Linked Data Approach to the PSI Directive Implementation: Supporting Tools and Lessons Learned

Valentina Janev, *Member, IEEE,* Vuk Mijović, and Sanja Vraneš, *Member, IEEE*

***Abstract* — This paper addresses the technical aspects and challenges of implementation of the revised European Directive on the Public Sector Information (2013/37/EU), emphasizing the role of Linked Data approach for improved interoperability and re-use. It points to several tools that have been developed at the Mihajlo Pupin Institute to support the implementation process i.e. the Serbian CKAN that can be used for cataloguing of open data, the RDF Data Cube Validation tool and the Exploratory Spatio-Temporal Analysis of Linked Data component for publishing and consuming of statistical data. The paper places a particular emphasis on the lessons learned and proposes recommendations for improving the interoperability of e-government services in Serbia.**

***Keywords* — e-government, Interoperability, Linked Data, Open Data, PSI Directive, RDF, Serbian CKAN.**

I. Introduction

THE Directive on the re-use of Public Sector Information (known as the 'PSI Directive'), which revised the Directive 2003/98/EC and entered into force on 17 July 2013, provides a common legal framework for a European market for government-held data (public sector information). It is built around two key pillars of the internal market: transparency and fair competition, and focuses on the economic aspects of re-use of information rather than on the accessibility of information to citizens [1]. Member States were obliged to transpose the Directive into national legislation by 18 July 2015, however the transposition of the revised PSI Directive into national legislations across Europe is not completed yet.[1]

The PSI Directive is a legislative document and does not specify technical aspects of its implementation. Article 5, point 1 of the PSI Directive says '*Public sector bodies shall make their documents available in any pre-existing format or language, and, where possible and appropriate, in open and machine-readable format together with their metadata. Both the format and the metadata should, in so far as possible, comply with formal open standards*.' Therefore, in July 2014, the Commission published guidelines [2] related to licenses (encourages the use of open licenses), datasets (asks for availability, quality, usability and interoperability of 'high-demand' datasets) and charges where Commission prefers the least restrictive re-use regime possible i.e. limits any charges to the marginal costs incurred for the reproduction, provision and dissemination of documents. The Commission will also facilitate the roll-out of *open data infrastructures* under the Connecting Europe Facility.

This paper primarily refers to the technical aspects related to the implementation of the Directive, especially the issues that directly or indirectly contribute to the *semantic interoperability* of open data. So far the European Commission has conducted several interoperability solutions programmes, where the last one "*Interoperability Solutions for European Public Administrations*" (*ISA*) will be active during the next five years (2016-2020) under the name ISA^2 [5]. *Semantic data interoperability* aims at supporting the implementation of the PSI Directive in the best possible way and "make documents available through open and machine-readable formats together with their metadata, at the best level of precision and granularity, in a format that ensures interoperability, re-use and accessibility" [5].

It is organized as follows. First, Section 2 introduces the Linked Data approach to semantic interoperability of data in public administration across Europe. Using an example from Serbia, Sections 3 analyses the challenges and introduces tools for publishing, flexible integration and visualization of statistical data [6]. Section 4 outlines the lessons learned, while Section 5 concludes the paper.

II. Linked Data Approach

Analyzing the PSI Directive, elements can be extracted related to policies and legislation, software tools and platforms, selection of data for publication / dataset criteria, charging, techniques for opening data, organizational issues, formats, re-use, persistence, data quality issues, documentation of open data and data discoverability [7].

Herein, we will discuss the following technical aspects related to the PSI Directive: techniques for opening data, formats, platforms/tools, and quality issues. We are also pointing to some results of our team, while the tools that

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

All authors are with the "Mihajlo Pupin" Institute, University of Belgrade, Volgina 15, 11060 Belgrade, Serbia (phone: 381-11- 6774024; e-mail: valentina.janev, vuk.mijovic, sanja.vranes@pupin.rs).

[1] https://ec.europa.eu/digital-agenda/en/news/open-data-commission-launches-infringement-cases-due-late-transposition-revised-psi-directive

have been developed in the Institute Mihajlo Pupin in the last four years are briefly presented in the Section III.

A. Open Data Formats

In order to share datasets between users and platforms, the datasets need to be accessible (regulated by licence), discoverable (described with metadata) and retrievable (modelled and stored in a recognizable format). According to the *World Bank Group* definition "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)." According to the PSI Directive, Open Data can be charged at marginal costs i.e. does not mean free of charge. Acceptable file formats for publishing data are CSV, XML, JSON, plain text, HTML and others. Recommended by the W3C consortium, international Web standards community, is the RDF format that provides a convenient way to directly interconnect existing open data based on the use of URIs as identifiers.

B. Linked Data Approach for Opening and Sharing Data

The term Linked Data (http://linkeddata.org/) refers to a set of best practices for publishing and connecting structured data on the Web. These best practices have been adopted by an increasing number of data providers over the past five years, leading to the creation of a global data space that contains many billions of assertions - the Linked Open Data cloud[2]. The government data respresents a big portion in this cloud.

Semantic approach to publishing and sharing information (based on the Linked Data paradigm) is already accepted on European level. In order to minimize incompatibilities between pools of PSI, the European Commission in collaboration with the W3C consortium has accepted a set of standard vocabularies that should be used for publishing data and metadata (used for describing the datasets, semantic services or repositories).

So far, the EU recommendations have been exemined by all EU countries, however there are differences in the effort of establishing semantic repositories on country level (see e.g. Germany[3], Denmark[4], and Estonia[5]), as well as in the amount of resources that are published and shared with other member states via the JoinUp platform (see federated repositories[6]). In 2012 the Commission announced the European Union Open Data portal [8], and currently under development is the pan-European Portal that will connect governments of the EU Member States. In the period 2012-2015, the Mihajlo Pupin Institute tested the Linked Data approach, developed several tools that support the data use and re-use (these tools are presented in Section IV), and established the Serbian CKAN [9].

C. Semantic platforms, tools and standards

Recognizing the big financial contribution of the EU Commission for development of tools that cover different aspects of the Linked Data life cycle [10], the following platforms are already in place for building Linked Data applications: the Linked Data Stack (http://stack.linkeddata.org/), fluidOps Information Workbench[7], Graphity Platform[8], Ontos Linked Data Information Workbench[9] and others. Common to these platforms is that data is represented using "W3C Linked Data" standards[10] and that developed applications can run both as cloud services or enterprise applications.

D. Quality Issues

Analyzing quality issues in the ODS project framework, the authors concluded that "Data quality has multiple dimensions including accuracy, availability, completeness, conformance, consistency, credibility, processability, relevance and timeliness" [11].

III. Example from Serbia

A. Challenges related to PSI Directive Implementation

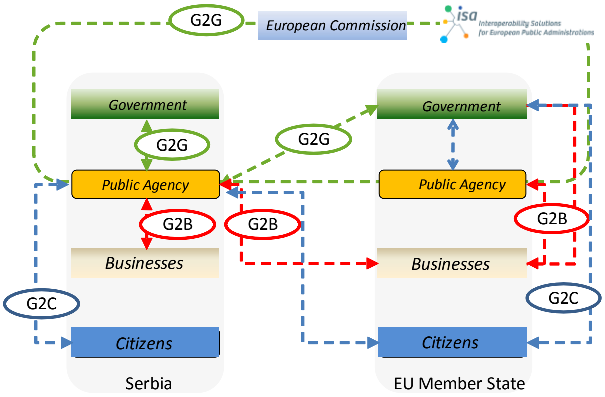

Fig. 1. PSI Re-use and Interoperability in Government Services.

The Draft of the Serbian e-Government Strategy[11] for the period 2015-2018 foresees implementation of the PSI Directive, where the eUPRAVA portal[12] is a central point of access to e-services for all Serbian citizens (G2C), businesses (G2B) and employees (G2G) in public administration. Currently, more than 500 services are available on the portal, and about 130 bodies announced their services there. According to our analysis, the greatest challenge in implementing the PSI Directive in Serbia will be to open and express the description (metadata) about the available data/services on the eUPRAVA portal or elsewhere, in a compliant machine-processable format as is required for federation with other EU portals. Technical

[2] http://lod-cloud.net
[3] XRepository, https://www.xrepository.deutschland-online.de/
[4] Digitalisér.dk, http://digitaliser.dk/
[5] RIHA, https://riha.eesti.ee/riha/main
[6] https://joinup.ec.europa.eu/catalogue/repository

[7] http://www.fluidops.com/information-workbench/
[8] http://graphityhq.com/technology/graphity-platform
[9] http://www.ontos.com/products/ontosldiw/)
[10] http://www.w3.org/2013/data/
[11] Directorate for Electronic Government, http://digitalnaagenda.gov.rs/en/
[12] http://www.euprava.gov.rs/

requirements for federation include[13]:

- Access to the harvested sites (login account or CKAN-API, FTP)
- DCAT Application Profile (DCAT-AP) for the government open data portal that will enable exchange of descriptions of datasets in eUPRAVA with other portals, as well as aggregation of and search for datasets across data portals in Europe
- INSPIRE metadata for geospatial data, etc.

B. *Semantic Interoperability with the Linked Data Stack*

In the period 2012-2015, funded by the EU Commission in the LOD2 and GeoKnow framework, the Mihajlo Pupin Institute was involved in development of state-of-the-art components for the *Linked Data Stack*[14] aimed at publishing Open Data in RDF format based on Linked Data principles, where semantic interoperability is achieved through standard vocabularies for representing data /metadata and processing information, re-usable open-source components and services (see Figure 2).

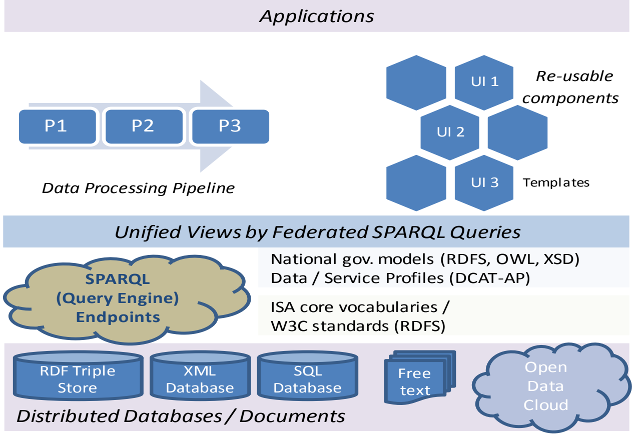
Applications
P1
P2
P3
Data Processing Pipeline
UI 1
UI 2
UI 3
Re-usable components
Templates
Unified Views by Federated SPARQL Queries
SPARQL (Query Engine) Endpoints
National gov. models (RDFS, OWL, XSD)
Data / Service Profiles (DCAT-AP)
ISA core vocabularies / W3C standards (RDFS)
RDF Triple Store
XML Database
SQL Database
Free text
Open Data Cloud
Distributed Databases / Documents

Fig. 2. Semantic integration and interlinking approach.

C. *Customizing the Linked Data Stack for Publishing and Consuming Statistical Data in Serbia*

Business requirements for Serbia were discussed and elaborated with two Serbian institutions: the Statistical Office of the Republic of Serbia (SORS) and the Serbian Business Registers Agency. As a result of this collaboration, the following activities were finalized:

- Localization of the CKAN catalog system in Serbian language and establishment of the *Serbian CKAN*[15],
- Integration of components needed for statistical data *processing* into a unique solution named *Statistical Workbench* [12] and its deployment at the Statistical Office[16],
- Development of a component for analysis of statistical data named *ESTA-LD*[17] (*Exploratory Spatio-Temporal Analysis of Linked Data* [13]) that can be easily *combined* with other components from the Stack and used for different purposes.

From our overall contribution to the Linked Data Stack in the last four years, we would like to point to these two reusable open-source components:

- the *RDF Data Cube Validation tool* [14], and
- *Exploratory Spatio-Temporal Analysis of Linked Data tool (ESTA-LD)* [13].

D. *RDF Data Cube Validation tool*

This tool enables quality assessment of the statistical data published in Linked Data format. It is implemented as a web applications that can work on top of any RDF store. Once the store is specified, the tool runs a set of queries to find and repair inconsistencies, such as missing links, issues with structure of the cube, etc. Moreover, along with the list of problems, the user is also provided with a description of the problem at hand, and when possible, a solution to the problem. The tool is built using Vaadin and Sesame frameworks, which were used to implement the user interface and evaluate queries respectively. Demonstrator is available online[18].

Example: In order to achieve better transparency and interoperability, in addition to the data in CSV and XML, SORS decided to publish some datasets in the RDF format as well. This effort included creation of SORS code lists [15], transformation of SORS data to RDF, preparation of the metadata descriptions, and finally validation of the dataset with the *RDF Data Validation tool* in order to ensure high quality of published data cubes and consistent use of the recommended RDF Data Cube vocabulary on data and metadata level before publishing data on the Serbia CKAN portal.

E. *Exploratory Spatio-Temporal Analysis of Linked Data Data*

Similar to the *RDF Data Validation tool*, ESTA-LD is a re-usable component that can be used with any RDF store. Once the data from different sources is integrated in a single RDF Triple Store, this tool allows:

- visualization of statistical data on a chart with an emphasis on temporal and spatial dimensions, creating different intersections in a multidimensional space of statistical indicators,
- showing geographical data on multiple levels of granularity, intuitive comparison of indicator across geographical regions, as well as combining the temporal and spatial dimensions, thus allowing to observe differences of the selected indicator between geographical regions and its evolution through time.

ESTA-LD's user interface is implemented partly in Vaadin, a Java server-side framework, while JavaScript libraries Leaflet and Highcharts are used to visualize statistical indicators on a map and chart respectively. Data is acquired from the store using Sesame. Information about how to install and run the tool as well as a link to the demo

[13] http://ec.europa.eu/digital-agenda/en/news/development-pan-european-open-data-portal-and-related-services-presentation-wendy-carrara

[14] Serbian CKAN - Comprehensive Knowledge Archive Network http://stack.linkeddata.org/

[15] http://rs.ckan.net/

[16] http://lod2.stat.gov.rs/lod2statworkbench

[17] http://geoknow.imp.bg.ac.rs/ESTA-LD

[18] http://jpo2.imp.bg.ac.rs/rdf-data-cube-validation-demo/

are available on GitHub[19]. Future work is planned around enrichment of the statistical datasets with external datasets such as DBpedia and linking the data with similar data from EUROSTAT, the statistical office of the European Union.

Example: In order to illustrate the potential and value of published information, and spark development of new services on top of it, as well as to offer a valuable service to the public, we decided to build ESTA-LD, a tool that provides highly interactive analysis, searching and filtering capabilities over the statistical RDF datasets coming from different publishers (SORS and SBRA in our case). By analyzing different socio-economic indicators at the same time, business users can spot trends that cannot be identified with the analysis of only one indicator.

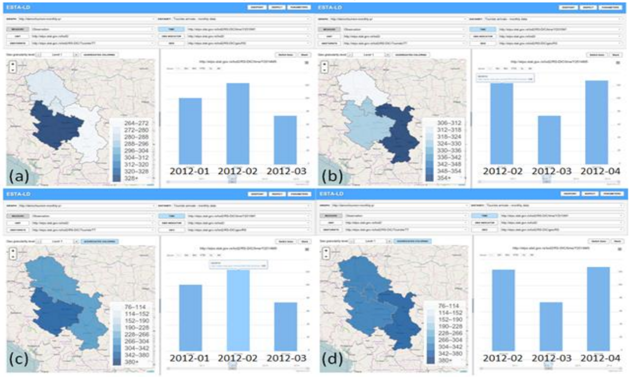

Fig. 3. Advanced user interface in ESTA-LD.

IV. Lessons Learned and Recommendations

In the LOD2 and GeoKnow framework, the PUPIN team studied the EU strategies for scalable and interoperable Open Government Data ecosystem and developed open-source tools by using the latest advances in Linked Open Data. Following are the main lessons learnt from this study:

- Linked Data approach (e.g. Linked Data Stack) offers a novel, flexible way of integrating and interlinking data across Europe. The use of open-source tools and non-proprietary, machine-readable formats fosters new business models and re-use of data outside the public administration.
- The recommended standards (core vocabularies, INSPIRE metadata) and tools (CKAN) are vastly used with open data portals across Europe and the PUPIN initiative around the Serbian CKAN is a good start towards building a semantic data catalogue of open data from Serbia.

V. Conclusion

This paper discusses technical aspects related to the implementation of the PSI Directive and points to Open Data activities in Serbia. According to the Draft of the Serbian e-Government Strategy, Serbia foresees to implement the PSI Directive in the next period 2016-2020. The Directive envisions publishing of the public/private datasets in machine readable format, thus, making sharing, using and linking of information easy and efficient.

Tools presented in the paper allow governments and their agencies to publish their data based on open standards and machine-readable formats. Herein, the emphasis is on statistical data, however the approach is generic and, using domain-specific vocabularies, applicable to other business areas. In future, a significant effort will be put into further adaptation of the EC recommendations for building interoperable tools and services, while taking into consideration different aspects, such as scalability, flexibility and ease-of-use/friendliness.

References

[1] "European legislation on reuse of public sector information", European Comission. Available: http://ec.europa.eu/digital-agenda/en/european-legislation-reuse-public-sector-information.

[2] Guidelines on recommended standard licences, datasets and charging for the reuse of documents (2014/C 240/01). Official Journal of the European Union C240/1-10 24.7.2014.

[3] https://ec.europa.eu/digital-agenda/en/digital-agenda-europe-2020-strategy

[4] "Open data An engine for innovation, growth and transparent governance", European Commission, COM(2011) 882 final. Access: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:PDF

[5] ISA2, http://ec.europa.eu/isa/isa2/index_en.htm

[6] Mijović, V., Janev, V., Paunović, D. (2015) ESTA-LD: enabling spatio-temporal analysis of linked statistical data, In Zdravković, M., Trajanović, M., Konjović, Z. (Eds) Proceedings of the 5th ICIST Conference (Kopaonik, Serbia), Society for Information Systems and Computer Networks.

[7] Elements of the Revised PSI Directive, SHARE-PSI 2.0. Available: https://www.w3.org/2013/share-psi/elements

[8] http://open-data.europa.eu/en/linked-data

[9] Janev, V., Milošević, U., Mijović, V., Milojković, J., Ivković, M., Vraneš, S. (2013) Towards Semantic Interoperability of Statistical Data, In: Proceedings of the INFOTECH 2013 ICT Conference & Exhibition (Aranđelovac, Serbia), JURIT- the Association for Computing, Information Technology, Telecommunications, Automation and Management of Serbia ISBN 978-86-82831-19-8

[10] 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

[11] Makx Dekkers, Nikolaos Loutas, Michiel De Keyzer Sijn Goedertier, Open Data and Metadata Quality, http://www.slideshare.net/OpenDataSupport/open-data-quality-29248578

[12] Van Nuffelen, B., Janev, V., Martin, M., Mijovic, V., Tramp, S. (2014) Supporting the Linked Data Life Cycle Using an Integrated Tool Stack. In S. Auer, V. Bryl, S. Tramp (Eds) Linked Open Data -- Creating Knowledge Out of Interlinked Data. Lecture Notes in Computer Science vol. 8661, pp. 108–129. Springer International Publishing. ISBN: 978-3-319-09846-3 (Online)

[13] Mijović, V., Janev, V., Paunović, D. (2015) GeoKnow Deliverable 4.6.2 Advanced GUI for GeoKnow prototype for Exploratory Spatiotemporal Analysis, http://svn.aksw.org/projects/GeoKnow/Public/D4.6.2_ESTA-LD.pdf

[14] Janev, V., Mijović, V., Vraneš, S. (2013) LOD2 Tool for Validating RDF Data Cube Models, In V. Trajkovik, A.Mishev (Eds) Web Proceedings of 5th ICT Innovations Conference 2013, (Ohrid, Macedonia). Published on-line by Macedonian Society on Information and Communication Technologies, pp. 1-9, ISSN 1857-728

[15] Vraneš, S., Janev, V., Spasić, M., Milošević, U. (2012) LOD2 Deliverable 9.5.1 Establishment of the Serbian CKAN Open Government Metadata Repository, http://static.lod2.eu/Deliverables/LOD2_D9.5.1_Serbian_CKAN.pdf

[19] https://github.com/GeoKnow/ESTA-LD