 2016
Accepted 24 October 2016
Available online 1 November 2016

*Keywords:*
Linked data
Statistics
Spatio-temporal
Exploration
Visualization
Interoperability

ABSTRACT

Publishing and sharing open government data in Linked Data format provides many opportunities in terms of data aggregation/integration and creation of information mashups. Statistical data, that contains various performance indicators and their evolution through time, is an example of data that can be used as the foundation for policy prediction, planning and adjustments, and can be re-used in different applications. However, due to Linked Data being relatively a new field, currently there is a lack of tools that enable efficient exploration and analysis of linked geospatial statistical datasets. Therefore, ESTA-LD (Exploratory Spatio-Temporal Analysis) tool was developed to address some of the Linked statistical Data management issues, such as crossing the statistical and the geographical dimensions, producing statistical maps, visualizing different measures, and comparing statistical indicators of different regions through time. This paper discusses the modeling approach that was adopted so that the published data conform to the established standards for representing statistical, spatial and temporal data in Linked Data format. The main contribution is related to the delivery of state-of-the-art open-source tools for retrieving, quality assessment, exploration and analysis of statistical Linked Data that is made available through a SPARQL endpoint.



## 1. Introduction

Statistical data is often used as the foundation for policy prediction, planning and adjustments, and therefore has a significant impact on the society (from citizens to businesses to governments). In the last few years, with the rise of the open data movement, a large and increasing number of governments and organizations have started to make information freely available and easily accessible online. In order to increase transparency, the information is also published as Linked Open Data [1].

From the government systems perspective, the Linked Data approach can be observed as a technique for making the data interoperable and ready for consumption. In order to harmonize approaches used for describing the datasets, semantic services or repositories, the European Commission, in collaboration with the W3C consortium, has accepted a set of standard vocabularies that should be used to build public administration services [2]. In the ISA programme framework, the European Commission supports the development of tools, services and frameworks in the area of e-Government through more than 40 actions.[1] Currently, the JOINUP repository is used for storing the descriptions of schemata used in the publicly available datasets, as well as services that enable access/retrieval of data.

In general, the wider adoption of standards for representing and querying semantic information, such as RDF(s) and SPARQL, along with increased functionalities and improved robustness of modern RDF stores, have established Linked Data and Semantic Web technologies in the areas of data and knowledge management. However, these technologies are still quite novel, and a lot of the tooling and standards are either missing, still in development, or not yet widely accepted. For example, the GeoSPARQL [3] standard that supports representing and querying geospatial data on the Semantic Web was published in June 2012, but the Spatial Data on the Web Working Group is still working on clarifying and formalizing the relevant standards landscape with respect to integrating spatial information with other data on the Web, discovering of different facts related to places, and identifying and assessing existing methods and tools in order to create a set of best practices. The RDF

* Corresponding author.
*E-mail addresses:* Vuk.Mijovic@pupin.rs (V. Mijović), Valentina.Janev@pupin.rs (V. Janev), Dejan.Paunovic@pupin.rs (D. Paunović), Sanja.Vranes@pupin.rs (S. Vraneš).

http://dx.doi.org/10.1016/j.websem.2016.10.002
1570-8268/

[1] http://ec.europa.eu/isa/ready-to-use-solutions/index_en.htm.

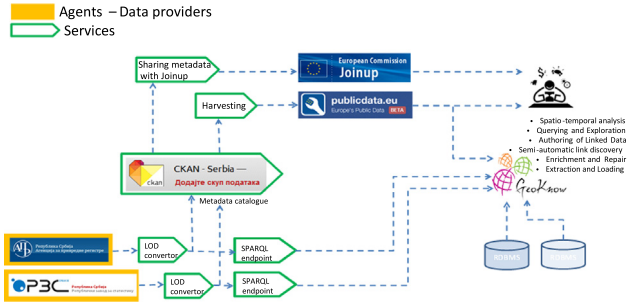


**Fig. 1.** Integrating public data from Serbia in EU data space.

Data Cube vocabulary (QB) [4] which enables modeling of statistical data as Linked Data is a W3C recommendation since January 2014, and it has been used in several pilot applications by public authorities [5], but representation of spatio-temporal concepts vary across the published datasets.

This paper describes several tools developed within the recent EU projects LOD2 [6] and GeoKnow [7] for managing Linked statistical Data. Section 2 describes the global context of sharing and reusing open data. The RDF Data Cube vocabulary, which is the basis for modeling statistical data, is discussed in Section 3, along with the challenges for managing spatial and temporal information in Linked Data format. The Linked Data tools for managing statistical data are briefly introduced in Section 4, while Section 5 gives examples of ESTA-LD use with data from a business case in Germany and a government use case in Serbia. Section 6 concludes the paper and gives an outlook on future work.

The work described in this paper builds upon and extends previous efforts elaborated in [8,9].

## 2. Motivation

### 2.1. Problem statement

In the recent years, various Open Government Data[2] initiatives, such as the Open Government Partnership,[3] have pushed governments to open up their data by insisting on opening non-sensitive information, such as core public data on transport, education, infrastructure, health, and environment. Consequently, the amount of the public sector information, which is mostly statistical in nature and often refers to different geographical regions and points in time, has increased significantly in the recent years, and this trend is very likely to continue.

In order to make the use of open data more efficient and less time-consuming, standardized approaches and tools are needed. For instance, in the process of integration of Open Data from Serbia in EU data space, the end user (consumer on the right of the Fig. 1) needs uniform solutions for accessing, describing, and re-using the data coming from different publishers, e.g. through a CKAN catalog. In order to share datasets between users and platforms, the datasets need to be accessible (regulated by license), discoverable (described with metadata) and retrievable (modeled and stored in a recognizable format). Therefore, the CKAN *data catalog* can be viewed as an electronic library index that structures descriptions (meta-data) about the actual data that facilitate discovery of public datasets.

Publishing data in Linked Data format involves different operations such as

- efficient transformation / conversion of traditional data stores (e.g. CSV, XML, relational databases) into linked, machine readable formats;
- building and querying triple stores containing RDF data cubes;
- validating RDF data cubes;
- interlinking and adding meaning to data;
- visualization and exploration of multi-dimensional RDF data cubes;
- publishing statistical data and respective metadata within a selected portal (i.e. a CKAN instance).

### 2.2. Related work

In the last three years, in the framework of the LOD2 and GeoKnow projects, the Institute Mihajlo Pupin (IMP) has been involved in maintaining the *Linked Data* stack,[4] an integrated set of tools for managing the Linked Data life-cycle. All tools are RDF based and enable developers to build custom applications on the top of the public sector data. In the LOD2 project, IMP was involved in the development of a specific web interface (*Statistical Workbench*) that aggregated several components of the stack, organized in an intuitive way to support the specific business context of a statistical office. The *Workbench* contains several dedicated extensions for manipulating RDF data according to the Data Cube vocabulary: validation, merging and slicing of cubes. The *Workbench* was tested with users from the Statistical Office of the Republic of Serbia[5] and was used to publish datasets from different statistical data in Linked Data format. These datasets are available as RDF dump files via the Serbian CKAN.[6] The *Workbench* also integrates CubeViz, a Data Cube visualization component [10], however this tool does not allow linking of the statistical indicators with polygons nor their visualization on a geographic map. In the GeoKnow framework, IMP worked on further extending the capabilities of the *GeoKnow Generator*,[7] an integrated solution for managing geospatial data, by providing the *RDF Data Cube Validation*[8] and *ESTA-LD*[9] components.

Apart from the tools available in the *Linked Data* stack, there are few attempts across Europe to work on specialized tools for managing Linked statistical Data. The *Sextant* tool [11] allows visualizing time-evolving linked geospatial data, but no computational/analysis options e.g. summarization across dimensions, have been reported. We stress that the visualization of the temporal dimension of geospatial data has not been captured by other semantic web tools due to lack of a standard temporal extension of RDF that practitioners could utilize when publishing RDF data. The *LSD* repository [12,13] investigates Semantic Similarities and Correlations of Linked Statistical Datasets, but this tool has no support for linking the results to a geographic map. Within the OpenCube project, the *OpenCube Toolkit*[10] has been released as open source software components that can be reused for building applications with their help of the open source Information Workbench *Information Workbench*[11] Community Edition platform.

The main question related to processing Linked Statistical Data raised during the LOD2 and GeoKnow projects were:

2 http://webfoundation.org/projects/ogd/.
3 http://www.opengovpartnership.org/.
4 http://stack.linkeddata.org/.
5 http://lod2.stat.gov.rs/lod2statworkbench.
6 http://rs.ckan.net/.
7 http://generator.geoknow.eu.
8 https://github.com/GeoKnow/DataCubeValidation.
9 https://github.com/GeoKnow/ESTA-LD.
10 http://opencube-toolkit.eu.
11 http://www.fluidops.com/information-workbench/.