Supporting the Linked Data publication process with the LOD2 Statistical Workbench

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Valentina Janev[a,*], Bert Van Nuffelen[b], Vuk Mijović[a], Karel Kremer[b], Michael Martin[c], Uroš Milošević[a] and Sanja Vraneš[a]
[a]*Institute "Mihailo Pupin", University of Belgrade, Volgina 15, 11060 Belgrade, Serbia*
[b]*TenForce, Haachtsesteenweg 378, 1910 Kampenhout, Belgium*
[c]*Agile Knowledge Engineering and Semantic Web, Leipzig University, Augustusplatz 10, 04109 Leipzig, Germany*

Abstract. 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. Additionally, in order to increase transparency and improve interoperability and interaction with citizens and society as a whole, but also create new businesses and job opportunities, national governments publish their data in a machine-readable and future-proof format. In this paper we present the LOD2 Statistical Workbench, an integrated set of professional tools for accessing, manipulating, exploring and publishing statistical data. The data representation and processing is based on the W3C standard vocabularies (RDF Data Cube as a main model) and open source components delivered by the LOD2 consortium. The system meets the needs of both publishers and consumers of statistical data and directs the potential of the LOD2 tools to the specific domain of the statistical office. Using an illustrative case study of the Statistical Office of the Republic of Serbia, the paper introduces the user requirements, gives an overview of possible scenarios and shows examples of its use. The first results indicate that wider adoption of the LOD2 tools in practice can be foreseen.

Keywords: LOD2, Open government data, RDF Data Cube, statistical Linked Data, tools

1. Introduction

Statistical data is often used as the foundation for policy prediction, planning and adjustments, and therefore has a significant impact on 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 (LOD). The term Linked Data [1] here refers to a set of best practices for publishing and connecting structured data on the Web.

The Government Linked Data Working Group (http://www.w3.org/2011/gld/) has paid special attention to the provision of standards for publishing government data as Linked Data. Thus, the *RDF Data Cube vocabulary* has been proposed as a model for publishing multi-dimensional data on the Web (currently a W3C Candidate Recommendation). The model builds upon the core of the SDMX 2.0 Information Model [2]. The SDMX information model (see also ISO 17369:2013) has been developed to support statistics as collected and used by governmental and supra-national statistical organizations, as well as to be applicable to other organizational contexts involving statistical data and related metadata.

* Corresponding author. E-mail: Valentina.Janev@institutepupin.com.

Aimed at providing a set of professional tools for accessing, manipulating, exploring and publishing statistical data in a Linked Data format, the LOD2 consortium developed the LOD2 Statistical Workbench based on the LOD2 Stack [3], a collection of open-source tools. The LOD2 Stack[1] comprises tools from partners of the LOD2 project, as well as third parties, for managing the life-cycle of Linked Data. The LOD2 project [2] ("Creating knowledge out of interlinked data") 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.

This paper reports on the current state of the LOD2 Statistical Workbench and describes what has been achieved in the Serbian government use case in the LOD2 framework. The paper is structured as follows. First, we introduce the main challenges in Section 2 that motivated the development of the LOD2 Statistical Workbench. Subsequently, we present the application architecture and components integrated into the LOD2 Statistical Workbench in Section 3. Some illustrative examples will be sketched in Section 4, while Section 5 points to early adoption of the tool by Serbian government institutions. We conclude this work with envisioned enhancements.

2. Motivation

In the last decade, the European Commission (EC) has done considerable investments to improve efficiency in the provision of public services, increase transparency [4] and interaction with citizens and society as a whole, but also define better strategies for delivering large amounts of trusted data to the public and improve interoperability (see 'Interoperability Solutions for European Public Administrations' program[3] for the period from 2010-2015 [5]).

2.1. Related work

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, LATC[4];
- publishing and maintaining Linked geo-spatial data, e.g. TELEIOS[5], PlanetData[6], GeoKnow[7];
- facilitating professional training for data practitioners, who aim to use Linked Data in their daily work e.g. project EUCLID[8].

As a result, several repositories of open source toolkits, as well as platforms for building Linked Data applications have emerged recently:

- LATC Data Publication & Consumption Tools Library[9] and *LATC 24/7 Interlinking Platform*[10],
- *Linked Data Stack*[11],
- *PlanetData Tool Catalogue*[12],
- *Information Workbench*[13] [6].

The LOD2 consortium approached the problem of creating a more scalable and interoperable Open Government Data ecosystem by considering the latest advances in Linked Open Data and tools provided by the LOD2 partners (see the *Linked Data Stack*). On the publisher side, LOD2 tools can be used to build public services (e.g. LOD convertors) that will deliver trusted, open and rich collections of interlinked datasets to the public, while on consumer side, LOD2 tools can be used to explore and reuse public data. In Figure 1 the use of the CKAN[14] catalogue, a web-based open source data management system for building government portals, storage and distribution of open data is illustrated.

Evaluation of LOD2 tools for building an infrastructure for public sector information is given in [7].

[1] http://stack.lod2.eu
[2] http://lod2.eu
[3] http://ec.europa.eu/isa/
[4] http://latc-project.eu
[5] http://www.earthobservatory.eu/
[6] http://planet-data.eu
[7] http://geoknow.eu
[8] http://euclid-project.eu
[9] http://wifo5-03.informatik.uni-mannheim.de/latc/toollibrary/
[10] http://latc-project.eu/platform
[11] http://stack.linkeddata.org/
[12] http://planet-data.eu/planetdata-tool-catalogue
[13] http://www.fluidops.com/information-workbench/
[14] http://ckan.org/

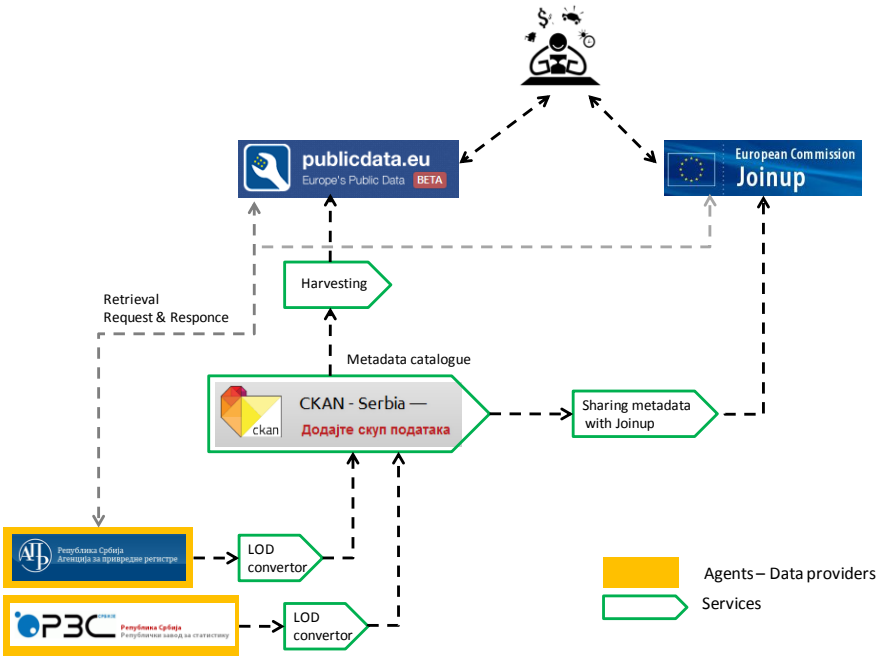

Fig. 1. Government Use Case.

Table 1

Overview of scenarios

Goal	Scenario	Benefits / Expected added value
Metadata management	Code lists - creating and maintaining	Standardization on the metadata level (a) will allow harmonization of specific concepts and terminology, (b) will improve interoperability, and (c) will support multilinguality in statistical information systems across Europe
Export	Export to different formats	Data Exchange with other semantic tools, as well as other commonly used spreadsheet tool e.g. Microsoft Excel.
RDF Data Cube - Extraction, Validation and Initial Exploration	CSV Data Extraction	Standardization of the extraction (CSV2DataCube, XML2DataCube, SDMX2RDFDataCube) process
	XML Data Extraction	
	SDMX-ML 2 RDF/XML Extraction	
	RDF Data Cube Quality Assessment (validation and analysis of integrity constraints)	Building well-formed RDF Data Cubes, where statistical data has been assigned an unique URI, meaning and links to similar data. This approach facilitates search and enables re-use of public statistical data. The well-formed RDF Data Cubes satisfy a number of integrity constraints and contain metadata thus enabling automation of different operations (exchange, linking, exploration)
RDF Data Cube - Transformation, Exploratory Analysis and Visualization	Merging RDF Data Cubes	Data fusion i.e. creation of a single dataset and different graphical charts that supports the exploratory analysis (e.g. indicator comparison)
	Slicing RDF Data Cubes	Facilitate creation of intersections in multidimensional data
	Visualization of RDF Data Cubes	Efficient analysis and search for trends in statistical data
Interlinking	Code lists - Interlinking	Assigning meaning, improved interoperability of data with similar governmental agencies
	CSV Data Extraction and Reconciliation with DBpedia	Assigning meaning
Publishing	Publishing to CKAN	Increased transparency, improved accessibility of statistical data

Analysing the Linked Data software frameworks that meet the needs of the official statistical production process, we realized that there are only few tools that work with RDF Data-Cube-compliant datasets[15].With the Linked SDMX project[16], the authors [8] provide XSLT 2.0 templates and scripts to transform Generic SDMX 2.0 data and metadata to RDF/XML using the RDF Data Cube and related vocabularies for statistical Linked Data. Conversion of statistical data from CSV into Linked Data format is possible with *CSVImport*[17] and T*abLinker*[18]. Exploration and visualization of RDF Data-Cubes are main functionalities of *CubeViz*[19] and *Tabels* tool[20]. In the LD-Cubes project[21] framework, the authors [9] have developed the Linked Data Cubes Explorer for statistical dataset analysis.

2.2. Linked Data publication process for statistical data

The work on the LOD2 Statistical Workbench was motivated by the need to support the process of publishing statistical data in the RDF format using common vocabularies such as the RDF Data Cube [10]. The aim here was to provide support for performing 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 using a LOD publication strategy and respective metadata about the RDF data cube within a selected portal (i.e. a CKAN instance).

The potential benefits of converting statistical data into Linked Data format were studied through several scenarios for the *National Statistical Office* use case [11], see Table 1.

3. Application Architecture and Scenarios

The development of the LOD2 Statistical Workbench[22] is based upon the LOD2 Demonstrator[23], an interface to explore and use all the different stack tools in an integrated way [2]. The LOD2 Stack components that have been integrated into the LOD2 Statistical Workbench are listed in the Appendix, as well as presented in Figure 2. The LOD2 Statistical Workbench also introduces a number of new components such as the *RDF Data Cube Validation tool* [12], *RDF Data Cube Slicing tool*, *RDF Data Cube Merging tool*, *LOD2 authentication component*, *LOD2 provenance component* and *CKAN Publisher*. All components in the LOD2 Stack act upon RDF data and are able to communicate via SPARQL with the central system-wide RDF quad store (i.e. OpenLink *Virtuoso* RDF Triple store). This way of communicating ensures a very loose coupling of the components in the stack.

The software environment features intuitive graphical user interface where the components are organized and grouped within the five topics named *Manage Graph*, *Find more Data Online*, *Edit & Transform*, *Enrich Datacube*, and *Present & Publish*.

3.1. Import features

The LOD2 Statistical Workbench is a framework for managing Linked Data stored in the RDF Data Cube format. However, it supports importing data from CSV and XML files. The CSV2RDF component[24] allows the end users to transform tabular data from a CSV file into a multidimensional RDF Data Cube. On the other hand, *LODGrefine*[25] can be used to import all kinds of structured formats including CSV, ODS and XSL(X) and transform them to RDF graphs which can be based on arbitrary vocabularies.

[15] http://www.w3.org/2011/gld/wiki/Data_Cube_Implementations
[16] https://github.com/csarven/linked-sdmx
[17] https://github.com/AKSW/csvimport.ontowiki
[18] https://github.com/Data2Semantics/TabLinker
[19] http://aksw.org/Projects/CubeViz
[20] http://idi.fundacionctic.org/tabels/
[21] http://www.linked-data-cubes.org

[22] http://demo.lod2.eu/lod2statworkbench
[23] http://demo.lod2.eu/lod2demo
[24] https://github.com/AKSW/csvimport.ontowiki
[25] http://code.zemanta.com/sparkica/

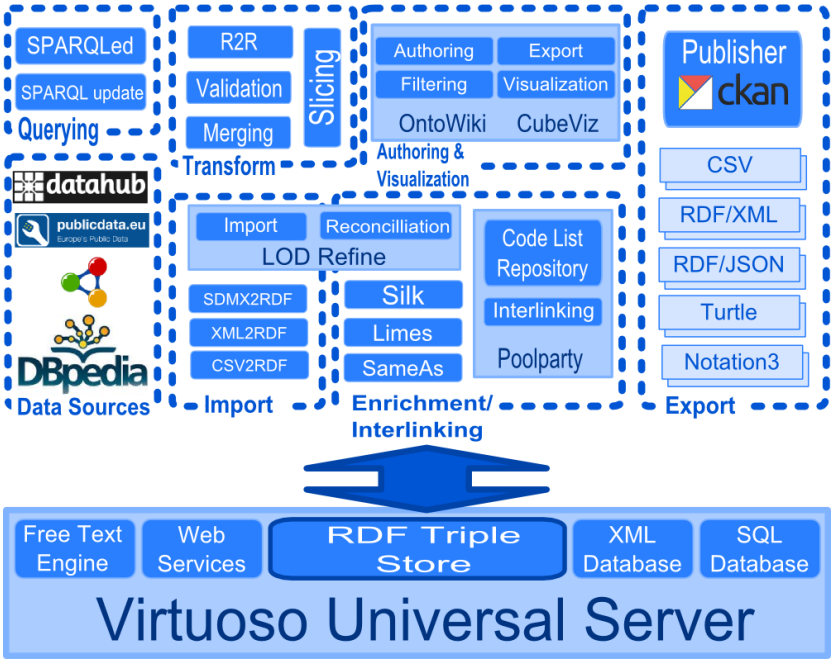

Fig. 2. LOD2 Statistical Workbench - Application architecture.

The users have possibilities to pass XML data as input to the XSLT processor and transform into RDF, however custom XSLT scripts are needed that take common vocabularies (RDF Data Cube, SDMX-RDF, SKOS, Dublin Core Terms, VoID) into consideration. Additionally, using the *Find more Data Online* submenu, the user is able to find and import more / similar data into the local RDF store using the respective tool of Statistical Workbench.

3.2. Semantic integration and storage

Linked Data applications are based on server platforms that enable RDF triple storage, semantic data integration and management, semantic interoperability based on W3C standards (XML, RDF, OWL, SOA, WSDL, etc). The *Virtuoso* Universal Server[26] is used for this purpose in the LOD2 Statistical Workbench.

3.3. RDF Data Cube transformation features

Specialized components have been developed to support the most common operations for manipulating statistical data such as merging datasets, creating slices and data subsetting (*Edit & Transform* submenu). As each dataset defines components (e.g. dimensions used to describe the observations), the merging algorithm checks the adequacy of the input datasets for merging and compiles a new RDF Data Cube to be used for further exploration and analysis. Additionally, the slicing component can be used to group subsets of observations where one or more dimensions are fixed. This way, slices are given an identity (URI) so that they can be annotated or externally referenced, verbosity of the data set can be reduced because fixed dimensions need only be stated once, and consuming applications can be guided in how to present the data.

3.4. RDF Data Cube validation

The RDF Data Cube Validation tool [12] supports

[26] http://www.openlinksw.com/wiki/main/

the identification of possibly not well-formed parts of an RDF data cube. The therein integrated analysis process consists mostly of integrity constraints represented as SPARQL queries published by the W3C[27]. The validation operation is applicable at several steps in the Linked Data publishing process e.g. on import / extraction / transformation from different sources or after fusion and creation of new RDF Data Cubes.

3.5. Authoring, querying and visualization

The *OntoWiki*[28] authoring tool facilitates the authoring of rich semantic knowledge bases, by leveraging Semantic Wiki technology, the WYSIWYM paradigm (What You See Is What You Mean [13]) and distributed social, semantic collaboration and networking techniques.

CubeViz[29], an extension of OntoWiki, is a facetted browser and visualization tool for statistical RDF data. It facilitates the discovery and exploration of RDF data cubes while hiding its complexity from users.

In addition to using the browsing and authoring functionality of *OntoWiki*, advanced users are able to query the data directly (SPARQL) using one of the following offered SPARQL editors: *OntoWiki* query editor, Sindice's *SparQLed* component [30] and the OpenLink *Virtuoso* SPARQL editor.

3.6. Enrichment and interlinking

Linked Data publishing isn't just about putting data on the web, but also about creating links, so that a person or machine can explore the web of data. Therefore, the enrichment and interlinking features are very important as a pre-processing step in integration and analysis of statistical data from multiple sources. LOD2 tools such as *SILK*[31] and *Limes*[32] facilitate mapping between knowledge bases, while *LODGRefine* can be used to enrich the data with descriptions from DBpedia or reconcile with other information in the LOD cloud.

PoolParty[33] allows users to create their own high quality code lists and link the concepts therein to external sources as well. Once the code lists have been established, they can be reused as dimension values in Data Cubes or linked to Cubes that have been created separately.

3.7. Export and Linked Data publishing

The LOD2 Statistical Workbench export features are reachable via the *Manage Graph* and *Present & Publish* submenus. The *Manage Graph* option allows exporting of a graph with all its content in RDF/XML, RDF/JSON, Turtle, Notation 3. *CubeViz* supports subsetting of the data and extraction of a portion that is interesting for further analysis in CSV and RDF/XML format. The *Publisher* component aims at automating the upload and registration of new data with existing CKAN instances.

3.8. Application Scenarios

In order to illustrate the use of the LOD2 Statistical Workbench for different data management operations, we have provided online tutorials[34] for the scenarios summarized in Table 2.

4. LOD2 Statistical Workbench in use

4.1. The RDF Data Cube vocabulary

A statistical data set comprises a collection of observations (see Figure 3) made at some points across some logical space. Using the RDF Data Cube vocabulary, 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.

The collection can must be characterized by a set of dimensions (`qb:DimensionProperty`) that define what the observation applies to (e.g. time `rs:time`, observed sector `rs:obsSector`, country `rs:geo`) along with metadata describing what has been measured (e.g. economic activity, prices) through *measurements*. Optionally additional information can be

[27] http://www.w3.org/TR/vocab-data-cube/#wf-rules
[28] http://aksw.org/Projects/OntoWiki.html
[29] http://aksw.org/Projects/CubeViz
[30] http://sindicetech.com/sindice-suite/sparqled/
[31] http://www4.wiwiss.fu-berlin.de/bizer/silk
[32] http://aksw.org/Projects/LIMES
[33] http://www.poolparty.biz
[34] http://wiki.lod2.eu/display/LOD2DOC/LOD2+Statistical+Workbench

provided on how the observation or cube was measured and how the observations are expressed through the use of *attribute* (qb:AttributeProperty) elements (e.g. units, multipliers, status) .

We can think of a statistical data set as a multi-dimensional space, or hyper-cube, indexed by those a set of 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.

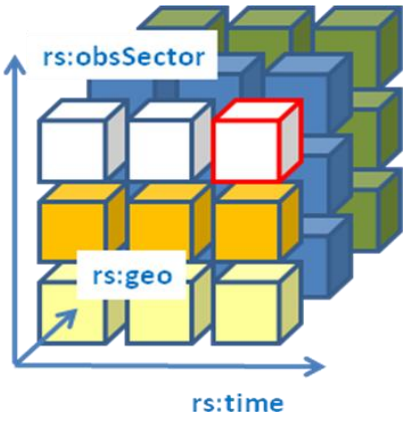

Fig. 3. RDF Data Cube – graphical respresentation.

The qb:dataSet property (see excerpt below) indicates that a specific qb:Observation instance is a part of a dataset. In this example, the primary measure, i.e. observation value (represented here via sdmx-measure:obsValue), is a plain decimal value. To define the units the observation in question is measured in, the sdmx-attribute:unitMeasure property which corresponds to the SDMX-COG concept of UNIT_MEASURE was used. In the example, it is an ESA95 code [14], MIO_NAT_RSD, corresponding to *millions of national currency* (Serbian dinars). The values in the time and location dimensions (rs:geo and rs:time), indicate that the observation took place in the Republic of Serbia (geographic region code RS), and in 2003 (time code Y2003), respectively.

Each data set has a set of structural metadata (see Table 2). These descriptions are referred to in SDMX and the RDF Data Cube Vocabulary as Data Structure Definitions (DSD). Such DSDs include information about how concepts are associated with the measures, dimensions, and attributes of a data cube along with information about the representation of data and related metadata, both identifying and descriptive (structural) in nature. DSDs also specify which code lists provide possible values for the dimensions, as well as the possible values for the attributes, either as code lists or as free text fields. A data structure definition can be used to describe time series data, cross-sectional and multidimensional table data. Because the specification of a DSD is independent of the actual data that the data cube is about, it is often possible to reuse a data structure definition over multiple data cubes.

.

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

<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 .
```

Table 2
Sample data structure

Dimension or Attribute or Measure	Concept description	Identifier	Code list
Dimension	Geographical region	`rs:geo`	`cl:geo`
Dimension	Time	`rs:time`	`cl:time`
Dimension	Economic activity	`rs:activityNACEr2`	`cl:nace_rev2`
Attribute	Unit of measurement	`sdmx-attribute:unitMeasure`	`cl:esa95-unit`
Measure	Observed value	`sdmx-measure:obsValue`	

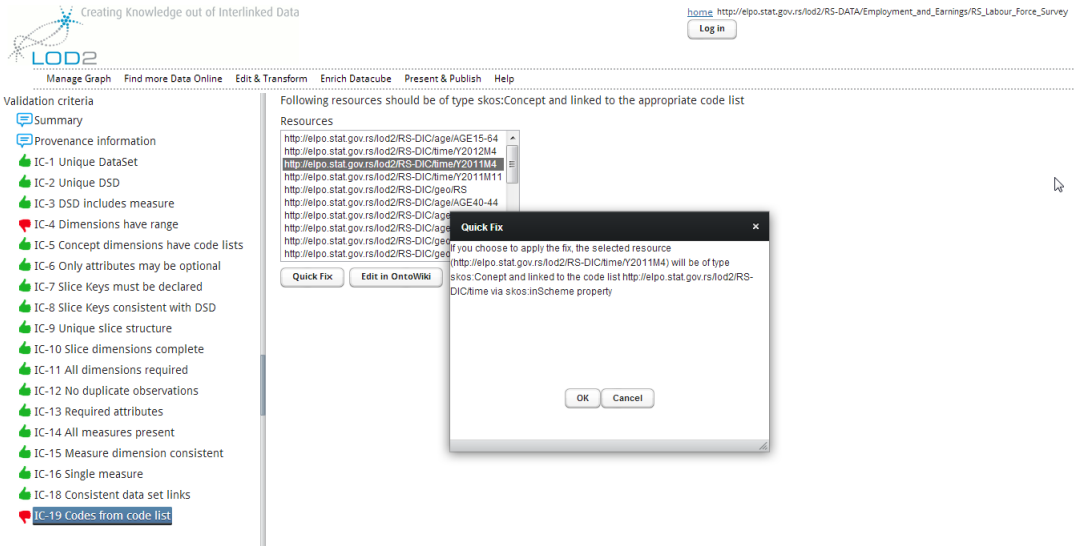

Fig. 4. RDF Data Cube – quality assessment.

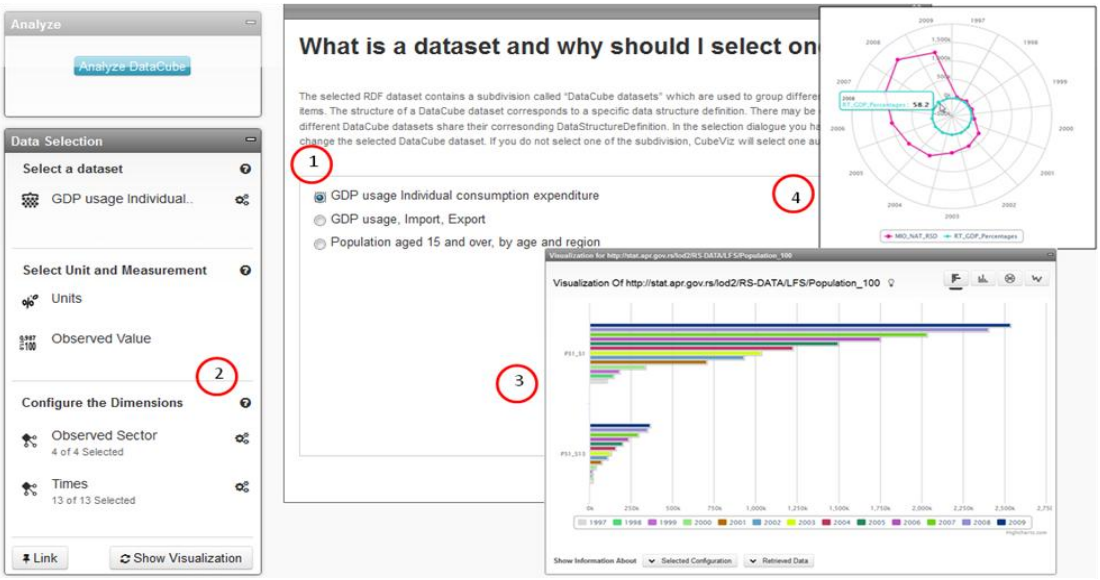

Fig. 5. RDF Data Cube – exploration and analysis.

4.2. Example 1: Quality assessment of RDF Data Cubes

Prior to publishing RDF data on an existing CKAN and thus enabling other users to download and exploit the data for various purposes, every dataset should be validated to ensure it conforms to the RDF Data Cube model. The data validation step is covered by the LOD2 Tool Stack, i.e. through the following software tools:

- The *RDF Data Cube Validation Tool* [12];
- The *CubeViz* tool for visualization of RDF Data Cubes [15].

The *RDF Data Cube Validation Tool* aims at speeding-up the processing and publishing of Linked Data in RDF Data Cube format. It's main use is validating the integrity constraints defined in the RDF Data Cube specification. It works with the *Virtuoso* Universal Server as a backend and can be run from the LOD2 Statistical Workbench environment[35].

The main benefits of using this component are improved understanding of the RDF Data Cube vocabulary and automatic repair of identified errors. Figure 4 shows the component in action: the user can select from the list of criteria corresponding to integrity constraints on the left side, while the results of analysis are shown on the right. A list of resources that violate the constraint, an explanation about the problem, and if possible, a quick solution to the problem is offered to the user. Once an RDF Data Cube satisfies the standard integrity constraints, it can be visualized with the *CubeViz* tool. A more detailed quality analysis scenario is included in the LOD2 Stack Documentation[36].

4.3. Example 2: Filtering, visualization and export of RDF Data Cubes

The *CubeViz* facetted browser and visualization tool can be used to filter observations to be visualized in charts interactively. Aggregation methods supported are SUM, AVG, MIN and MAX. Step-by-step interactions with the tool in an exploration session can be explained as follows:

- Select one from the available datasets in the graph;
- Choose the observations of interest by using the available dimensions;
- Visualize the statistics by using groups, or
- Visualize the statistics in two different measure values (millions of national currency and percentages).

For a more detailed explanation of the filtering and export options we refer to the LOD2 Stack Documentation[37].

4.4. Example 3: Merging RDF Data Cubes

Merging[38] is an operation of creating a new dataset (RDF Data Cube) that compiles observations from the original datasets (two or more), and additional resources (e.g. data structure definition, component specifications) that will allow visualization of the newly created dataset. In order to obtain meaningful charts the observed phenomena (i.e serial data) have to be described on the same granularity level (e.g. year, country) and expressed in same units of measurement (e.g. euro, %). Therefore alignment of the code lists used in the input data is necessary before the merging operation is performed.

5. Challenges of broader adoption

Linked Data principles have been introduced into a wide variety of application domains, e.g. publishing statistical data and interpretation of statistics [16], improving tourism experience [17], pharmaceutical R&D data sharing [18], crowdsourcing in emergency management [19], etc.

A few years ago, our analysis of the adoption of Semantic Web technologies by enterprises [20] has shown that companies can achieve benefits such as data share and re-use (57%), improved search (57%),

[35] http://fraunhofer2.imp.bg.ac.rs/lod2statworkbench/

[36] LOD2 Documentation Wiki, http://wiki.lod2.eu/display/LOD2DOC/RDF+Data+Cube+Quality+Assessment

[37] LOD2 Documentation Wiki, http://wiki.lod2.eu/display/LOD2DOC/Data+Subsetting+and+Export+Scenario

[38] LOD2 Documentation Wiki, http://wiki.lod2.eu/display/LOD2DOC/Eurostat+Merge+and+Enhance+Scenario

incremental modelling (26%), explicit content relation (24%), identifying new relationships (17%), dynamic content generation (14%), personalization (10%), open model (12%), rapid response to change (10%), reduced time to market (5%), and automation (5%). The features provided in the LOD2 Statistical Workbench address the following challenges: data share and re-use, improved search, explicit content relation, identifying new relationships, open model and automation. The LOD2 Statistical Workbench is a customized version of the LOD2 Stack for statistical data that can be used by both publishers and consumers of Linked Data such as national statistical offices (institutes), national banks, publication offices, etc. Herein, we point to three government institutions, as potential adopters of the LOD2 Stack.

5.1. Use case 1: Statistical Office of the Republic of Serbia (SORS)

In the course of the LOD2 PUBLINK 2011 activities, the SORS public data was integrated into the LOD cloud via the Serbian CKAN[39]. The Serbian CKAN is a metadata repository to be used for dissemination purposes by Serbian national institutions (see Figure 1). Maintenance activities include identifying changes in the dissemination data (new public data, changes on metadata level) and fixing the mapping process (from XML to RDF) accordingly. The SORS is in the process of adopting the LOD2 Statistical Workbench that will allow the users to automatically publish data (in the existing and new packages[40]) to the Serbian CKAN.

5.2. Use case 2: Business Registers Agency

In the course of the LOD2 PUBLINK 2012 activities, example data from the *Regional Development Measures and Incentives* Register was triplified using the LOD2 Statistical Workbench and registered with the Serbian CKAN. The visualization of the data with the *CubeViz* tool is also reachable via the Serbian CKAN resources[41].

[39] http://rs.ckan.net
[40] http://rs.ckan.net/group/rzs
[41] http://rs.ckan.net/dataset/apr-register-of-regional-development-measures-and-incentives

5.3. Use case 3: National Bank

Most of the public data coming from the Serbian National Bank and registered with the Serbian CKAN is in Excel format. In the LOD2 framework, the data[42] will be transformed to CSV, and after then it will be automatically harvested and triplified with the CSV to RDF Mass Conversion extension for the PublicData.eu portal [21].

5.4. Challenges faced by early adopters

In the course of the LOD2 Statistical Workbench development, questions have risen such as

- Is the integration of the LOD2 Stack into a coherent solution for a specific domain possible ?
- What is configurable ?
- How that can be configured ?

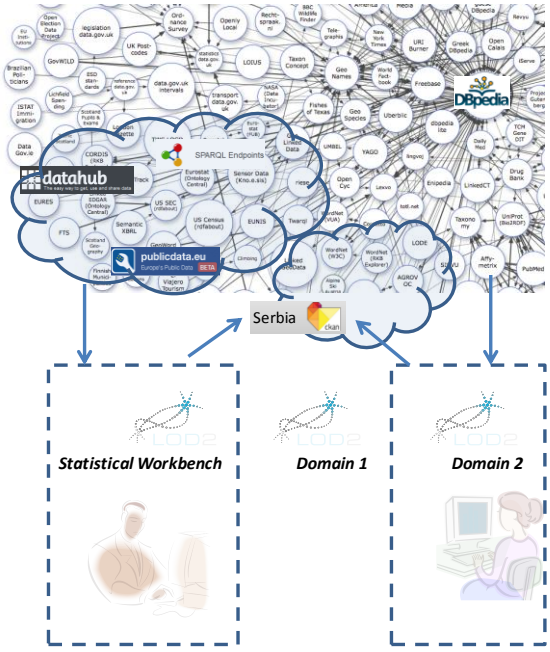

Fig. 6. Adoption of the LOD2 Stack for different domains.

The first results of using the LOD2 Statistical Workbench have shown that LOD2 components are well integrated in a unified solution for conversion, exploration and publishing of statistical data into Linked Data format [11]. Configuration of the LOD2 stack for the statistical domain required:

[42] http://rs.ckan.net/group/nbs

- Specifying the set of components that will be used;
- Identifying the set of vocabularies that are applicable for the domain ;
- Customizing the graphical user interface for the preselected group of users;
- Configuring (enabling/restricting) user access to different on-line services.

Once the integration is proved within one domain (in our case the Statistical Office domain), it will work for other domains as well e.g. drug data, tourism data, etc. General instructions how to install and use the LOD2 Stack are given in the LOD2 Documentation[43].

6. Conclusions and Future work

The LOD2 Statistical Workbench is a result of the integration effort of three institutions: TenForce, the "Mihajlo Pupin Institute" and the University of Leipzig (Agile Knowledge Engineering and Semantic Web Group). It is based on the state-of-the-art technologies and tools for managing Linked Data delivered by partners of the LOD2 consortium. It contributes to the standardization of the Linked Data processing in statistical data domain in organizations such as national statistical offices (institutes), national banks, publication offices, etc.

The paper discusses the main objectives, the functionalities provided and the possibilities for adoption of the solution by government institutions. The system meets the needs of both publishers and consumers of statistical data and directs the potential of the LOD2 tools to one specific domain. The integrated graphical user interface allows the user intuitive completion of common operations in statistical data processing including integration with RDBMS, import/export in standard formats (e.g. CSV), search for similar data in Linked Data format, editing, transformation and fusion of Linked statistical data, enrichment and reconciliation with DBpedia, navigating through Linked Data modes and visualization and analysis of multidimensional data. Moreover, the LOD2 Statistical Workbench can be integrated with a preselected CKAN based government portal thus enabling the user automatic publishing and cataloguing newly created linked statistical data on local level. By scheduling periodical "harvesting" at an international level (e.g. by PublicData.eu), the data become interrelated with similar data and globally visible.

Challenges that will be addressed in future are related to improving the integration of tools and achieving better user experience. New components from the LOD2 stack will be integrated into the LOD2 Statistical Workbench if they fulfil a business need in the statistical domain. Additional features, including reconciliation with standard code lists and a module for Dataset Structure Definition management are also planned. Methods for manipulation, fusion and visualization of data at different levels of granularity will be improved in order to have solid basis for advanced analysis and visualization of statistical Linked Data.

Appendix

- CKAN - Comprehensive knowledge archive network, http://ckan.org/
- CSVImport – OntoWiki CSV import and transformation extension, https://github.com/AKSW/csvimport.ontowiki
- CubeViz - RDF DataCube Brower, http://aksw.org/Projects/CubeViz.html
- DBpedia Knowledge Base, http://dbpedia.org
- Limes Link discovery framework, http://aksw.org/Projects/LIMES.html
- LOD2Webapi - Graph and Prefix management API
- LODRefine, http://code.zemanta.com/sparkica/
- Mondeca SPARQL Endpoint status, http://labs.mondeca.com/sparqlEndpointsStatus/
- OntoWiki - Generic Data Wiki, http://aksw.org/Projects/OntoWiki.html
- PoolParty - SKOS Taxonomy Editor, http://www.semantic-web.at/poolparty-semantic-information-management
- R2R, http://wifo5-03.informatik.uni-mannheim.de/bizer/r2r/spec/
- SDMX-ML to RDF/XML, https://github.com/csarven/linked-sdmx/
- SemMap Spatial RDF Browser, http://linkedgeodata.org/LGD%20Browser
- Silk Linking Workbench, http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/
- Sindice - the Semantic Web Index, http://sindice.com/

43 http://wiki.lod2.eu/display/LOD2DOC/How+To+Start

- SparQLed - Assisted SPARQL Editor, http://sindicetech.com/sindice-suite/sparqled/
- Virtuoso Universal Server, http://virtuoso.openlinksw.com/
- Xalan - Java XSLT processor, http://xml.apache.org/xalan-j/

References

[1] T. Berners-Lee. Linked Data. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html

[2] SDMX Information Model: UML Conceptual Design (Version 2.0), November 2005, Statistical Data and Metadata Exchange Initiative, http://sdmx.org/docs/2_0/SDMX_2_0%20SECTION_02_InformationModel.pdf

[3] S. Auer, et all. Managing the Life-Cycle of Linked Data with the LOD2 Stack. In: Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Xavier Parreira, J., Hendler, J., Schreiber, G., Bernstein, A., Blomqvist, E. (Eds.) *International Semantic Web Conference 2*, (Book 7650):1-16, Springer, 2012.

[4] M. Martin, C. Stadler, P. Frischmuth, and J. Lehmann. Increasing the Financial Transparency of European Commission Project Funding. *Semantic Web – Interoperability, Usability, Applicability*. Retrieved from http://www.semantic-web-journal.net/system/files/swj435.pdf

[5] "Decision no 922/2009/EC of the European parliament and of the Council of 16 September 2009 on Interoperability Solutions for European Public Administrations (ISA)", *Official Journal of the European Union, L 260/20* (3.10.2009). Retrieved from http://ec.europa.eu/isa/documents/isa_lexuriserv_en.pdf

[6] P. Haase, C. Hütter, M. Schmidt, and A. Schwarte. The Information Workbench as a Self-Service Platform for Developing Linked Data Applications. *WWW 2012 Developer Track*, April 18-20, 2012, Lyon, France. Retrieved from http://www2012.wwwconference.org/proceedings/nocompanion/DevTrack_016.pdf

[7] V. Janev, U. Milošević, M. Spasić, J. Milojković, S. Vraneš. Linked Open Data Infrastructure for Public Sector Information: Example from Serbia. In: S. Lohmann, T. Pellegrini (Eds.) Proceedings of the 8th Int. Conference on Semantic Systems September (I-SEMANTICS 2012, Messecongress Graz, Austria, September, 5-7, 2012). CEUR Workshop Proceedings (CEUR-WS.org). ISBN: 978-1-4503-1112-0, 2012

[8] S. Capadisli, S. Auer, and A. Ngonga Ngomo. Linked SDMX Data: Path to high fidelity Statistical Linked Data. *Semantic Web – Interoperability, Usability, Applicability*. Retrieved from http://www.semantic-web-journal.net/content/linked-sdmx-data

[9] B. Kämpgen, S. O'Riain, and A. Harth. Interacting with Statistical Linked Data via OLAP Operations. In: C. Unger, P. Cimiano, V. Lopez, E. Motta, P. Buitelaar, R. Cyganiak (Eds) *Proceedings of the Workshop on Interacting with Linked Data* (ILD 2012, Workshop co-located with the 9th Extended Semantic Web Conference May 28, 2012, Heraklion, Greece). CEUR Workshop Proceedings, pp. 36-49, http://ceur-ws.org/Vol-913/ILD2012.pdf

[10] R. Cyganiak, and D. Reynolds. The RDF Data Cube vocabulary, http://www.w3.org/TR/vocab-data-cube/, 2013.

[11] V. Janev, U. Milošević, M. Spasić, S. Vraneš, J. Milojković, and B. Jireček. Integrating Serbian Public Data into the LOD Cloud. In: Budimac, Z., Ivanović, M., Radovanović, M. (Eds.) *Proceedings of the 5th Balkan Conference in Informatics* (BCI'12, September 16–20, 2012, Novi Sad, Serbia). New York: ACM International Conference Proceeding Series vol. 641, pp.94-99., 2012.

[12] V. Janev, V. Mijović, and S. Vraneš. LOD2 Tool for Validating RDF Data Cube Models. *Web Proceedings of the 5th ICT Innovations Conference*, Ohrid, Macedonia, September 12-15, 2013. Retrieved from http://ict-act.org/proceedings/2013/htmls/papers/icti2013_submission_01.pdf

[13] A. Khalili, and S. Auer. WYSIWYM authoring of structured content based on Schema.org. In: Lin, X., Manolopoulos, Y., Srivastava, D., Huang, G. (Eds) *Web Information Systems Engineering* – WISE 2013. Lecture Notes in Computer Science vol. 8181, pp. 425-438, 2013. http://dx.doi.org/10.1007/978-3-642-41154-0_32

[14] SDMX. 2009. *SDMX Content-oriented Guidelines: Cross-domain code lists.* (2009). Retrieved from http://sdmx.org/wpcontent/uploads/2009/01/02_sdmx_cog_annex_2_cl_2009.pdf

[15] 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.

[16] H. Paulheim. Generating Possible Interpretations for Statistics from Linked Open Data. In: *Proceedings of the 9th Extended Semantic Web Conference* (ESWC 2012, Heraklion, Crete, Greece). The Semantic Web: Research and Applications, Lecture Notes in Computer Science Volume 7295, pp. 560-574, 2012.

[17] M. Sabou, A. M. P. Brasoveanu, and I. Arsal. Supporting Tourism Decision Making with Linked Data. In: *Proceedings of the 8th Int. Conference on Semantic Systems* (I-SEMANTICS/I-CHALLENGE, Graz, Austria). ACM International Conference Proceedings Series, pp. 201-204, 2012.

[18] M. Samwald, et all. Linked open drug data for pharmaceutical research and development. *Journal of Cheminformatics* 3:19, 2011, http://www.jcheminf.com/content/3/1/19

[19] J. Ortmann, M. Limbu, D. Wang, and T. Kauppinen. Crowdsourcing Linked Open Data for Disaster Management. *Terra Cognita 2011 Workshop*, In Conjunction with the 10th International Semantic Web Conference, (ISWC 2011, Bonn, Germany), 2011.

[20] V. Janev, and S. Vraneš. Applicability assessment of Semantic Web technologies. *Information, Processing and Management* 47(4): 507-517, 2011.

[21] I. Ermilov, S. Auer, and C. Stadler. User-driven Semantic Mapping of Tabular Data. In: *Proceedings of the I-Semantics 2013* (September 4–6, 2013, Graz, Austria). ACM International Conference Proceedings Series.