Statistical Metadata Management in European e-Government Systems

Valentina Janev, Vuk Mijović, and Sanja Vraneš
"Mihajlo Pupin" Institute, University of Belgrade, Volgina 15, 11060 Belgrade, Serbia
Valentina.Janev, Vuk.Mijovic, Sanja.Vranes@pupin.rs

***Abstract*—The goal of this paper is, from one side, informative i.e. to introduce the existing activities in the European Commission related to metadata management in e-government systems, and from the other, to describe our experience with metadata management for processing statistical data in RDF format gained through development of Linked Data tools is recent EU projects LOD2, GeoKnow and SHARE-PSI. The statistical domain has been selected in this analysis due to its relevance according to the EC Guidelines (2014/C 240/01) for implementing the revised European Directive on the Public Sector Information (2013/37/EU). The lessons learned shortly described in this paper have been included in the SHARE-PSI collection of Best practices for sharing public sector information.**

Keywords: *Linked Data, metadata, RDF Data Cube Vocabulary, PSI Directive, Best Practices*

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 17th of July 2013, provides a common legal framework for a European market for government-held data (public sector information) [1,2].

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*."

Analyzing the metadata management requirements and existing solutions in EU Institutions and Member States, the authors [3] have found that 'activities around metadata governance and management appear to be in an early phase'.

A. Related Work

The most common definition of metadata is "data about data." Metadata management can be defined as "as a set of high-level processes for structuring the different phases of the lifecycle of structural metadata including design and development of syntax and semantics, updating the structural metadata, harmonisation with other metadata sets and documentation" [1] Commercial software providers e.g. IBM, make difference between business, technical, and operational metadata. Hence, metadata management includes tools, processes, and environment that enable organization to answer different questions related to resources they own. Samsung Electronics [4], for instance, is looking at three types of issues related to metadata management: (1) metadata definition and management, (2) metadata design tools, and (3) metadata standards.

B. Paper structure

The goal of this paper is, from one side, informative i.e. to introduce the existing activities in the European Commission related to metadata management based on our knowledge obtained through participation in recent EU projects LOD2, GeoKnow and SHARE-PSI. From the other, it describes our experience with metadata management for processing statistical data in RDF format gained through development of Linked Data tools.

Section 2 introduces the European holistic approach to interoperability in eGovernment Services. Next, Section 3 shows the latest trends to semantic interoperability in public administration across Europe based on the Linked Data approach. Using an example from Serbia, Sections 4 further analyses the challenges of metadata management of statistical data and points to tools developed in the Mihajlo Pupin Institute. The Pupin team contributed also to sharing the existing experiences with European partners as is described in Section 5. Section 6 concludes the paper.

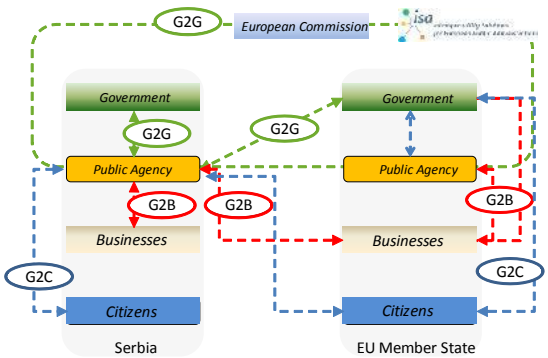

Figure 1. E-Government services across Europe

II. PROBLEM STATEMENT

A. Holistic Approach to PSI Re-use and Interoperability in European eGovernment Services

Since 1995, 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]. The holistic approach (G2G, G2C, G2B, see Figure 1) foresees four levels of interoperability, namely legal, organizational, semantic and technical interoperability. In our research, we take special interest in methods and tools for semantic data interoperability that support 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*". Up to now, the ISA programme has provided a wide range of supporting tools, see Repositories of reusable software, standards and specifications.

B. Other Recommendations: What Truly Open Data Should Look Like?

In order to share datasets between users and platforms, the datasets need to be accessible (regulated by license), 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)*" ("Open Data Essentials"). According to the PSI Directive, open data can be charged at marginal costs i.e. it does not have to be 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.

C. What do we need for efficient data sharing and re-use?

The data can be exposed for download and/or exploration in different ways. Although there are "Best Practices for Publishing Linked Data" (2014), the metadata of published datasets can be of low quality leading to the questions such as:

- 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?
- Is it possible to fuse heterogeneous data and formats used by different publishers and what are the necessary steps? Are there standard approaches / services for querying government portals?
- What is the quality of the data / metadata, i.e., do we have a complete description of the public datasets? Does the publisher track changes on data and schema level? Is the publisher reliable and trustful?

In order to that make the use of open data more efficient and less time-consuming, standardized approaches and tools are needed e.g. the Linked Data tools that work on top of commonly accepted models for describing the underlying semantics.

III. LINKED DATA APPROACH FOR OPEN DATA

A. Linked Data Principles

The Linked Data principles have been defined back in 2006 [5], while nowdays the term Linked Data[2] is used to refer to a set of best practices for publishing and connecting structured data on the Web. These principles are underpinned by the graph-based data model for publishing structured data on the Web – the Resource Description Framework (RDF) [6], and consist of the following: (1) using URIs as names for things, (2) making the URIs resolvable (HTTP URIs) so that others can look up those names, (3) when someone looks up a URI, providing useful information using the standards (RDF, SPARQL), and (4) including links to other URIs, so that they can discover other things 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 thousands of datasets and billions of assertions - the Linked Open Data cloud[3]. The government data respresents a big portion of this cloud.

Some governments around the world[4] [7,8] have adopted the approach and publish their data as Linked Data using the standards and recommendations issued by the Government Linked Data[5] (GLD, Working Group,one of the main providers of Semantic Web standards.

B. Metadata on the Web

Metadata, or structured data about data, improves discovery of, and access to such information[6]. The effective use of metadata among applications, however, requires common conventions about semantics, syntax, and structure. The Resource Description Framework (RDF) is indeed the infrastructure that enables the encoding, exchange, and reuse of structured metadata.

In the last twenty years standardization organizations such as the World Wide Web Consortium (W3C) are working on defining conventions e.g. for describing government data[7] (see RDF Data Cube Vocabulary[8], Data Catalog Vocabulary[9] and the Organization Vocabulary[10]).

C. Re-using ISA Vocabularies for Providing Metadata

The ISA programme supports the development of tools, services and frameworks in the area of e-Government through more than 40 actions[11]. In the area of metadata management, the programme recommends using Core Vocabularies (Core Person, Registered organisation, Core

[2] http://linkeddata.org/
[3] http://lod-cloud.net
[4] http://linkedopendata.jp/
[5] http://www.w3.org/2011/gld/)
6
[7] https://www.w3.org/blog/news/archives/3591
[8] https://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/
[9] https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/
[10] https://www.w3.org/TR/2014/REC-vocab-org-20140116/
[11] http://ec.europa.eu/isa/ready-to-use-solutions/index_en.htm

Location, Core Public service)[12] as '*simplified, re-usable and extensible data models that capture the fundamental characteristics of an entity in a context-neutral fashion'*. These vocabularies should support the description of the base registries that are maintained by EU public administrations (i.e. *a base registry is a trusted, authentic source of information under the control of an appointed public administration.*) Moreover, they should support harmonization of base registries across Europe, as well as additional registries, e.g. see *The notion of Linked Land Administration* [9][13].

IV. EXAMPLE: METADATA MANAGEMENT IN STATISTICAL DATA PROCESSING

A. *SDMX and RDF Data Cube standards*

In January 2014, W3C recommended the *RDF Data Cube* vocabulary[14], as a standard vocabulary for modeling statistical data. The vocabulary focuses purely on the publication of multi-dimensional data on the Web. The model builds upon the core of the *SDMX 2.0 Information Model*[15] realized in 2001 by the Statistical Data and Metadata Exchange (SDMX[16]) Initiative with the aim to achieve greater efficiencies in statistical practice.

SDMX Information model differentiates between

- "structural metadata" - those metadata acting as identifiers and descriptors of the data, such as names of variables or dimensions of statistical cubes.
- "Reference metadata" - metadata that describe the contents and the quality of the statistical data (concepts used, metadata, describing methods used for the generation of the data, and metadata, describing the different quality dimensions of the resulting statistics, e.g. timeliness, accuracy).

B. *Structural Metadata (DSDs)*

Each data set has a set of *structural metadata*. These descriptions are referred to in SDMX as Data Structure Definitions (DSD), which 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 identifying and descriptive (structural) metadata. DSD also specifies which code lists (conceptual schemas, see Figure bellow) 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.

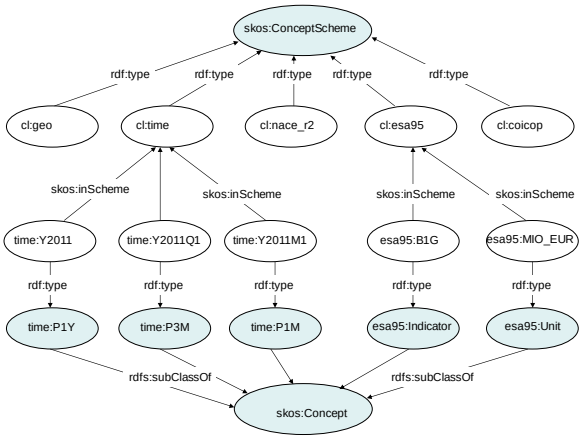

Figure 3. Example of concepts and instances in code lists

Once defined on a national level, and then registered in the EU JOINUP platform, code lists can be used for publishing statistical data coming from different public sector agencies. For more information on how to define a DSD for a statistical dataset, we refer to results from the LOD2 project see D9.5.1.

C. *Quality Assessment of Structural Metadata of RDF Data Cubes*

According to W3C recommendations, a statistical dataset in RDF should be modeled with the *RDF Data Cube* vocabulary and should adhere to the integrity constraints defined in the standard. To that aim, we developed a specialized tool (RDF Data Cube Validation Tool, 2014) to be used prior to publishing the statistical data in RDF format. The RDF *Data Cube Validation component* checks if the statistical dataset (RDF graph) is valid according to the integrity constraints defined in the RDF Data Cube specification (http://www.w3.org/TR/vocab-data-cube). Each constraint in the W3C document is expressed as narrative prose, and where possible, a SPARQL ASK query or query template that returns true if the graph contains one or more Data Cube instances which violate the corresponding constraint. Our tool runs slightly modified versions of these queries which allow it to not only show if the constraint is violated or not, but also to list the offending resources, provide information about the underlying issue and if possible offer a quick solution in order to repair the structural metadata.

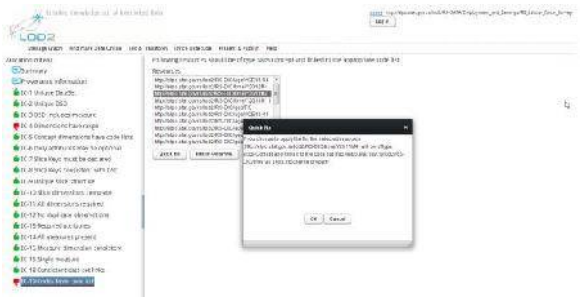

Figure 4. RDF Data Validation Tool - GUI

D. *Improving Quality by Storing and Re-using Metadata*

In order to support reuse of code lists and DSD

12 https://joinup.ec.europa.eu/site/core_vocabularies/Core_Vocabularies_user_handbook/ISA%20Hanbook%20for%20using%20Core%20Vocabularies.pdf

13 http://www.yildiz.edu.tr/~volkan/INSPIRE_2014.pdf

14 http://www.w3.org/TR/vocab-data-cube/.

15 *SDMX Content-oriented Guidelines: Cross-domain code lists.*(2009). Retrieved from http://sdmx.org/wp-content/uploads/2009/01/02_sdmx_cog_annex_2_cl_2009.pdf

16 http://www.sdmx.org/

descriptions defined by one publisher (e.g. Statistical Office of the Republic of Serbia), we have designed a new service that stores the DSDs that have been used in published open data and offers a possibility to

- build and maintain a repository of unique DSDs,
- create a new DSD based on the underlying statistical dataset,
- refer to, and reuse a suitable DSD from the repository.

Thus the tool has potential to uniform the creation and re-use of structural metadata (DSDs) across public agencies, reduce duplicates and mismatches and improve harmonization on national level (we assume here that public agencies in one country will use the same code lists).

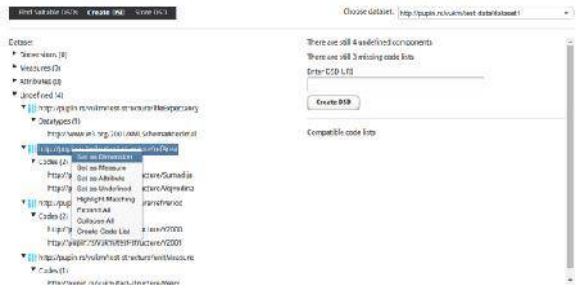

Figure 5. DSD Repository - GUI

V. BEST PRACTICES FOR OPEN DATA

A. About the SHARE-PSI project

Financed by the EU Competitiveness and Innovation Framework Programme 2007-2013, in the last two years, the SHARE-PSI network is involved in analysis of the implementation of the PSI Directive across Europe. The network is composed of experts coming from different types of organizations (government, academia, SME/Enterprise, citizen groups, standardization bodies) from many EU countries. Through a series of public workshops, the experts were involved in discussing EU policies and writing best practices for implementation of the PSI Directive. In the project framework, a collection of Best Practices[17] was elaborated that should serve the EU member states to support the activities around PSI Directive implementation.

Curious about the adoption of the Linked Data concepts in European e-government systems, we carried out an anlaysis which produced findings that are presented bellow.

B. Findings about EU OGD Initiatives (related to metadata management)

1) Semantic Asset Repositories

According to our research, there are differences in the effort of establishing semantic metadata repositories on country level (see e.g. Germany[18], Denmark[19], and Estonia[20]), as well as in the amount of resources that are published and shared with other member states via the JoinUp platform (see federated repositories[21]).

Currently in the EU, still ongoing is the adoption of the ISA Core Vocabularies and Core Public Service Vocabulary on national level (see e.g. implementation in Flanders[22] or Italy [10]) and the exchange of data inside a country (see Estonian metadata reference architecture)[23].

2) Federation of Data Catalogues

The DCAT-AP is a specification based on the Data Catalogue vocabulary (DCAT) for describing public sector datasets in Europe. Its basic use is to enable cross-data portal search for data sets and make public sector data better searchable across borders and sectors. This can be achieved by the exchange of descriptions of datasets among data portals. Nowadays, an increasing number of EU Member States and EEA countries are providing exports to the DCAT-AP or have adopted it as the national interoperability solution. The European Data Portal[24] implements the DCAT-AP and thus provides a single point of access to datasets described in national catalogs (376,383 datasets retrieved on January 5th 2016).

The hottest issue regarding federation of public data in in EU is the quality of metadata associated with the national catalogues [11].

C. SHARE-PSI Best Practices

Having strong technical background in semantic technologies, besides the activities in the SHARE-PSI network, our team was involved in other European projects that delivered open source tools for Linked Open Data publishing, processing and exploitation. In LOD2 project we tested the LOD2 tools with Serbian government data and the knowledge gain was communicated at the SHARE-PSI workshops [12].

Additionally, we contributed to publishing more than 100 datasets from the Statistical Office of the Republic of Serbia to the Serbian CKAN. Based on that experience we contributed to formulation of the following Best Practices:

- Enable quality assessment of open data[25] (PUPIN contributed with the experience with the RDF Data Cube Validation);
- Enable feedback channels for improving the quality of existing government data[26];
- Publishing Statistical Data In Linked Data Format[27] (PUPIN contributed with the experience with Statistical Workbench [12]).

While the first two Best Practices are well recognized and already in practice across EU countries, the third one still has a status *draft*, meaning that consensus across EU is needed.

17 https://www.w3.org/2013/share-psi/bp/
18 XRepository, https://www.xrepository.deutschland-online.de/
19 Digitalisér.dk, http://digitaliser.dk/
20 RIHA, https://riha.eesti.ee/riha/main

21 https://joinup.ec.europa.eu/catalogue/repository
22 https://www.openray.org/catalogue/asset_release/oslo-open-standards-local-administrations-flanders-version-10
23 http://www.w3.org/2013/share-psi/workshop/berlin/EEmetadataPilot
24 http://www.europeandataportal.eu/
25 https://www.w3.org/2013/share-psi/bp/eqa/
26 https://www.w3.org/2013/share-psi/bp/ef/
27 https://www.w3.org/2013/share-psi/wiki/Best_Practices/Publishing_Statistical_Data_In_Linked_Data_Format

VI. CONCLUSION

According to 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.

This paper introduced the latest open data and interoperability initiatives in the EU, including ISA recommendations, and described how Linked Data technologies can be used to publish open data on the Web in a machine readable format that makes it easily accessible, and discoverable. In this process, metadata plays an important role as it provides a way to describe the actual contents of the dataset which can then be published on well-known portals and catalogues, thus allowing data consumers to easily discover datasets that satisfy their specific criteria. Described principles were demonstrated on statistical data, however the approach (enhancing the data with metadata, quality assessment, reuse of metadata on national level) is generic and, using domain-specific vocabularies, applicable to other areas as well. In the 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.

ACKNOWLEDGMENT

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).

REFERENCES

[1] European legislation on reuse of public sector information. (2013, June 23). Official Journal of the European Union L 175/1. Retrieved from European Commission, http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32013L0037&from=EN

[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] Makx Dekkers, Stijn Goedertier, Audrius Leipus, Nikolaos Loutas, Metadata management requirements and existing solutions in EU Institutions and Member States, SC17DI06692, http://ec.europa.eu/isa/documents/metadata-management-requirements-and-existing-solutions-in-eu-institutions-and-member-states_en.pdf

[4] Won Kim, On Metadata Management Technology:Status and Issues, JOURNAL OF OBJECT TECHNOLOGY, Vol. 4, No.2, March-April

[5] Berners-Lee, Tim. (2006). Design Issues: Linked Data. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html

[6] F. Manola, E. Miller, B. McBride,"RDF Primer", 2004, February 10, Retrieved from http://www.w3.org/TR/rdf-primer/ 2005

[7] Hendler, J., Holm, J., Musialek, C. ; Thomas, G. (2012) US Government Linked Open Data: Semantic.data.gov, Intelligent Systems, IEEE (Volume:27 , Issue: 3), IEEE Computer Society.

[8] Wood, David (Ed.), Linking Government Data, Springer-Verlag New York, 2011

[9] Volkan Çağdaş, Erik Stubkjær, Supplementing INSPIRE through e-Government Core Vocabularies, http://inspire.ec.europa.eu/events/conferences/inspire_2014/pdfs/20.06_4_09.00_Volkan_%C3%87a%C4%9Fda%C5%9F.pdf

[10] Ciasullo, G., Lodi, G., Rotundo, A. (2015) Core Public Service Vocabulary: The Italian Application Profile. SHARE-PSI Workshop. Retrieved from http://www.w3.org/2013/share-psi/wiki/images/7/73/AgID_BerlinWorkshop.pdf

[11] Carera, W. (2015) The Role of the European Data Portal, http://www.w3.org/2013/share-psi/wiki/images/4/46/Share_PSI_2_0_EDP_paper_v1_1.pdf

[12] V.Janev, Publishing and Consuming Linked Open Data with the LOD Statistical Workbench, SHARE-PSI Workshop, Samos, Greece, 2014 https://www.w3.org/2013/share-psi/workshop/samos/agenda