Transformation and Analysis of Spatio-Temporal Statistical Linked Open Data with ESTA-LD

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*—Recent open data initiatives have contributed to opening of non-sensitive governmental data, thereby accelerating the growth of the LOD cloud and the open data space in general. A lot of this data is statistical in nature and refers to different geographical regions (or countries), as well as various performance indicators and their evolution through time. Publishing such information as Linked Data provides many opportunities in terms of data aggregation/integration and creation of information mashups. However, due to Linked Data being relatively new field, currently there is a lack of tools that enable exploration and analysis of linked geospatial statistical datasets. This paper presents ESTA-LD (Exploratory Spatio-Temporal Analysis), a tool that enables exploration and visualization of statistical data in linked data format, with an emphasis on spatial and temporal dimensions, thus allowing to visualize how different regions compare against each other, as well as how they evolved through time. Additionally, the paper discusses best practices for modeling spatial and temporal dimensions so that they conform to the established standards for representing space and time in Linked Data format.**

I. INTRODUCTION

As stated by the OECD (The Organization for Economic Co-operation and Development), *"Open Government Data (OGD) is fast becoming a political objective and commitment for many countries"*. In the recent years, various OGD initiatives, such as the Open Government Partnership[1], 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, environment, etc. Moreover, the vision for ICT-driven public sector innovation [1] 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. 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 parallel, 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 [2]. 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 RDF Data Cube vocabulary [3] which enables modeling of statistical data as Linked Data is a W3C recommendation since January 2014, and the GeoSPARQL [4] standard that supports representing and querying geospatial data on the Semantic Web was published in June 2012, while 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, discovery of different facts related to places, and identifying and assessing existing methods and tools in order to create a set of best practices. As a consequence, the tools that are based on these standards are scarce, and representation of spatio-temporal concepts may vary across different datasets.

This paper describes ESTA-LD (Exploratory Spatio-Temporal Analysis of Linked Data), a tool that enables exploration and analysis of spatio-temporal statistical linked open data. The RDF Data Cube vocabulary, which is the basis of this tool, is discussed in Section 2, along with the best practices for representing spatial and temporal information as Linked Data and transformation services that help to transform different kinds of spatial and temporal dimensions into a form that is compliant with ESTA-LD. The tool itself, and its functionalities are given in Section 3, while the conclusions and outlook on future work are given in Section 4.

The work described in this paper builds upon and extends previous efforts elaborated in [5, 6].

II. CREATING SPATIO-TEMPORAL STATISTICAL DATASETS AND TRANSFORMING THEM WITH ESTA-LD

This section will discuss modeling of spatio-temporal statistical datasets as Linked Data. First, standard well-known vocabularies will be introduced, followed by recommendations based on these standards. Finally, the section will describe the approach taken in ESTA-LD and introduce its services that help to transform different spatial and temporal dimensions into an expected form.

A. Modeling statistical data as Linked Data

The best way to represent statistical data as Linked Data is to use the RDF Data Cube vocabulary, a well-known vocabulary recommended by the W3C for modelling statistical data. This vocabulary is based on the *SDMX 2.0 Information Model* [7] which is the result of the Statistical

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

Data and Metadata Exchange (SDMX[2]) Initiative, an international initiative that aims at standardizing and modernizing ("industrializing") the mechanisms and processes for the exchange of statistical data and metadata among international organizations and their member countries. Having in mind that SDMX is an industry standard, and backed up by influential organizations such as Eurostat, World Bank, UN, etc., it is of crucial importance that RDF Data Cube vocabulary is compatible with SDMX. Additionally, the linked data approach brings the following benefits:

- Individual observations, as well as groups of observations become web addressable, thus enabling third parties to link to this data,
- Data can be easily combined and integrated with other datasets, making it integral part of the broader web of linked data,
- Non-proprietary machine readable means of publication with out-of the-box web API for programmatic access,
- Reuse of standardized tools and components.

Each RDF Data Cube consists of two parts: structure definition, and (sets of) observations. The Data Structure Definition (DSD) provides the cube's structure by capturing specification of dimensions, attributes, and measures. Dimensions are used to define what the observation applies to (e.g. country or region, year, etc.) and they serve to identify an observation. Therefore, a statistical data set can be seen as a multi-dimensional space, or hyper-cube, indexed by those dimensions, hence the name cube (although the name cube should not be taken literally as the statistical dataset can contain any number of dimensions, not just exactly three). Attributes and measures on the other hand provide metadata. Measures are used to denote what is being measured (e.g. population or economic activity), while attributes are used to provide additional information about the measured values, such as information on how the observations were measured as well as information that helps to interpret the measured values (e.g. units). The explicit structure captured by the DSD can then be reused across multiple datasets and serve as a basis for, validation, discovery, and visualization.

However, in order to spur reuse and discoverability, the data structure definition should be based on common, well-known concepts. To tackle this issue, the SDMX standard includes a set of content oriented guidelines (COG) which define a set of common statistical concepts and associated code lists that are meant to be reused across datasets. Thanks to the efforts of the community group, these guidelines are also available in linked data format and they can be used as a basis for modeling spatial and temporal dimensions. Although they are not part of the vocabulary and do not form a Data Cube specification, these resources are widely used in existing Data Cube publications and their reuse in newly published datasets is highly recommended.

B. Representing space and time

One of the earliest efforts for representing spatial information as linked data is the Basic Geo Vocabulary by W3C. This vocabulary does not address many of the issues covered in the professional GIS world, however it provides a namespace for representing latitude, longitude and other information about spatially-located things, using the WGS84 CRS as the standard reference datum.

Since then, GeoSPARQL has emerged as a promising standard [2]. The goal of this vocabulary is to ensure consistent representation of geospatial semantic data across the Web, thus allowing both vendors and users to achieve uniform access to geospatial RDF data. To this end, GeoSPARQL defines an extension to the SPARQL query language for processing geospatial data, as well as a vocabulary for representing geospatial data in RDF. The vocabulary is concise and among other things, it enables to represent features and geometries, which is of crucial importance for spatial visualizations, such as ESTA-LD. Following is an example of specifying a geometry for a particular entity using GeoSPARQL:

```
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
eg:area1 geo:hasDefaultGeometry eg:geom1 .
eg:geom1 geo:asWKT "MULTIPOLYGON(…)"^^geo:wktLiteral
```

Therefore, GeoSPARQL enables to explicitly link a spatial entity to a corresponding serialization that can be encoded as WKT (Well Known Text) or GML (Geography Markup Language). Alternatively, one can refer to geographical regions by referencing well-known data sources, such as GeoNames. This approach is simpler and less verbose, however in this case the dataset doesn't contain the underlying geometry and is therefore not self-sufficient and requires any tool that operates on top of it to acquire the geometries from other sources.

In order to represent time, the two most common approaches are the use of the OWL Time ontology and using the XSD date and time data types. The OWL time ontology presents an ontology of temporal concepts, and at the moment, it is still a W3C draft. It provides a vocabulary for expressing facts about topological relations among instants and intervals, together with information about durations, and about datetime information. On the other hand, the XSD date and time data types can be used to represent more basic temporal concepts, such as points in time, days, months, and years, however they cannot be used to denote custom intervals (e.g. a period from 15th of January till 19th of March). Although they are clearly less expressive than the OWL time ontology, XSD date and time data types are widely used and supported in existing tools and libraries, thus requiring little to no effort for type transformation when third party libraries are used.

C. ESTA-LD approach

First and foremost, ESTA-LD is based on the RDF Data Cube vocabulary, and any dataset based on it can be visualized on a chart. However, ESTA-LD offers additional features (visualizations) for datasets containing a spatial and/or temporal dimension. In order for these features to be enabled, the dataset needs to abide to certain principles that were described earlier. Namely, spatial entities need to be linked to their corresponding geometries using GeoSPARQL vocabulary, while the serialization needs to be encoded as a WKT string. On the other hand, with regards to the temporal dimension, ESTA-LD can handle values encoded as XSD date and time data types. Furthermore, even though it is not

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

required, the use linked data version of the content oriented guidelines for specifying and explicitly indicating spatial and temporal dimensions is highly encouraged. An example observation, along with the definition of the temporal and the spatial dimension is given in Figure 1.

```
@prefix qb: <http://purl.org/linked-data/cube#>
@prefix geo: <http://www.opengis.net/ont/geosparql#>

eg:refArea a rdf:Property, qb:DimensionProperty;
  rdfs:label "reference area"@en;
  rdfs:subPropertyOf sdmx-dimension:refArea ;
  rdfs:range eg:Area, geo:Feature ;
  qb:concept sdmx-concept:refArea .

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

eg:obs1 a qb:Observation ;Blah,
  qb:dataset eg:ds1 ;
  eg:refArea eg:area1 ;
  eg:refPeriod "2014-12"^^xsd:gYearMonth ;
  sdmx-measure:obsValue "123623" .

eg:area1 geo:hasDefaultGeometry eg:geom1 .
eg:geom1 geo:asWKT "MULTIPOLYGON(…)"^^geo:wktLiteral .
```

Figure 1 Spatio-Temporal Data Cube Example

This example shows the spatial dimension `eg:refArea` that is derived from the dimension `sdmx-dimension:refArea` and associated with the concept `sdmx-concept:refArea`, both of which are available in the linked data version of the content oriented guidelines. Similarly, the temporal dimension is derived from `sdmx-dimension:refPeriod` and associated with the concept `sdmx-concept:refPeriod`. Finally, the observation uses the defined dimensions to refer to the particular time period and the geographical region, which is in turn linked to the geometry and its WKT serialization by using the GeoSPARQL vocabulary.

D. ESTA-LD Data Cube Transformation Services

The modelling principles for temporal and spatial dimensions are still in early stages and therefore not so well known and widespread, meaning that many Data Cubes may vary slightly when it comes to modelling these two dimensions. In other words, there is a reasonable chance that there are spatio-temporal datasets that do not clearly express the presence of spatial and temporal dimensions or that values of spatial and temporal dimensions are represented in a custom ("non-standard") way, thus requiring slight modifications in order to enable all of ESTA-LD's functionalities. To address this issue, ESTA-LD is accompanied with an "Inspect and Prepare" component that provides services for transforming spatial and temporal dimensions. This component provides a visual representation of the structure of the chosen data cube. The structure is displayed as a tree that shows dimensions, attributes, and measures, as well as their ranges, code lists (if a code list is used to represent values of that particular dimension/measure/attribute), and values that appear in the dataset. Furthermore, the tree view can be used to select any of the available dimensions and initiate transformation services.

1) Transforming Temporal Dimensions

Many temporal dimensions may miss a link to the concept representing a time role. Furthermore, in some cases, organizations may decide to use their own code lists to represent time. However, even if this is the case, the URIs representing time points usually contain all the information needed to derive the actual time. Take for example the code list used by the Serbian statistical office, where code URIs take the following form: http://elpo.stat.gov.rs/RS-DIC/time/Y2009M12. This URI clearly denotes the December of 2009, and it can be parsed in order to transform it to an XSD literal such as "2009-12"^^xsd:gYearMonth. To achieve this with ESTA-LD's *Inspect and Prepare* view (see Figure 2), one only needs to provide the pattern by which to parse the URIs and the target type, after which the component transforms all values, changes the range of the dimension, removes a link to the code list since code list is not used any more, and links the dimension to the concept representing the time role. The target type is selected from the drop-down list, while the pattern is provided in the text

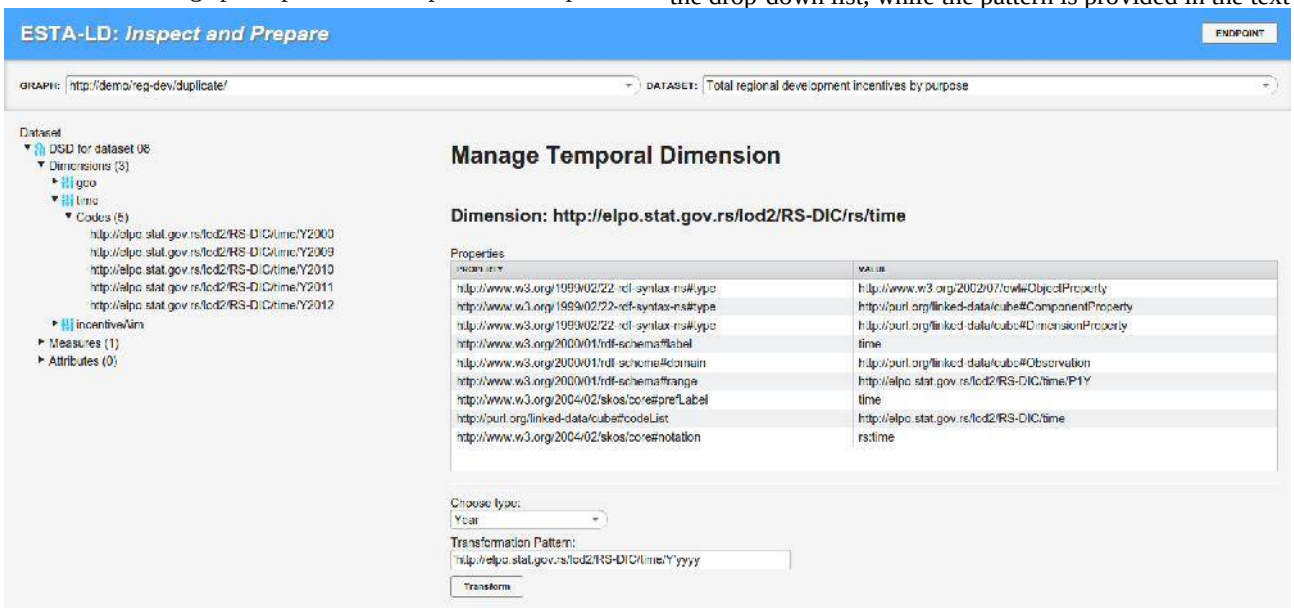

Figure 2 ESTA-LD Inspect and Prepare Component - Transformation of the Temporal Dimension

field.

2) Transforming Spatial Dimensions

In many cases, a Data Cube may contain a spatial dimension but miss the polygons which are required for visualization on the choropleth map. If this is the case, ESTA-LD's *Inspect and Prepare component* can be used to enrich the Cube with polygons acquired from LinkedGeoData. All that is needed on the user's part is to supply a pattern that the tool will use to extract one of the identifiers that can be used for the lookup, as well as to specify identifier's type, which can be one of the following: name, two-letter code (ISO 3166-1 alpha2) and three-letter code (ISO 3166-1 alpha 3). Similarly to the temporal dimension transformation service, the pattern is supplied in a text field, while the identifier type can be selected in a drop-down list. At the moment, this functionality can only be used to acquire polygons of countries.

III. ESTA-LD

ESTA-LD is a tool for visualizing statistical data in linked data format, i.e. data cubes modeled with the RDF Data Cube vocabulary. However, unlike other tools that treat any data cube in the same manner, such as CubeViz [8], ESTA-LD distinguishes spatial and temporal dimensions from the rest, and provides specialized visualizations based on their specific properties. Namely, if a Cube contains observations related to different geographic regions, i.e. it contains a spatial dimension, then the data can be visualized on a map where regions are colored in different grades/shades of blue based on observation values, thus giving intuitive insight into the disparities across regions. On the other hand, if a Cube contains measurements at different points in time, all measurements are organized on the time axis where a user can choose any time interval he or she wants to analyze and/or slide through time, thereby gaining insights into the evolution of the indicator under analysis over time.

A. Architecture and Implementation

ESTA-LD is a web application that can be deployed on any servlet container. Furthermore, it can operate on top of any SPARQL endpoint and accepts query string parameters for specifying the default endpoint and graph, thus ensuring that it can be used as a standalone tool, but at the same time easily integrated into other environments such as the GeoKnow Generator. It is based on the following frameworks and libraries:

- Vaadin: a Java framework for building web applications,
- Sesame: an open-source framework for querying and analysing RDF data,
- Leaflet: an open source JavaScript library for mobile-friendly interactive maps,
- Highcharts: a charting library written in pure HTML5/JavaScript, offering intuitive, interactive charts to a web site or web application,
- wellknown: a Javascript library for parsing and stringifying Well-Known Text into GeoJSON,
- jQuery: a Javascript library that makes things like HTML document traversal and manipulation, event handling, animation, and Ajax much simpler with an easy-to-use API that works across a multitude of browsers

Most of the user interface components, including the drop-down lists that are used to associate dimensions to particular values, are implemented in Vaadin, while the choropleth map and the chart are implemented using LeafletJS and Highcharts respectively, due to lack of adequate UI components in Vaadin. Sesame is used to query the selected SPARQL endpoint for available data cubes, as well as for the selected cube's structure, including contained dimensions, attributes, and measures. This information is used to populate Vaadin drop-down lists which allow the user to specify desired visualization, i.e. which dimension(s) will be visualized, and the values of other dimensions. After the user specifies the desired visualization, selection is passed to the Javascript layer, and jQuery is used to query the endpoint for observations that satisfy the selected criteria. After the endpoint returns the desired observations, the data is transformed and fed to the Leaflet map and Highcharts chart in the expected format. During the transformation process, wellknown is used to parse the WKT strings returned by the endpoint into GeoJSON that is required by Leaflet.

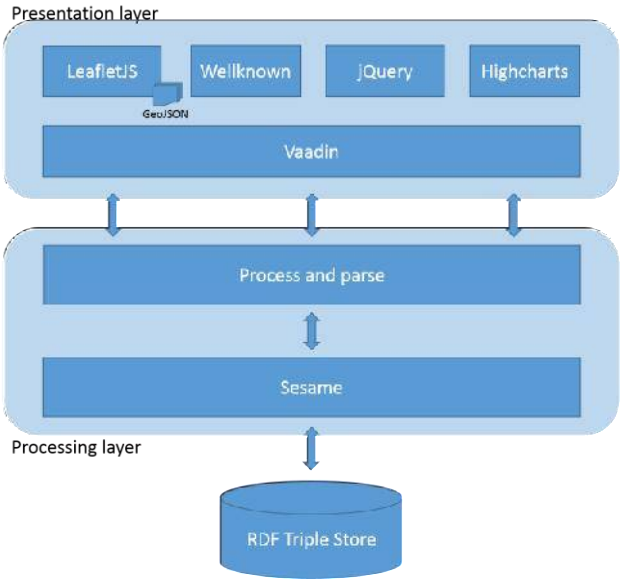

Figure 3 ESTA-LD Architecture

B. Chart Functionalities

The chart allows to analyze how observations vary across the selected dimension. If two dimensions are selected, values of the first dimension are laid on the X axis as categories, while values of the second dimension can be chosen in the legend as series (see Figure 4), thus allowing comparison between the selected dimensions. When two dimensions are visualized, it is also possible to swap series and categories, as well as to stack the values of the selected series. Furthermore, the chart allows to switch the axes, and to change its size by dragging the separator on the left side and showing/hiding the parameters section. Finally, in case the cube contains multiple measures, any two measures can be visualized in parallel in order to enable comparison, as shown in Figure 5. This example shows that measure comparison can be used to find correlations between different measures.

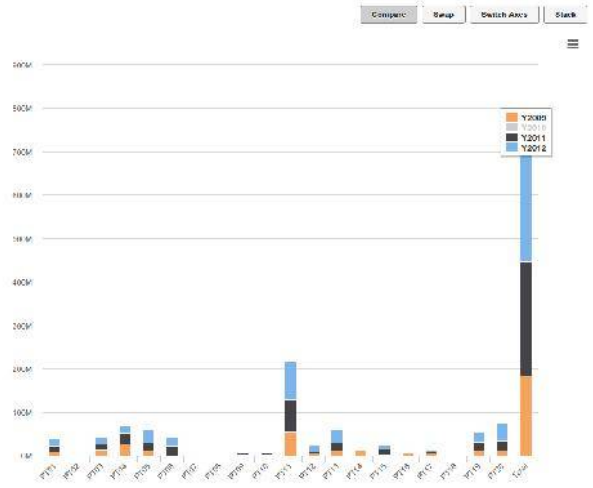

Figure 4 Chart Visualization - Two Dimensions

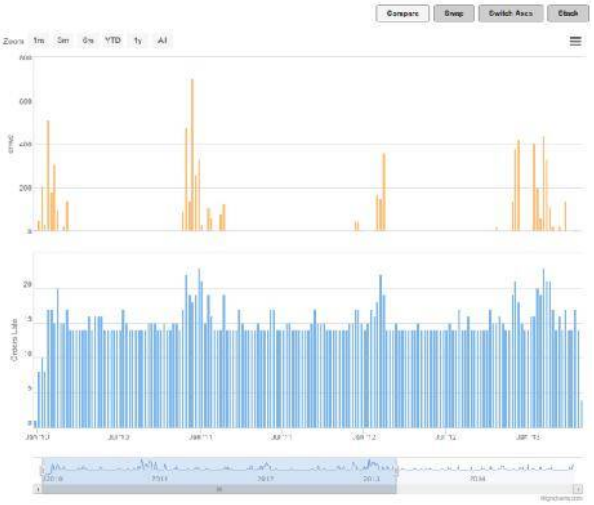

Figure 5 Time Chart - Comparing Two Measures

C. Spatio-Temporal Visualization/Analysis

The choropleth map always visualizes the spatial dimension, i.e. it shows the same information as the chart would if the only selected dimension were the spatial dimension. However, while the chart visualization would place a separate bar for each region, the choropleth map paints regions on the map in different shades of blue based on observation values. There are nine shades available, and each one represents a certain value range, where ranges are calculated based on the maximum and minimum observation value. This way, it is much easier to note disparities across different geographical regions (see Figure 6). The map also allows the user to select a particular region on the map, thus changing the region to be visualized in the chart. Finally, the choropleth map supports multiple hierarchy levels. Namely, if the hierarchical structure of geographical regions is given using the SKOS vocabulary, the tool allows to select which hierarchy level is to be visualized.

If the dataset contains a temporal dimension, and it is selected for visualization, the tool uses a specialized time chart that includes a timeline at the bottom. The timeline can be used to specify a particular time window to be visualized which is very useful when the underlying dataset contains huge number of time points, such as for example, daily data for a period of four years. For convenience, the time chart provides a shortcut for setting the duration of the time window to commonly used values such as 1 month, 3 months, 6 months, and 1 year. It is also possible to drag the time window, thus gaining an insight into the evolution of the selected indicator through time (see Figure 6).

Finally, the choropleth map and the time chart are fully synchronized. This means that whenever a region is selected on the map, the time chart is updated to show how the chosen measure evolved through time in that particular region. Similarly, whenever a time window is changed in the time chart, the map is updated to show only the selected period in time. Moreover, this happens immediately as the time window gets changed. As a

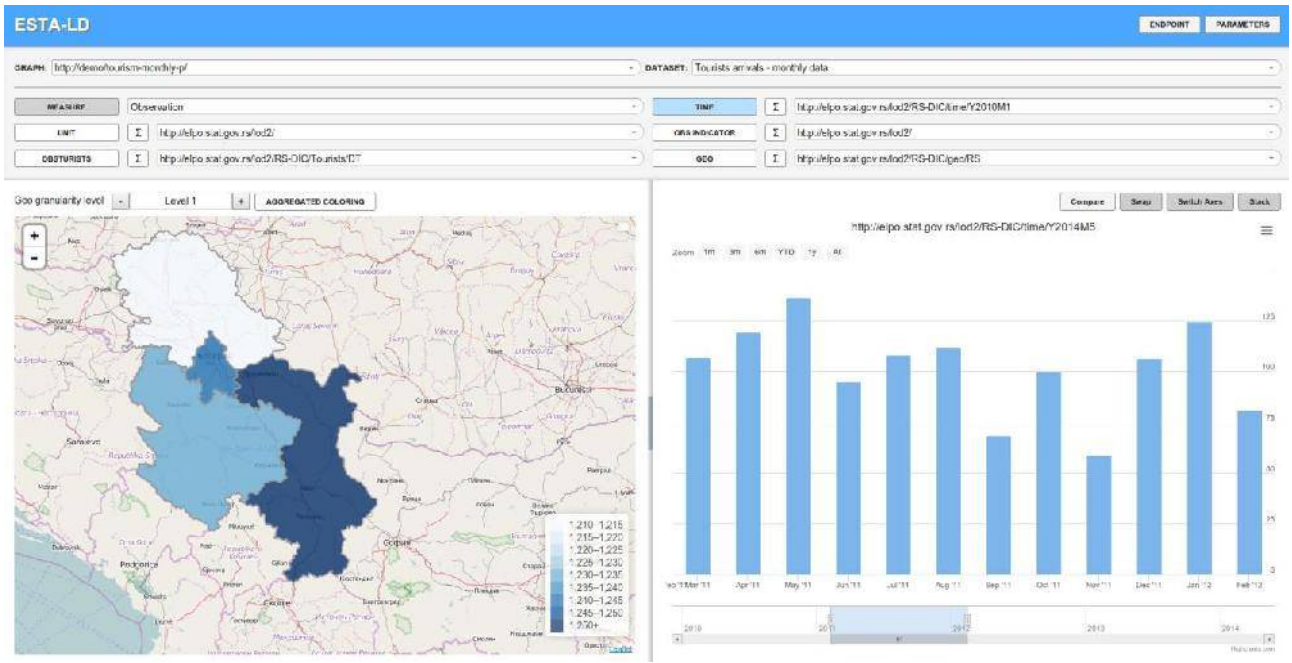

Figure 6 ESTA-LD Spatio-Temporal Visualization

consequence, by dragging the time window it is possible to get an insight into how different regions evolved over time. To further support this functionality, the map supports "aggregated coloring" which ensures that value ranges for different shades of blue do not change unless the duration of the window changes. Namely, without the aggregated coloring, whenever the time window is moved, the underlying set of observations to be visualized on the map changes, and with it the maximum and minimum observation value change as well. Consequently, value ranges for different shades get recalculated whenever the time window is dragged, making it impossible to determine how a particular region fairs against the previously selected period (since now every shade represents a different value range). With aggregated coloring employed, the tool calculates the minimum and maximum values based on every possible time window of the same duration and calculates the value ranges accordingly. This way, dragging of the time window doesn't impact the coloring scheme, making it unnecessary to recalculate the value ranges unless the duration of the time window changes. Therefore, it not only provides insight into disparities across regions through time, but also into the evolution of the chosen measure in each region that is shown on the map.

IV. Conclusions and Future Work

ESTA-LD is a tool that enables exploration and analysis of statistical linked open data. While it can visualize any statistical dataset on a chart, the tool puts an emphasis on spatial and temporal dimensions in order to enable spatio-temporal analysis. Namely, if the dataset contains a spatial and a temporal dimension, it is visualized on the choropleth map and the time chart respectively. Furthermore, these two views are synchronized, meaning that every selection in one of the views updates the other, thus providing insights into disparities of the chosen indicator across different geographical regions as well as their evolution through time. Furthermore, this paper showed how statistical data can be modeled as linked open data and discussed different approaches to representing spatio-temporal information within statistical data cubes, including the approach adopted by ESTA-LD. Having in mind that linked data is a relatively new technology where standards for representing spatial and temporal concepts are still shaping up, the paper described how ESTA-LD's *Inspect and Prepare* component can be used to transform different types of spatial and temporal dimensions to a form that is compliant with ESTA-LD.

In the future, ESTA-LD will be extended with additional types of graphs, and possibly with a data structure definition (DSD) repository that would reduce replication of information since DSDs can be reused and shared across different datasets. Finally, we intend to examine how to best leverage enrichment of statistical datasets with external sources of information such as DBpedia in order to provide advanced search and filtering capabilities over the data cubes.

Acknowledgment

The research presented in this paper is partly financed by the European Union (FP7 GeoKnow, Pr. No: 318159; CIP SHARE-PSI 2.0, Pr. No: 621012), and partly by the Ministry of Science and Technological Development of the Republic of Serbia (SOFIA project, Pr. No: TR-32010).

References

[1] 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.

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

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

[4] M. Perry, J. Herring, "OGC GeoSPARQL – A Geographic Query Language for RDF Data", http://www.opengis.net/doc/IS/geosparql/1.0, July 2012.

[5] 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.

[6] V. Mijović, V. Janev, D. Paunović: "ESTA-LD: Enabling Spatio-Temporal Analysis of Linked Statistical Data", *In Proceedings of the 5th International Conference on Information Society Technology and Management, Information Society of the Republic of Serbia*, pp 133-137, March 2015, Kopaonik, Serbia.

[7] SDMX Standards, "Information Model: UML Conceptual Design", https://sdmx.org/wp-content/uploads/SDMX_2-1-1_SECTION_2_InformationModel_201108.pdf , July 2011.

[8] M. Martin, K. Abicht, C. Stadler, A. Ngonga, T. Soru, S. Auer, "CubeViz: Exploration and Visualization of Statistical Linked Data", *In Proceedings of the 24th International Conference on World Wide Web*, pp 219-222, May 2015, Florence, Italy.