ESTA-LD: enabling spatio-temporal analysis of linked statistical data

Vuk Mijović*, Valentina Janev**, Dejan Paunović**
* School of Electrical Engineering, University of Belgrade, Institute Mihailo Pupin, Belgrade, Serbia
** University of Belgrade, Institute Mihailo Pupin, Belgrade, Serbia
{Vuk.Mijovic, Valentina.Janev, Dejan.Paunovic}@pupin.rs

***Abstract*—In the recent years, Linked Data has become widely adopted and became established in the areas of data and knowledge management. Furthermore, various open government initiatives contributed to the availability of governmental data which is largely statistical in nature and often refers to different geographical regions and points in time. However, semantic technology has not influenced spatial data management yet. In this paper we discuss the possibilities for utilizing Linked Data in this domain and argue that it would facilitate integration of geospatial data with external datasets which is cumbersome in existing GIS systems. The paper addresses modelling of statistical linked data with the focus on representing spatial and time dimensions, and describes the current prototype of the Exploratory Spatio-Temporal Analysis tool for Linked Data developed by the Institute Mihailo Pupin within the GeoKnow framework.**

I. INTRODUCTION

With the wider adoption of standards for representing and querying semantic information, such as RDF(s) and SPARQL, Semantic Web technologies gained traction in the recent years and became established in the areas of data and knowledge management [1]. This process was also supported by the advances of RDF stores which have become more robust and now offer functionalities that are similar and comparable to those of traditional databases.

Geospatial data makes for a large portion of available knowledge bases and presents a highly valuable source of information for variety of applications. It can be loaded into GIS systems, however it is quite cumbersome to integrate external datasets into them and thereby leverage additional knowledge that is available. Open Geospatial Consortium enables and provides a way to share, reuse and integrate data between different GIS systems, but this data is still isolated in the GIS realm and therefore disconnected from the Web of Data. In order to tackle this issue, the EU FP7 project GeoKnow aims to make geospatial data a first-class citizen of the Web of Data and provide tools that would enable easy integration of variety of data sources and enable decision making powered by this data.

On the other hand, in the recent years global Open Government Data (OGD) initiatives, such as the Open Government Partnership[1], have helped to open up governmental data for the public, by insisting on opening non-sensitive information, such as core public data on transport, education, infrastructure, health, environment, etc. The vision for ICT-driven public sector innovation [2] refers to the use of technologies for the creation and implementation of new and improved processes, products, services and methods of delivery in the public sector. These efforts contributed to the availability of large volumes of public sector information which is mostly statistical in nature and often refers to different geographical regions and points in time.

ESTA-LD (Exploratory Spatio-Temporal Analysis of Linked Data) is a tool developed within The GeoKnow projects which aims to enable exploration and analysis of spatio-temporal linked statistical data, and demonstrate that by publishing statistics as Linked Data, it is easier to integrate external data, compare different datasets, and link to non-tabular data. This tool will also demonstrate usefulness of other tools developed within the project which will be leveraged to extract and transform data from other formats, interlink and integrate it with external datasets, and finally validate and improve its quality.

Main basis of the tool is the RDF Data Cube vocabulary, a W3C recommendation for modeling statistical data as Linked Data. The vocabulary and modeling of spatial and time dimension will be discussed in Section 2. GeoKnow project and the role of ESTA-LD will be described in Section 3, while the tool itself will be elaborated in Section 4. Finally, we will provide conclusions and give insights about the future work in Section 5.

The work described in this paper builds upon and extends previous efforts elaborated in [3].

II. MODELING SPATIO-TEMPORAL DATA

In this section we will introduce the RDF Data Cube vocabulary and how it can be used to model statistical data. Afterwards, we will go further into details and discuss how to model spatial and time dimensions in a way that will denote the role of these dimensions and enable exploration and analysis across space and time.

A. RDF Data Cube vocabulary

In January 2014, W3C recommended the *RDF Data Cube* vocabulary [4] as a standard vocabulary for modelling statistical data. The vocabulary focuses purely on the publication of multi-dimensional data on the Web. It builds upon the core of the *SDMX 2.0 Information Model* [SDMX] which is the result of the Statistical Data and Metadata Exchange (SDMX[2]) Initiative established in 2001 by seven international organizations (BIS, ECB,

[1] http://www.opengovpartnership.org/

[2] http://www.sdmx.org/

Eurostat, IMF, OECD, World Bank and the UN) with the aim to introduce and support greater efficiencies in statistical practice.

The vocabulary sees a statistical data set as a collection of observations made at some points across some logical space. The collection can be characterized by a set of dimensions that define what the observation applies to (e.g. time, country) along with metadata describing what is measured (e.g. economic activity, prices), how it is measured and how the observations are expressed (e.g. units, multipliers, status). Therefore, a statistical data set can be seen as a multi-dimensional space, or hyper-cube, indexed by those dimensions. Consequently, the vocabulary refers to statistical datasets as data cubes though this name shouldn't be taken literally since 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. Explicit definition of the cube's structure is represented by a Data Structure Definition (DSD) that enables validation, visualization, discovery, and abbreviation. DSD consists of a set of dimensions, attributes and measures, where dimensions serve to identify the observations, measures are used to describe phenomena being observed, while attributes allow to qualify and interpret the observed values. The vocabulary also allows to group subsets of observations together by creating slices through the cube in which one or more dimension values are fixed. In this case, explicit structure of a slice is given by associating it with an appropriate slice key, much like DSDs are used to describe structure of datasets.

B. Modelling hierarchucal data

In order to formalize the conceptualization of hierarchical dimensions we can use the Simple Knowledge Organization System (SKOS)[3]. SKOS Core is a model and an RDF vocabulary for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, 'folksonomies', other types of controlled vocabulary, and also concept schemes embedded in glossaries and terminologies. Concepts represented as skos:Concept are grouped into concept schemes (skos:ConceptScheme) that serve as code lists from which the dataset dimensions draw on their values. Semantic relation used to link a concept to a concept scheme is skos:inScheme. Herein, we will present an example of coding a geographical dimension in RDF.

```
geo:RS21 rdf:type geo:Region ;
  owl:sameAs
     <http://dbpedia.org/page/%C5%A0umadija_
and_Western_Serbia> ;
  skos:broader geo:RS ;
  skos:narrower geo:RS212, geo:RS216,
geo:RS211, geo:RS215, geo:RS213, geo:RS218 ,
geo:RS214 ,geo:RS217 ;
  skos:notation "RS21"^^xsd:string ;
  skos:prefLabel "Region of Sumadija and
Western Serbia"@en , "REGION ŠUMADIJE I
ZAPADNE SRBIJE"@sr-rs .
```

C. Modelling spatial and time dimensions

As part of the content oriented guidelines (COG), SDMX standard defines a set of common statistical concepts and associated code lists which are meant to be reused across different datasets. These guidelines are also available in RDF, as part of an effort by the community group[4]. Resources defined therein are not a part of the Data Cube specification, however they are used by a number of existing Data Cube publications and represent a solid foundation for modeling new dimensions. Among the provided concepts are dimensions sdmx-dimension:refPeriod and sdmx-dimension:refArea which may be used as a basis for defining time and spatial dimensions respectively. For example, dimension from the COG can be used to derive a new time dimension in the following way:

```
eg:refPeriod a rdf:Property,
qb:DimensionProperty;
  rdfs:label "reference period"@en;
  rdfs:subPropertyOf sdmx-
dimension:refPeriod ;
  rdfs:range interval:Interval;
  qb:concept sdmx-concept:refPeriod .
```

Furthermore, it is convenient to be able to easily identify which dimension is the time dimension, which component represents spatial dimension, which is a primary measure and so forth. To enable this, the vocabulary allows to denote which role a dimension plays within the structure definition. Roles are encoded as subclasses of skos:Concept and associated with dimensions through the qb:concept property. In the above example eg:refPeriod is linked to the concept it represents through the qb:concept property. This concept is of type sdmx:TimeRole, thereby denoting that this dimension represents a time dimension within the structure definition:

```
sdmx-concept:refPeriod a sdmx:TimeRole.
```

Another issue to consider is the range of these two dimensions, i.e. how to encode the dimension values. One way of representing time would be to define a code list. The problem with this approach is that it is too specialized and would limit the usability. Namely, if any tool was to be used to process this data, it would need to be aware of that particular code list in order to be able to interpret the codes. However, this problem can be overcome easily since there are two standard ways of representing the time. One solution is to use the OWL time ontology which is at the moment W3C working draft, and the other is to use xsd types such as xsd:gYear, xsd:gYearMonth, xsd:date, etc. On the other hand, the most common way to refer to geographic entities is to use resources defined in the GeoNames[5] database or link to them. Since almost every geographical dataset links to GeoNames this would enable to easily acquire any additional information that is needed. For example, if in our statistical dataset the countries were

[3] http://www.w3.org/TR/2005/WD-swbp-skos-core-spec-20051102/

[4] https://code.google.com/p/publishing-statistical-data/

[5] http://www.geonames.org/

represented with, or linked to GeoNames resources, it would be possible to acquire a polygon for any country with a simple query against the LinkedGeoData endpoint, which is like many other geographical datasets linked to GeoNames.

III. GeoKnow Generator and Linked Data Stack

GeoKnow is an EU research project that was motivated by previous work on LinkedGeoData project. Within LinkedGeoData information from OpenStreetMap was made available in RDF and interlinked with GeoNames, DBpedia, and other data sources, while GeoKnow aims to complement these efforts by making geospatial data more easily accessible on the web and improving publishing, querying, interlinking and quality assessment of geospatial information that is based on Linked Data principles.

The goal of GeoKnow is to support all stages in the Linked Data lifecycle: storage, authoring, interlinking, classification/enrichment, quality assessment, evolution/repair and searching/browsing/exploration. To achieve this goal, it includes many tools, some of which existed prior to GeoKnow and are now being improved and/or extended, while some are being developed during the course of the project (which is the case with ESTA-LD). When used together, these tools ensure better quality, and more information, thus leading to better visualizations and greater possibilities. For example, let's look at one example from the perspective of ESTA-LD. This collection of tools can be used to extract and transform data from different formats to RDF, then link it to other datasets such as GeoNames, and finally validate and repair if needed, thus leading to more data being available for analysis and ensuring high quality of this data. Furthermore, as described in the previous section, links to GeoNames can be used by the tool to acquire polygons and visualize data on a map.

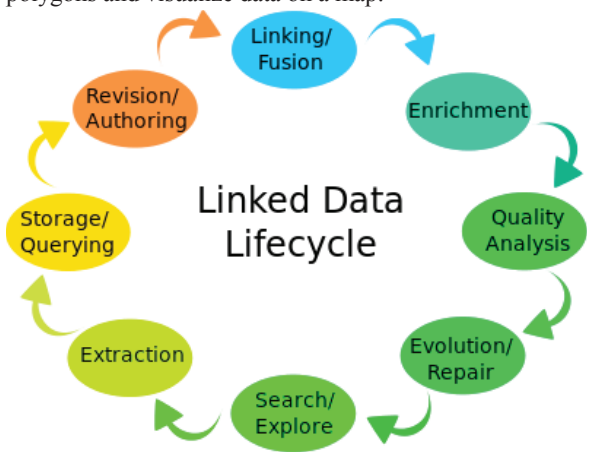

Figure. 1 Linked Data Lifecycle

Tools that are maintained and developed within GeoKnow are available as Debian packages and included in the Linked Data Stack which is a repository of Debian packages that targets Ubuntu 14.04 LTS operating system. This approach eases distribution and installation of all tools and components since the desired pieces of software can be installed with a single command without the need to deal with dependencies or configuration. The stack also includes GeoKnow Generator Workbench that integrates some of the components from the stack and provides various management functionalities such as access control and authorization, provenance, user management, or data source management, thus acting as an integrated environment that supports complete lifecycle of geospatial linked open data. We are currently in the process of creating a Debian package for ESTA-LD and making it a part of the Linked Data Stack, after which it will be integrated in the GeoKnow Generator Workbench.

IV. ESTA-LD

ESTA-LD is a tool that enables exploration and analysis of spatio-temporal linked statistical data (see Fig. 2) that is being developed within GeoKnow projects. The prototype can work on any SPARQL endpoint containing statistical data modeled with the RDF Data Cube vocabulary. It enables the user to select up to two arbitrary indicators for analysis First, it queries the endpoint for available graphs containing Data Cubes which are shown in the drop-down list on the left and when the user selects a graph, drop-down list on the right becomes populated with datasets contained in the chosen graph. Upon the selection of the dataset, its structure is analyzed and the user interface is updated accordingly. All dimensions are listed on the right side. For each dimension there is a toggle button which is used to select if the particular dimension is to be analyzed/visualized, and a drop-down list that is used to fix the dimension to a particular value.

Geographical dimension is visualized on the left side on the choropleth map. In this case, the prototype fires a query where all other dimensions are fixed to values selected in the drop-down lists on the right. This results in a set of observations for each geographical entity. Finally, the results are visualized on the map where regions for which the observed measurement is higher are depicted with a darker color then regions for which the observed measurement is lower. The map is also used to fix the geographic dimension to a particular value. Currently, the tool works with a custom defined code list of geographic areas, while in the future it will support any geographic entities linked to GeoNames.

All other dimensions are visualized on the chart positioned on the right side where up to two dimensions can be visualized/analyzed. The chart is refreshed every time a user changes the selection of dimensions to be analyzed or fixes any of the dimensions to a different value. Upon any of the mentioned changes, a query is executed in order to acquire all observations matching the selected criteria. If the user choses to visualize a single dimension, bar chart is used, while in the case of two dimensions, the component uses a histogram.

In case only a single dimension is selected for visualization and that dimension is a time dimension, a special kind of bar chart is used. This is actually a bar chart with additional controls that make it possible to select a period in time that will be shown on the chart. One of these controls is a ribbon placed below the chart. This ribbon is used to select the size and position of the time window that will be visualized. In this way it is possible to precisely set the period in time which is to be visualized, where size of the window determines duration, and its position determines the starting point. There are also pre-defined windows for periods of 1 month, 3 months, 6 months and 1 year which can be selected by clicking the dedicated buttons above the chart. At the moment, the tool is based on a custom defined code list

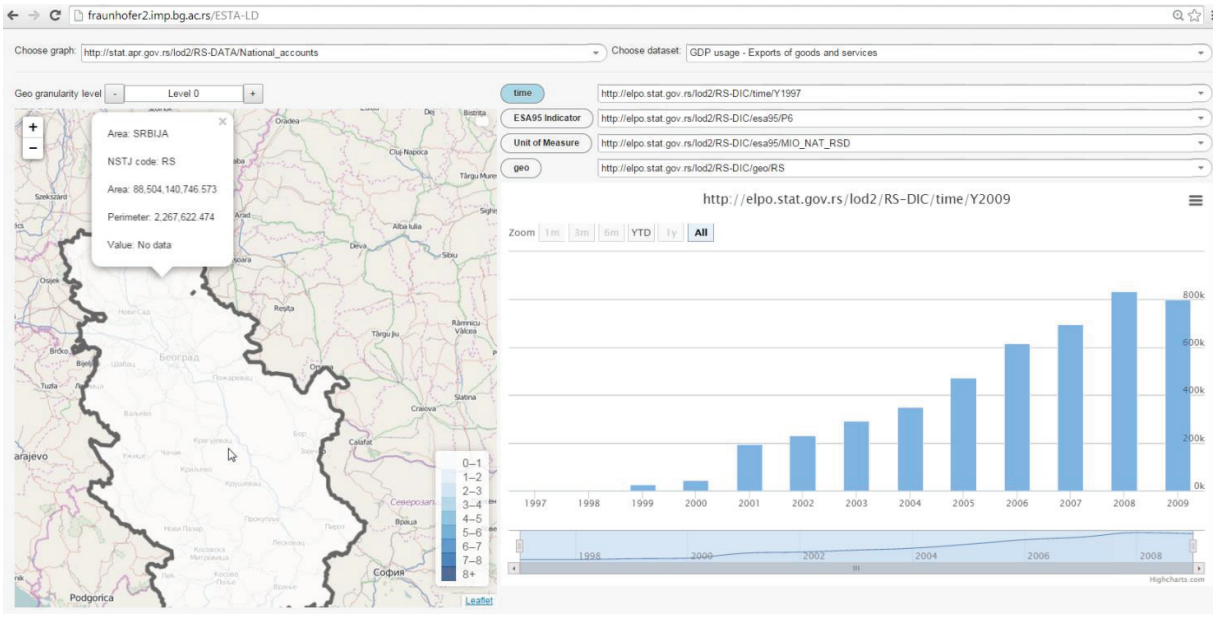

Figure 2. ESTA-LD prototype

for providing yearly and monthly data, while OWL Time and xsd types will be supported in the next version.

A. *Implementation*

The first prototype was developed in JavaScript and HTML5 which enabled early evaluation and testing of visualization components, while still ensuring easy integration in the *GeoKnow Generator*. Representation and interaction with geographic information were implemented using *Leaflet*[6], an open source JavaScript library for mobile-friendly interactive maps. Geographic data (such as region borders), originally available as shape files, was transformed and stored in GeoJSON format as required by Leaflet. This data is then modified using JavaScript and added to maps to create interactive visualizations. On the other hand, different statistical indicators, which are the subjects of the spatio-temporal analysis, are stored in the RDF Data Store which is queried using SPARQL query language. The actual retrieval of data from the SPARQL Endpoint was implemented using the jQuery library and its standard getJSON function. Finally, the results of the spatio-temporal analysis are visualized using *Highcharts*[7], a charting library written in pure HTML5/JavaScript, offering intuitive, interactive charts to a web site or web application.

Later on, ESTA-LD was generalized to enable the selection of indicators for analysis. In order to reuse existing Java module for querying RDF Data Cubes, new version was implemented using Vaadin[8], a Java framework for building modern web applications which also allowed for easy integration of existing functionalities. For querying the SPARQL endpoint we rely on the Sesame framework, while the user chooses the graph, dataset, and indicators to be analyzed using various Vaadin components (see Fig. 3). Since the *GeoKnow Generator* is a JavaScript web application which uses Java web servlets for the integration of Java components, and Virtuoso as an RDF store, this approach ensures straightforward integration of ESTA-LD. User interface can be easily integrated as HTML Iframe and parameters such as endpoint and initial graph to be analyzed can be specified as HTTP parameters, while the interaction and exchange of data with other components is achievable through the use of common RDF store.

B. *Evaluation*

Statistical data are often used as foundations for policy prediction, planning and adjustments, having a significant impact on the society (from citizens to businesses to governments) [5]. One such example is the Serbian Register of the Regional Development Measures and Incentives which is a unique, centralized electronic database of the taken measures and implemented incentives that are of significance for regional development. This register is essential for making new policies where various indicators need to be taken into

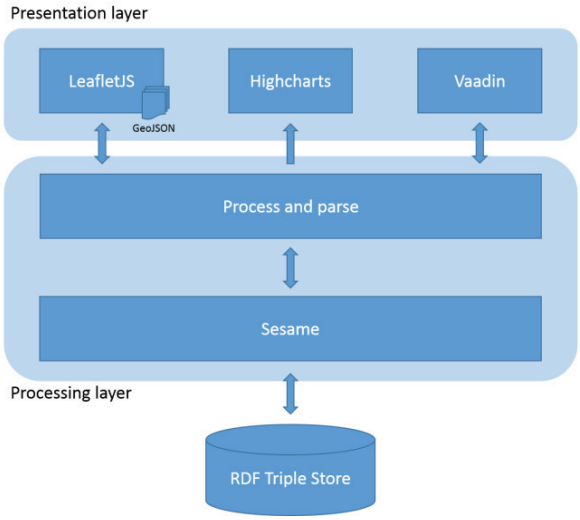

Figure 3. ESTA-LD Architecture

[6] http://leafletjs.com/

[7] http://www.highcharts.com/

[8] https://vaadin.com/home

account. In this process Linked Data technologies can be utilized to ensure interoperability and integration of data from other datasets, be it general purpose datasets such as GeoNames, or other statistical datasets such as a Register containing data from the Dissemination database of the Statistical Office of the Republic of Serbia (SORS). Data in these two registers contains both geographical and time dimensions, thus imposing challenges for analysis across geographic regions and different periods in time.

Statistics about regional development shows government investments for various purposes such as tourism, education and so forth. For each purpose, and each year, investments are captured on country, regional, and municipality level, thus allowing to evaluate the tool's ability to enable visualization and analysis across space and time where geographical data is captured on multiple levels of hierarchy. On the other hand, tourism data acquired from SORS captures different tourism indicators such as overnight stays on a monthly basis. This data allowed us to test and evaluate possibilities for analyzing different period in time. In the future, we will evaluate the benefits of analyzing data integrated from multiple sources and aggregating observations across space and time.

V. Conclusions and Future Work

This paper presented ESTA-LD, tool for exploratory spatio-temporal analysis of Linked Data. We demonstrated how statistical data can be modeled using the RDF Data Cube vocabulary and discussed different approaches to modeling spatial and time dimensions, followed by the discussion of ESTA-LD's place in the linked data lifecycle and relation to other tools for processing linked geospatial data.

Current prototype was described in detail and showed that ESTA-LD supports exploration and analysis of linked statistical data across space and time. Evaluation was primarily conducted using data published by Serbian governmental institutions that shows investments and progress of different economic indicators across geographic regions over time, thus catering to the main goal, which is to support policy makers by enabling them to utilize integration capabilities of Linked Data technologies and analyze different indicators over the integrated datasets. Currently, the tool is generic and allows exploration and analysis of arbitrary datasets and contained indicators. However, interpretation of spatial and time values is at the moment tied to custom defined code lists. Therefore, in external datasets that are not modeled with these code lists, utilizing visualization specifically tailored to spatial and time dimensions would require transformation. Consequently, the next step will be implementation of support for OWL Time and xsd types, while the generalization of the spatial dimension will be achieved by supporting geographic dimensions that contain links to GeoNames or use GeoNames resources directly. In the final stages we will take into consideration different aspects, such as scalability, flexibility and ease-of-use/friendliness.

Acknowledgment

The research presented in this paper is partly financed by the European Union (FP7 GeoKnow, Pr. No: 318159), and partly by the Ministry of Science and Technological Development of the Republic of Serbia (SOFIA project, Pr. No: TR-32010).

References

[1] J. Lehman, et al., "The GeoKnow Handbook", http://svn.aksw.org/projects/GeoKnow/Public/GeoKnow-Handbook.pdf, Accessed in December 2014.

[2] EC Digital Agenda, "Orientation paper: research and innovation at EU level under Horizon 2020 in support of ICT-driven public sector.",http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=2588, May 2013.

[3] D. Paunović, V. Janev, V. Mijović, "Exploratory Spatio-Temporal Analysis tool for Linked Data", *In Proceedings of 1st International Conference on Electrical, Electronic and Computing Engineering*, RTI1.2.1-6., June 2014, Vrnjačka Banja, Serbia.

[4] R. Cyganiak, D. Reynolds, J. Tennison, "The RDF Data Cube vocabulary", http://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/, January 2014.

[5] V. Janev, V. Mijović, D. Paunović, U. Milošević, "Modeling, Fusion and Exploration of Regional Statistics and Indicators with Linked Data tools", *In Proceedings of the Third International Conference, EGOVIS 2014, Lecture Notes in Computer Science,* vol. 8650, pp 208-221, September 2014, Munich, Germany.