Exploratory Spatio-Temporal Analysis tool for Linked Data

Dejan Paunović, Valentina Janev, Vuk Mijović

***Abstract*—Linked Data provides a publishing paradigm in which not only documents, but also data, can be a first class citizen of the Web, thereby enabling the extension of the Web with a global data space based on open standards - the Web of Data. This paradigm has been used in an increasing number of data stores in recent years, including data stores with spatio-temporal data. Analysis of spatio-temporal data is not a straightforward task due to the complexity of the data structures, together with the representation and manipulation of the data involved. This paper considers challenges for modelling and management of spatio-temporal Linked Data and describes the first prototype of the Exploratory Spatio-Temporal Analysis tool for Linked Data developed by the Institute Mihajlo Pupin in the GeoKnow framework.**

***Index Terms*— Linked Data; integration; modelling; spatio-temporal; data exploration**

I. Introduction

In recent years, Semantic Web methodologies and technologies have strengthened their positions in the areas of data and knowledge management. Standards for organizing and querying semantic information, such as RDF(S) [1] and SPARQL [2] are adopted by large academic communities, while corporate vendors adopt semantic technologies to organize, expose, exchange and retrieve their data as Linked Data [3].

Linked Data paradigm has been utilized recently in order to achieve linking of datasets together through references to common concepts. The approach recommends use of HTTP uniform resource identifiers (URI) to name the entities and concepts so that consumers of the data can look-up those URIs to get more information, including links to other related URIs. RDF stores, that are used to store Linked Data, have become robust enough to support volumes of billions of records (RDF triples). They offer data management and querying functionalities very similar to those of traditional relational database systems. Integrating Semantic Web with geospatial data management requires the scientific community to address two challenges: (i) the definition of proper standards and vocabularies that describe geospatial information according to RDF(S) and SPARQL protocols, that also conform to the principles of established geospatial standards, (e.g. OGC), (ii) the development of technologies for efficient storage, robust indexing, and native processing of semantically organized geospatial data. Recently, GeoSPARQL [4] has emerged as a promising standard from W3C for geospatial RDF, with the aim of standardizing geospatial RDF data modelling and querying.

In this paper we will describe the Exploratory Spatio-Temporal Analysis tool for Linked Data (ESTA-LD). We will start by describing the GeoKnow Generator [5], [6], a suite of tools for geospatial Linked Open Data (LOD), into which the final ESTA-LD component will be integrated, then we will discuss some challenges and directions for modelling spatio-temporal data and finally we will describe the functionalities and implementation of the ESTA-LD first prototype.

A. Related work

Several projects have been financed within the EU FP7 research program devoted to publishing and consuming data in Linked Data format. As a result, several repositories of open source toolkits, as well as platforms for building Linked Data applications have emerged recently, as is presented in Table I.

Currently, there are three major sources of open geospatial data in the Web: Spatial Data Infrastructures (SDI), open data catalogues, and crowdsourced initiatives. Among various efforts we highlight OpenStreetMap (www.openstreetmap.org), GeoNames (www.geonames.org), and Wikipedia (www.wikipedia.org) as the most significant.

The development of ESTA-LD was motivated by the GeoKnow requirements and by the features that were not available in the GeoKnow components Facete [7] and CubeViz [8].

II. Geospatial Linked Open Data Lifecycle Management

Applications based on Linked Data can be more flexible than their traditional counterpart. The tools integrated in the *LinkedData* Stack (see http://stack.linkeddata.org/), for example, enable developers to build custom applications on top of public sector data. Leveraging *LinkedData* Stack components, in the GeoKnow framework, the GeoKnow Generator, an integrated solution for managing geospatial data, has been developed.

Dejan Paunović is with the Institute Mihajlo Pupin, University of Belgrade, Volgina 15, 11060 Belgrade, Serbia (e-mail: dejan.paunovic@pupin.rs).

Dr Valentina Janev is with the Institute Mihajlo Pupin, University of Belgrade, Volgina 15, 11060 Belgrade, Serbia (e-mail: valentina.janev@pupin.rs).

Vuk Mijović is with the Institute Mihajlo Pupin, University of Belgrade, Volgina 15, 11060 Belgrade, Serbia (e-mail: vuk.mijovic@pupin.rs).

TABLE I
OPEN SOURCE TOOLKITS FOR LINKED DATA

FP7 project	Tool
LATC	*Data Publication & Consumption Tools Library*, http://wifo5-03.informatik.uni-mannheim.de/latc/toollibrary/
LATC	*LATC 24/7 Interlinking Platform*, http://latc-project.eu/platform
LOD2, http://lod2.eu, GeoKnow, http://geoknow.eu	*Linked Data Stack*, http://stack.linkeddata.org/,
PlanetData	*PlanetData Tool Catalogue*, http://planet-data.eu/planetdata-tool-catalogue
OpenCube, http://opencube-project.eu/	*Extensions of the Information Workbench*, http://www.fluidops.com/information-workbench/.

A. *GeoKnow Generator*

The *GeoKnow Generator* (http://generator.geoknow.eu) is a full suite of tools supporting the complete lifecycle of geospatial linked open data. It will enable publishers to:

- triplify geospatial data,
- interlink geospatial data with Linked Data sources,
- fuse and aggregate linked geospatial data,
- visualize and author linked geospatial data in the Web.

It aims to provide support to semantic interoperability, interlinking, querying, reasoning, aggregation, fusion, and visualization of geospatial data. It provides a comprehensive toolset of easy to use applications covering a range of possible usage scenarios (e.g. mobility/traffic, energy/water, culture, etc).

It allows the user to triplify geospatial data, such as ESRI shapefiles and spatial tables hosted in major DBMSs using the GeoSPARQL, WGS84 or Virtuoso RDF vocabulary for point features geospatial representations (TripleGeo). Non-geospatial data in RDF (local and online RDF files or SPARQL endpoints) or data from relational databases (via Sparqlify) can also be entered into the Generator's triple store. With these two sources of data it is possible to link (via LIMES), to enrich (via GeoLift), to query (via Virtuoso), to visualize (via Facete - see Fig. 1) and to generate light-weight applications as JavaScript snippets (via Mappify) for specific geospatial applications. Most steps in the Linked Data lifecycle [3] have been integrated in the Generator as a graph-based workflow that allows the user to easily manage new generated data. The components comprising it are available in the *Linked Data* Stack [6].

The *GeoKnow Generator* is currently still in development. As new tools are developed, or existing tools are improved, these will be incorporated both in the *Generator* and in the *Linked Data* Stack.

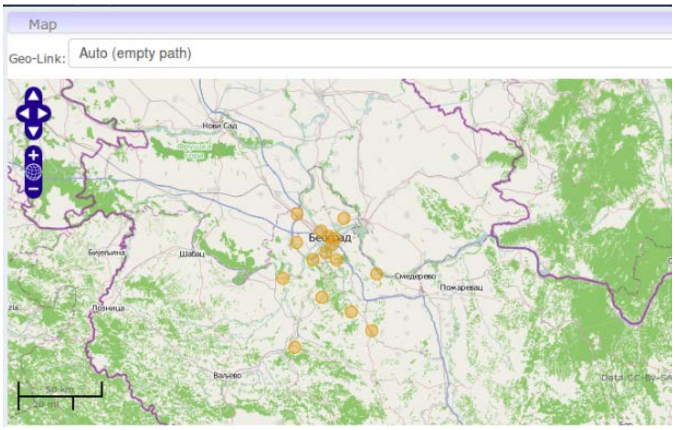

Geo-Link: Auto (empty path)

Fig. 1. Visualization with Facete.

B. *Challenges*

The GeoKnow Generator addresses the following challenges:

- Creation and maintenance of qualitative geospatial information from existing unstructured data such as OpenStreetMap, Geonames and Wikipedia,
- Reuse and exploitation of unforeseen discoveries found in geospatial data,
- Mapping and exposing existing structured geospatial information on the web of data, considering comprehensive and qualitative ontologies and efficient spatial indexing,
- Automatic fusing and aggregation of geospatial data by developing algorithms and services based on machine learning,
- Exploring, searching, authoring and curating the Spatial Data Web by using Web 2.0 and machine learning techniques.

Spatio-temporal data models and query languages are a topic of growing interest. They are used with applications where data types are characterized by both spatial and temporal semantics [9]. In the past, in spite of many similarities, research in spatial and temporal data models and databases has largely developed independently. Spatial database research has focused on modelling, querying, and integrating geometric and topological information in databases. Temporal database research has concentrated on modelling, querying, and recording the temporal evolution of facts under different notions of time (valid time, transaction time) and thus on extending the knowledge stored in databases about the current and past states of the real world. Both the temporal and spatial dimensions add substantial complexity to the problem, therefore, spatio-temporal data handling is not a straightforward task due to the complexity of the data structures, together with the representation and manipulation of the data involved.

III. Modelling spatio-temporal data

Many data objects in real world have attributes related to both space and time, thus imposing challenges for visualizing both dimensions on a geographical map. Moreover, these data objects are often multi-dimensional in nature meaning that the information can be represented on different granularity levels in space and time, as well as the type of information.

A statistical data set comprises 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). 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.

A. *Modelling statistical data (RDF Data Cube vocabulary)*

In January 2014, W3C recommended the RDFData Cube vocabulary [10], a standard vocabulary for modelling statistical data, see http://www.w3.org/TR/vocab-data-cube/. The vocabulary focuses purely on the publication of multi-dimensional data on the Web. The model builds upon the core of the SDMX 2.0 Information Model [11].

In the example of modelling socio-economic data that include spatio-temporal information that is given bellow, we have a coarse-grained representation of the indicator "Tourists arrivals" for the territory of country Serbia (geo:RS) and for year 2005 (time:Y2005). Additionally the indicator represents the "Total" number of tourists including "Domestic" and "Foreign". The RDF Data Cube vocabulary is used to represent one single observation as follows:

```
http://elpo.stat.gov.rs/lod2/RS-
   DATA/Tourism/Tourists_arrivals/data/obs1>
aqb:Observation ;
rs:geogeo:RS ;
rs:time time:Y2005 ;
rs:dataTypedatatype:number ;
rs:obsIndicator "Tourists arrivals" ;
rs:obsTurists "Total" ;
qb:dataSet<http://elpo.stat.gov.rs/lod2/RS-
   DATA/Tourism/Tourists_arrivals/data> ;
sdmx-measure:obsValue "1988469" .
```

B. *Representing spatial data in RDF (WGS84 vocabulary)*

W3C Semantic Web Interest Group has developed a very minimalistic RDF vocabulary for describing Points with latitude, longitude, and altitude properties from the reference datum specification (see http://www.w3.org/2003/01/geo/). This design allows for basic information about points to be described in RDF/XML, and augmented with more sophisticated or application-specific metadata. An example of using this vocabulary for specifying the location of a company on a geographical map is given bellow:

```
@prefix org: <http://www.w3.org/ns/org#> .
@prefix vcard:
   <http://www.w3.org/2006/vcard/ns#> .
@prefix wgs84:
   <http://www.w3.org/2003/01/geo/wgs84_pos#>

<http://rs.pupin.muplatexample/Beograd> a
   org:Organization ;
rdfs:label "SedišteUpave u Beogradu" ;
wgs84:lat "44.811314"^^xsd:float ;
wgs84:long "20.487436"^^xsd:float ;
org:Site<http://rs.pupin.muplatexample/vcard/0
   > .
```

C. *Modelling using SKOS vocabulary*

In order to formalize the conceptualization of hierarchical dimensions (space, time), we can use the Simple Knowledge Organization System (SKOS), see http://www.w3.org/TR/2005/WD-swbp-skos-core-spec-20051102/. 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 in 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:hasTopConcept.

Herein, we will present an example of coding the space and time dimension in RDF.

Example: Spatial dimension

```
geo:RS21
rdf:typegeo:Region ;
owl:sameAs
    <http://dbpedia.org/page/%C5%A0umadija_and
   _Western_Serbia> ;
skos:broadergeo:RS ;
skos:narrower geo:RS212 , geo:RS216 ,
   geo:RS211 , geo:RS215 , geo:RS213 ;
skos:narrower 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 .
```

SKOS properties skos:broader and skos:narrower can be used for relating concepts of same type, in our case geographical area (geo:Region). However, if the concepts are not of the same type (e.g. to regions and municipalities), the skos:relatedalignment can be applied.

```
geo:_70980
rdf:typegeo:Municipality ;
owl:sameAs<http://dbpedia.org/resource/Prijepo
   lje> ;
skos:notation "70980"^^xsd:string ;
skos:prefLabel "Prijepolje"@en ;
skos:related geo:RS211 .
```

Example: Time dimension

Observed data can be described with time information using different formats (e.g. seconds from the begging of an event, day-time, day, month, year).

One way to specify the frequency of data (or time granularity) in a dataset is to use the SDMX CONTENT-ORIENTED GUIDELINES, see http://sdmx.org/wp-content/uploads/2009/01/02_sdmx_cog_annex_2_cl_2009.pdf

```
@prefix rdf:   <http://www.w3.org/1999/02/22-
   rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-
   schema#> .
@prefix skos:
   <http://www.w3.org/2004/02/skos/core#> .
@prefix time:
   <http://elpo.stat.gov.rs/lod2/RS-DIC/time#>
   .

time:P1D
rdf:typeowl:Class ;
rdfs:subClassOf skos:Concept ;
skos:prefLabel "Daily (or Business)"@en .

time:P1M
rdf:typeowl:Class ;
rdfs:subClassOf skos:Concept ;
skos:prefLabel "Monthly"@en .

time:P1Y
rdf:typeowl:Class ;
rdfs:subClassOf skos:Concept ;
skos:prefLabel "Annual"@en .

time:Y1980
rdf:type time:P1Y ;
skos:notation "Y1980"^^xsd:string ;
skos:prefLabel "1980"@en .

time:Y1980M1
rdf:type time:P1M ;
skos:broader time:Y1980Q1 ;
skos:notation "Y1980M1"^^xsd:string ;
skos:prefLabel "1980/january"@en .
```

IV. ESTA-LD FIRST PROTOTYPE

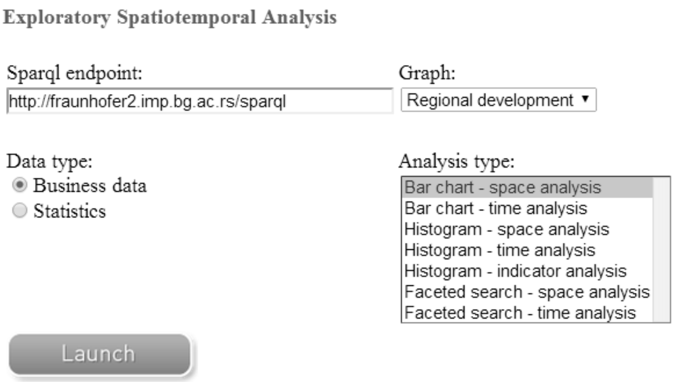
In order to demonstrate the applicability of the *GeoKnow Generator* for advanced spatio-temporal analysis of Linked data and statistics that appear at different levels of granularity, the ESTA-LD component was developed. This component deals with processing, presentation and exploration of collected data, uploaded in a form of a graph in the local RDF store.

A. *Experimental results*

ESTA-LD component supports common analysis in space and time dimension for resources described with standard vocabularies. These vocabularies are:

- RDF Data Cube vocabulary (http://www.w3.org/TR/vocab-data-cube/) for statistical data,
- Organization vocabulary (http://www.w3.org/ns/org#) and
- Registered Organization vocabulary (http://www.w3.org/ns/regorg#) for business entities.

When launching the ESTA-LD component, the user specifies the SPARQL endpoint, and selects the graph that contains the data to be explored and the required analysis type (see Fig. 2).



Exploratory Spatiotemporal Analysis

Sparql endpoint: http://fraunhofer2.imp.bg.ac.rs/sparql

Graph: Regional development ▾

Data type:
● Business data
○ Statistics

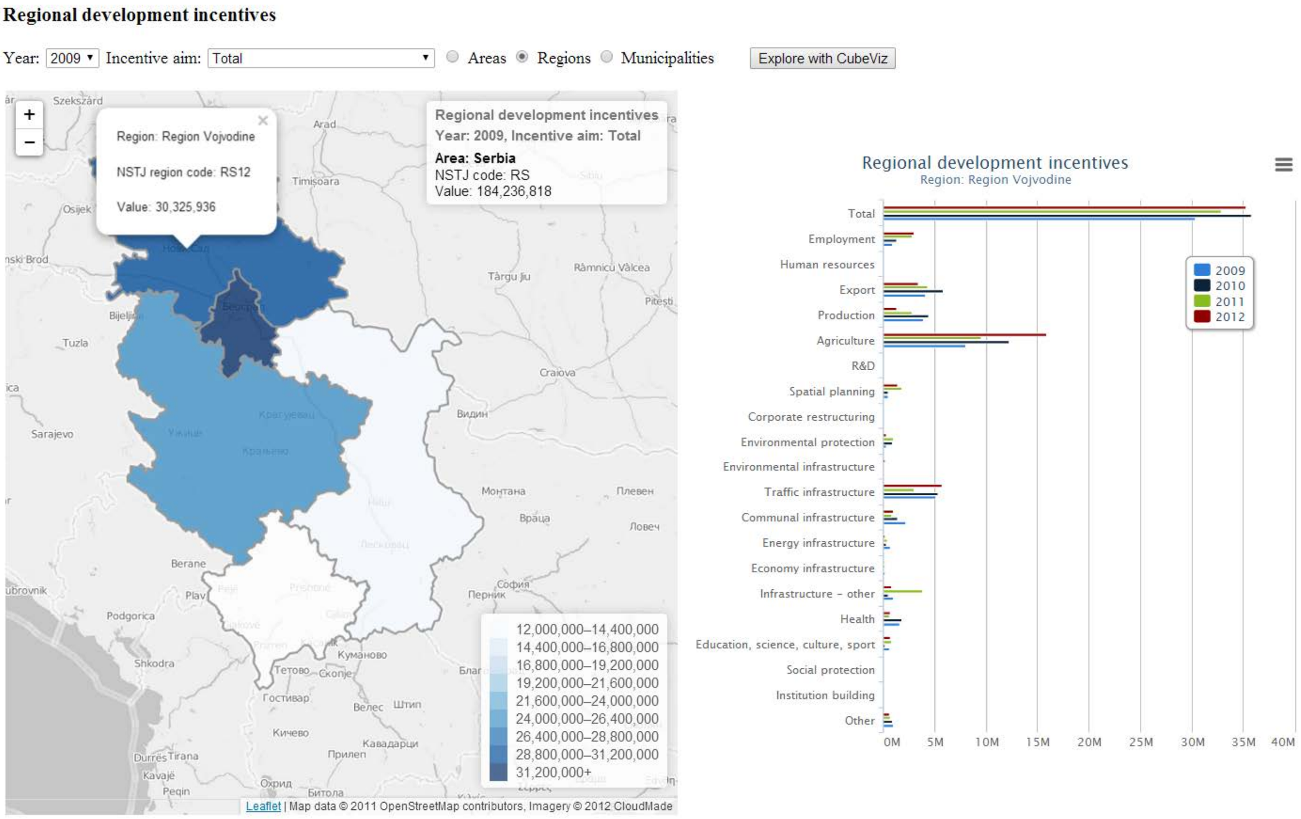
Analysis type:
Bar chart - space analysis
Bar chart - time analysis
Histogram - space analysis
Histogram - time analysis
Histogram - indicator analysis
Faceted search - space analysis
Faceted search - time analysis

Fig. 2. Launching the ESTA-LD component.

The data is then retrieved from the specified SPARQL endpoint and visualized on the choropleth map. The choropleth map provides an easy way to visualize how measurement varies across a geographic area. It is an ideal way to communicate spatial information quickly and easily, since the data is aggregated or generalized into classes or categories that are represented on the map by grades of colour. The ranges of data values for different colors are recalculated every time a new set of data is retrieved from the SPARQL endpoint.

After the data is retrieved, the user can utilize different filtering options that are currently implemented (see Fig. 3):

- For selecting values from the time dimension,
- For selecting the indicator under study,
- For selecting the granularity level for the space dimension,
- Interactive selection of the area of interest on the geographical map (for the selected area a bar-chart or histogram representation of the indicator is displayed).



Regional development incentives

Year: 2009 ▾ Incentive aim: Total ▾ ○ Areas ● Regions ○ Municipalities Explore with CubeViz

Region: Region Vojvodine

NSTJ region code: RS12

Value: 30,325,936

Regional development incentives
Year: 2009, Incentive aim: Total
Area: Serbia
NSTJ code: RS
Value: 184,236,818

Fig. 3. ESTA-LD Example of spatial analysis and bar chart.

B. *Implementation*

The first prototype of ESTA-LD component is currently available at http://fraunhofer2.imp.bg.ac.rs/esta-ld/. It was developed using HTML5 and JavaScript in order to enable ease of integration with the *GeoKnow Generator*. Representation and interaction with geographic information were implemented using *Leaflet*, an open source JavaScript library for mobile-friendly interactive maps (see http://leafletjs.com/). The geographic data (such as region borders), originally created from shape files, is stored in GeoJSON format. It is brought in and programmed with JavaScript and added to maps to create interactive visualizations. On the other side, different statistical indicators, which are the subjects of the spatio-temporal analysis, are stored in the RDF Data Store. This data repository is accessed and queried using SPARQL query language. The actual retrieval of data from the SPARQL Endpoint is implemented using the jQuery library and its standard getJSON function. Finally, the results of the spatio-temporal analysis are visualized using *Highcharts*, a charting library written in pure HTML5/JavaScript, offering intuitive, interactive charts to a web site or web application (see http://www.highcharts.com/). 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, the integration of ESTA-LD will be straightforward. ESTA-LD is a JavaScript web application that can be configured to work with any SPARQL endpoint, meaning that its user interface will be easy to integrate with the *Generator* and that it can be configured to work with the same RDF store as the other components.

C. *Benefits for early adopters*

In this phase, the ESTA-LD component has been tested with the Register of the Regional Development Measures and Incentives, maintained by the Serbian Business Registers Agency (SBRA, http://www.apr.gov.rs/). The Register is a unique, centralized electronic database of the taken measures and implemented incentives that are of significance for regional development. It aims at enabling a more comprehensive monitoring of investments and comparative analyses thereof, for the purpose of further planning and investments in the development of certain regions, with the ultimate goal of reducing regional disparities and improving regional competitiveness. Data (aggregated, interlinked with public datasets and fused) appear at different levels of granularity, thus, imposing challenges for data processing, presentation and exploration.

SBRA will be the first adopter of the developed ESTA-LD component and its innovative approach to collecting and visualizing data. ESTA-LD component will extend the map-based dashboard with interactive visualization and analysis of the multidimensional data, as well as interlinking the data from the Register with data from other public agencies. Geospatial visualization and analytics will help policy makers (public administration, businesses and citizens) to easily observe multiple factors at a single glance, and better understand how different characteristics relate to one other - and to a specific place (region).

V. CONCLUSION AND FUTURE WORK

In this paper we have presented the Exploratory Spatio-Temporal Analysis tool for Linked Data.

The ESTA-LD component contributes to further development and standardization of Linked Data technologies. It showcases the use of emerging W3C standards (RDF Data Cube vocabulary, SKOS vocabulary OWL-Time ontology) in the field of statistical and temporal geospatial information management.

Our initial testing has shown that the first prototype of the ESTA-LD tool, although still in development, proved to be a valuable instrument for advanced spatio-temporal analysis of Linked Data. Besides bug fixes, some future work will include improvements of the layout and some of the controls. An advanced GUI will be developed that will allow drill-down style online access to highly granular data. The whole geospatial information life cycle in the resulting first prototype for exploratory spatio-temporal analysis will be further tested with data provided by the Serbian Business Registers Agency, and different statistical indicators published by the Serbian Statistical Office at the Serbian CKAN (http://rs.ckan.net). A significant effort will be put into further generalization of the ESTA-LD filtering and exploration and integration of the component into the *GeoKnow Generator*. Finally, the tool is planned to be further enhanced taking 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 LOD2 project, Pr. No: 257943; 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] W3C - RDF Schema, http://www.w3.org/TR/rdf-schema/

[2] W3C - SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparql-query/

[3] S. Auer, J. Lehmann, “Making the web a data washing machine - creating knowledge out of interlinked data”, Semantic Web Journal, vol. 1, no. 12, pp. 97-104, IOS Press, 2010.

[4] M. Perry, J. Herring (eds), “OGC GeoSPARQL standard - A geographic query language for RDF data”, Open Geospatial Consortium Inc, v.1.0, 27/04/2012.

[5] A. Garcia-Rojas, S. Athanasiou, J. Lehmann, D. Hladky, “GeoKnow: Leveraging Geospatial Data in the Web of Data”, In Proc. Open Data on the Web (ODW13), 23-24 April 2013, Google Campus, Shoreditch, London/UK.

[6] D. Hladky, A. Garcia Rojas, J. Lehmann, “GeoKnow – Making the Web an Exploratory Place for Geospatial Knowledge”. ERCIM Journal 96, January 2014

[7] C. Stadler, M. Martin, S. Auer, “Exploring the Web of Spatial Data with Facete”, Companion proceedings of 23rd International World Wide Web Conference WWW, page 175--178. (2014)

[8] P. E Salas, F. Maia Da Mota, K. Breitman, M. A Casanova, M. Martin, S. Auer, “Publishing Statistical Data on the Web”, *International Journal of Semantic Computing* 06(04):373-388, 2012.

[9] N. Pelekis, B. Theodoulidis, I. Kopanakis, Y. Theodoridis, 2004 “Literature review of Spatio-temporal database models”, Knowl. Eng. Rev. J. 19, pp. 235–274.

[10] R. Cyganiak, D. Reynolds, J. Tennison, “The RDF Data Cube vocabulary”, 14. July 2010.

[11] “SDMX Information model: UML Conceptual Design (version 2.0)”, November 2005, http://sdmx.org/docs/2_0/SDMX_2_0%20SECTION_02_InformationModel.pdf