Publishing and consuming Linked Open Data with the LOD2 Statistical Workbench

Valentina Janev, Vuk Mijović, Uroš Milošević, Sanja Vraneš

Institute "Mihailo Pupin", University of Belgrade, Volgina 15, 11060 Belgrade, Serbia

Abstract. Statistical data is often used as the foundations for policy prediction, planning and adjustments, and therefore has a significant impact on the society (from citizens to businesses to governments). Linked Data paradigm has opened new possibilities and perspectives for the process of collecting and monitoring socio-economic indicators. The paper introduces 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. Using an illustrative case study of the Statistical Office of the Republic of Serbia, the paper gives an overview of possible scenarios and shows examples of its use. The first results indicate that wider adoption of the Linked Data tools in practice can be foreseen.

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

1. Problem statement

Data is open if it is technically open (available in a machine-readable standard format, which means it can be retrieved and meaningfully processed by a computer application) and legally open (explicitly licensed in a way that permits commercial and non-commercial use and re-use without restrictions)[1]. Although Open Government Data policies have spread fast (the number of open data initiatives has grown from two to over 300)[1], the availability of truly open data remains low, with less than 7% of the dataset surveyed in the Open Data Barometer published both in bulk machine-readable forms, and under open licenses [2]. Figure 1 shows, for instance, three types of open data portals:

- Open Data Portal in Country X contains data and metadata descriptions, but does not provide DCAT[2] support for harmonization of portal/catalog with similar data portals/catalogues. If an open data portal does not exist on country level, publishers can use the cross country portals for publishing data, e.g. the Engage[3] portal,
- Open Data Portal in Country Y contains data and metadata descriptions, as well as a Linked Data SPARQL endpoint, but it is isolated , that is, it is not integrated at the international level,
- Open Data Portal in Country Z is CKAN based, meaning that it can be easily harvested by other metadata catalogs on international level.

Assuming that a new user of open data or a developer of innovative services (see right side of Figure 1) has identified the resources to be re-used, the questions that could be raised are:

- Is the open data ready for exploration? Is the metadata complete? What about the granularity? Do we have enough information about the domain/region the data is describing?

---

[1] http://data.worldbank.org/about/open-government-data-toolkit/knowledge-repository

[2] DCAT, http://www.w3.org/TR/vocab-dcat/, is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web.

[3] http://www.engagedata.eu/

• Is it possible to fuse heterogeneous data and formats from different publishers and what are the necessary steps? Are there standard services for querying government portals?

Linked Data[4] paradigm [3,4] has been utilized recently in order to achieve publishing and linking of datasets together through references to common concepts. The standard for the representation of the information that describes those entities and concepts is RDF.
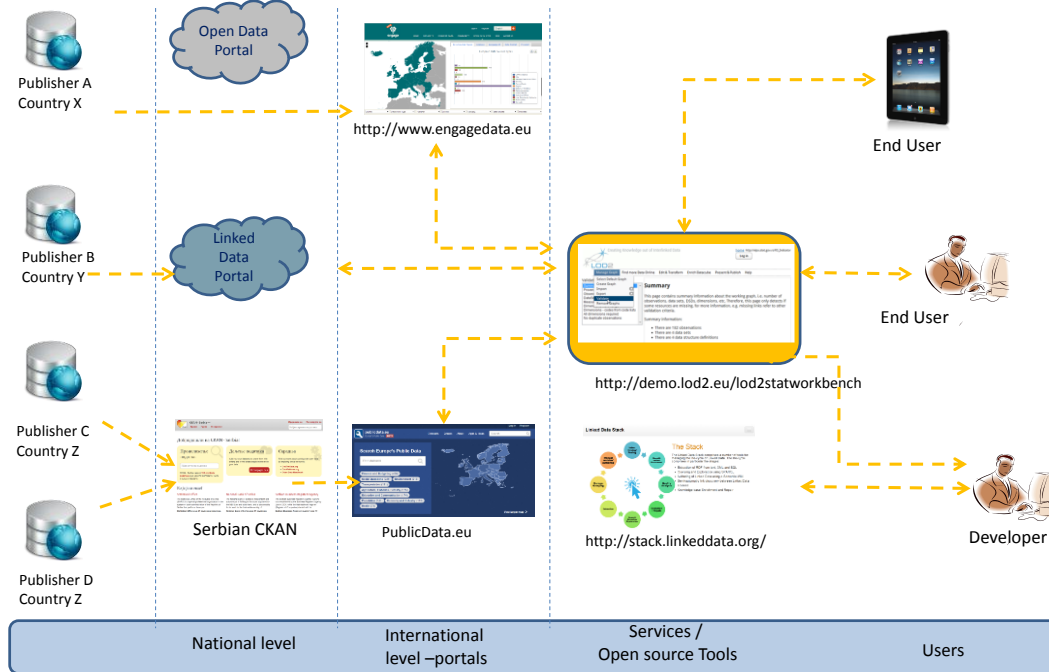

Fig. 1. Exploring and consuming open data.

2. Linked Data Life Cycle

The Linked Data Life Cycle consists of a set of steps or phases for:
• efficient transformation / conversion of traditional data stores (e.g. CSV, XML, relational databases) into linked, machine readable formats;
• managing triple stores containing Linked Data in RDF format;
• enriching, interlinking and adding meaning to data;
• quality assessment, evolution and repair of data;
• publishing data and respective metadata using the LOD publication strategy;
• searching, browsing, visualization, exploration and re-use of data.

Aimed at providing a set of professional tools for accessing, manipulating, exploring and publishing statistical data in a Linked Data format, the LOD2[5] consortium developed the LOD2 Statistical Workbench[6]. The work was motivated also by the number of open datasets with statistical data and the need to support the process of managing Linked Data modelled with the RDF Data Cube vocabulary[7]. 1:n relationship can be established between the tools and the Linked Data Life Cycle, meaning that one tool can support few phases of the process (e.g. LODRefine[8] can be used for extraction, enrichment and reconciliation) .

The potential benefits of converting statistical data into Linked Data format were studied through several scenarios for the National Statistical Office use case [5], see Table 1.

---

[4] The term Linked Data here refers to a set of best practices for publishing and connecting structured data on the Web.
[5] http://lod2.eu , FP7 project ("Creating knowledge out of interlinked data")
[6] http://fraunhofer2.imp.bg.ac.rs/lod2demo-test/stat or http://demo.lod2.eu/lod2statworkbench
[7] http://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/
[8] http://code.zemanta.com/sparkica/

Visualization with CubeViz
Geo-Spatial exploration
Publish to CKAN
Publish to Sindice

Creating Knowledge out of Interlinked Data

LOD2

home no current graph selected

Log in

Manage Graph   Find more Data Online   Edit & Transform   Enrich Datacube   Present & Publish   Help

Select Default Graph
Create Graph
Import
Export
Validate
Remove Graphs

Interlinking dimensions (Silk)
Data enrichment and reconciliation (LODRefine)
Interlinking with Limes
Interlinking with SameAs

hich comprises a
cle of Linked Data.
e stages

nd SQL
- Querying and Exploration using SPARQL
- Authoring of Linked Data using a Semantic Wiki
ni-automatic link discovery between Linked Data
rces
wledge-base Enrichment and Repair

cess tools for each of these stages using the menu

Stack is developed by the LOD2 project consortium
comprising 15research groups and companies. The LOD2
project is co-funded by theEuropean Commission within the
7th Framework Programme (GA no. 257934).

You can find further information about the LOD2 Stack at
http://stack.lod2.eu and the LOD2 project at http://lod2.eu.

Edit Graph (OntoWiki)
Edit Code Lists (PoolParty)
Merge datasets
Slice datasets
Transform and Update Graph (SPARQL Update Endpoint)
SparQLed - Assisted Querying
PoolParty Code Lists SPARQL endpoint

Inter-
linking/
Fusing

Classifi-
cation/

Manual
revision/
authoring

Storage/
Querying

Extraction

Search/
Browsing/
Exploration

Evolution/
Repair

Load RDF data from publicdata.eu
Load RDF data from Data Hub
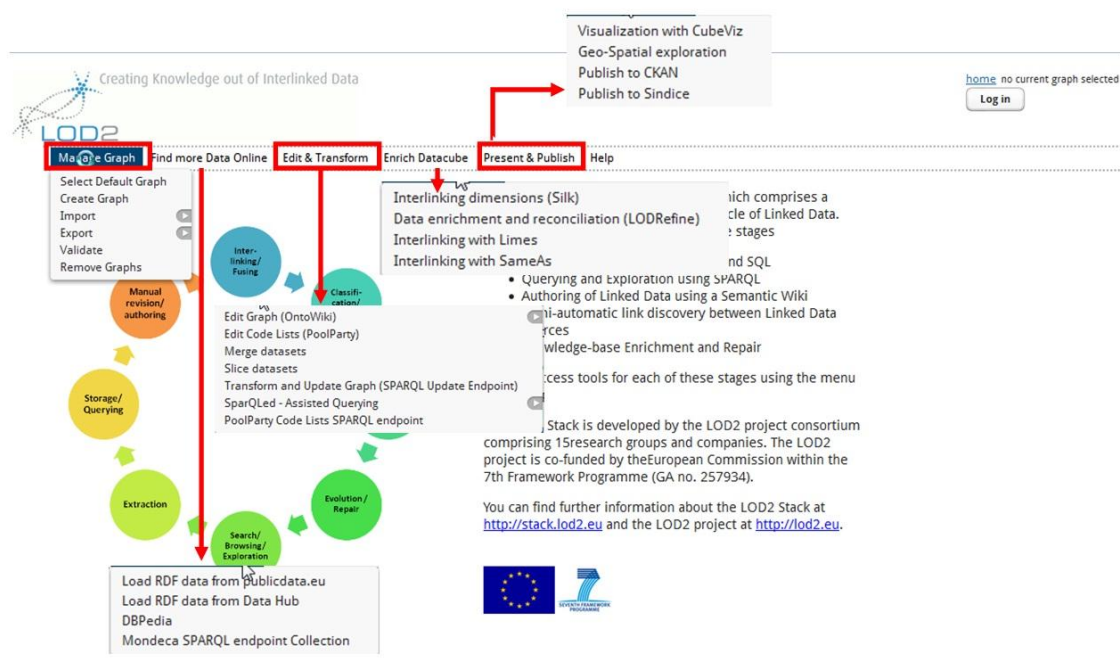DBPedia
Mondeca SPARQL endpoint Collection

Fig. 2. Linked Data Life Cycle as it is supported by LOD2 Statistical Workbench.

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 multi-linguality in statistical information systems across Europe
Export functionalities	Export to different formats	Data Exchange with other semantic tools, as well as other commonly used spread-sheet 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

3. The SORS PUBLINK project

Linked Data principles have been introduced into a wide variety of application domains, e.g. publishing statistical data and interpretation of statistics [6], improving tourism experience [7], pharmaceutical R&D data sharing [8], crowdsourcing in emergency management [9], etc.

In the course of the LOD2 activities, the Institute Mihajlo Pupin (IMP) used the LOD2 Statistical Workbench to publish statistical data from the Statistical Office of the Republic of Serbia (SORS) in Linked Data format via

the Serbian CKAN[9]. The Serbian CKAN is a metadata repository to be used by Serbian national institutions for dissemination purposes. The Statistical Office of the Republic of Serbia is in the process of adopting the LOD2 Statistical Workbench for automatic publishing data of data (in the existing and new packages[10]) to the Serbian CKAN. Herein, we would like to establish a relation between the activities carried out in the SORS project and the *Best Practices for Publishing Linked Data[11]* document.

3.1. STEP #1 PREPARE STAKEHOLDERS

Legal framework for starting the cooperation between IMP and SORS was the LOD2 2011 PUBLINK call[12]. SORS applied for LOD Consultancy after a couple of meetings where the IMP team had presented the LOD vision to the SORS top management.

3.2. STEP #2 SELECT A DATASET

SORS as a special professional organization in the system of state administration of the Republic of Serbia that publishes information on monthly, quarterly and yearly basis in the form of open, downloadable free of charge documents in PDF format, while raw data with short and long-term derived indicators are organized in a central statistical database[13]. The statistical database was selected as a respectable source of public data that should be made publicly available in Linked Data format as well.

3.3. STEP #3 MODEL THE DATA

More information about the modelling and publishing process can be found in the *LOD2 Deliverable 9.5.1 Establishment of the Serbian CKAN[14],* as well as in [5]. The SORS Linked Open Dataset[15] was prepared and published in accordance with the <u>Government Linked Data (GLD) Working Group</u>[16] standards and recommendations.

3.4. STEP #4 SPECIFY AN APPROPRIATE LICENSE

All datasets were published as Other (Public Domain) Licence, because the data was already public and no further restrictions were needed.

3.5. STEP #5 GOOD URIs FOR LINKED DATA

Decisions about the URIs were made with the representatives of the IT department. The SORS code list (prefix `cl:`) and the SORS LOD dataset namespaces for different domains were defined as follows

```
cl:          http://elpo.stat.gov.rs/lod2/RS-DIC/
accounts:    http://elpo.stat.gov.rs/lod2/RS-DATA/National_accounts/dsd/
tu:          http://elpo.stat.gov.rs/lod2/RS-DATA/Tourism/dsd/
```

Additionally a domain for common attributes was defined as

```
rs:          http://elpo.stat.gov.rs/lod2/RS-DIC/rs/
```

3.6. STEP #6 USE STANDARD VOCABULARIES

Standard vocabularies that were used to publish the data and to describe the datasets were:
- The Data Cube RDF vocabulary[17], a core foundation, focused purely on the publication of multi-dimensional data on the Web.

---

[9] http://rs.ckan.net
[10] http://rs.ckan.net/group/rzs
[11] http://www.w3.org/TR/ld-bp/
[12] http://lod2.eu/Article/Call_2011.html
[13] http://webrzs.stat.gov.rs/WebSite/public/ReportView.aspx
[14] http://static.lod2.eu/Deliverables/LOD2_D9.5.1_Serbian_CKAN.pdf
[15] http://elpo.stat.gov.rs/LOD2/rzs.ttl
[16] http://www.w3.org/2011/gld/
[17] http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html

- The Simple Knowledge Organization System[18] is a vocabulary that supports the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading lists and taxonomies within the framework of the Semantic Web.

- VoID – "Vocabulary of Interlinked Datasets"[19], an RDF based schema to describe linked datasets.

Thus, using Turtle syntax, an observation is expressed as follows:

```
<http://elpo.stat.gov.rs/lod2/RS-DATA/Tourism/Tourists_arrivals/data/obs1> a qb:Observation ;

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

3.7. STEP #7 CONVERT DATA

The SORS statistical data in XML form was passed as input to the XSLT processor and transformed into RDF using the aforementioned vocabularies and concept schemes (geographical, time, indicators, tourists' types). Examples of transformation scripts were also published, see http://rs.ckan.net/dataset/rzs-prices-retail-price-indices .

3.8. STEP #8 PROVIDE MACHINE ACCESS TO DATA

The SORS Linked datasets were made available via the Serbian CKAN portal[20]. The CKAN open-source data portal platform provides a powerful API for machine access to RDF data.

3.9. STEP #9 ANNOUNCE NEW DATA SETS

The SORS Linked datasets were announced through the LOD2 blog[21].

3.10. STEP #10 RECOGNIZE THE SOCIAL CONTRACT

Once the data is published, it sould be properly maintained. Maintaining activities include identifying changes in the dissemination data (regular updates of existing datasets, new public datasets, changes on metadata level) and fixing the mapping process accordingly.

4. LOD2 Statistical Workbench in use

In this Section, we would like to point to some functionalities that have been implemented in the Statistical Workbench and are of great importance for manipulating, publishing and re-using Linked Data.

4.1. 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* [10];
  – The *CubeViz* tool for visualization of RDF Data Cubes [11].

---

[18] http://www.w3.org/2004/02/skos/
[19] http://semanticweb.org/wiki/VoiD
[20] http://rs.ckan.net
[21] http://stack.linkeddata.org/stack-in-use/integrating-serbian-public-data-into-the-lod-cloud/

The main benefits of the *RDF Data Cube Validation Tool* are improved understanding of the RDF Data Cube vocabulary and automatic repair of identified errors. The tool points out resources that violate the constraint, provides an explanation about the problem, and if possible, offers a quick solution to the problem. 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.

4.2. 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 (see Figure 2):

– Select one of 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).

4.3. Example 3: Merging RDF Data Cubes

Merging 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 [12].

5. Conclusions

The LOD2 Statistical Workbench 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 first evaluation results showed that the RDF Data Cube vocabulary is mature enough to be used for publishing statistical data as it improves interoperability and allows comparison of data from different statistical sources. Furthermore, the LOD2 tools and technologies yield to establishment of interoperable Open Government Data ecosystem whose benefits are economic, through the identification of new business opportunities, and social, through increased transparency, participation and accountability.
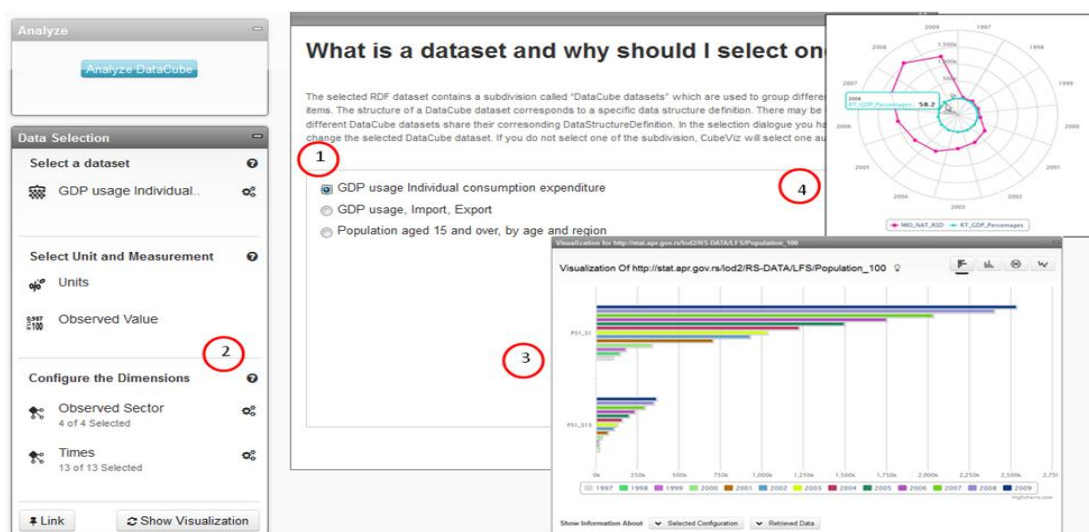

Fig. 3. RDF Data Cube – exploration and analysis.

References

[1] J. M. Alonso. Announcing the Global Open Data Initiative (GODI), World Wide Web Foundation, (June 11, 2013), http://www.webfoundation.org/2013/06/announcing-the-global-open-data-initiative-godi/

[2] *Open Data Barometer - 2013 Global Report*, http://www.opendataresearch.org/dl/odb2013/Open-Data-Barometer-2013-Global-Report.pdf

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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

[9] 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.

[10] 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

[11] 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.

[12] V. Janev, B., Van Nuffelen, V., Mijović, K., Kremer, M., Martin, U., Milošević, S., Vraneš. Supporting the Linked Data publication process with the LOD2 Statistical Workbench, *Semantic Web – Interoperability, Usability, Applicability* (under review, IOS Press, http://www.semantic-web-journal.net/content/supporting-linked-data-publication-process-lod2-statistical-workbench)