Towards a methodology for publishing Linked Open Statistical Data

Irene Petrou
IMIS / RC “Athena”
irene.p@imis.athena-innovation.gr

Marios Meimaris
IMIS / RC “Athena”
m.meimaris@imis.athena-innovation.gr

George Papastefanatos
IMIS / RC “Athena”
gpapas@imis.athena-innovation.gr

ABSTRACT

The number of open government initiatives and directives around the globe with focused interest on publishing large amounts of data on the Web as “open” is increasing rapidly in the recent years. Opening up data aims for citizens, scientists and organizations to easily access, discover and exploit the data and consequently to benefit out of them. As a result, there has been an emerging need of integrating and representing those data in transparent and reusable ways, with high degree of interoperability which will further facilitate the discovery of new connections and insights by linking data coming from disperse sources. Statistical data published either by government bodies or by national statistical authorities are used for policy and decision making purposes, as they present important socioeconomic indicators. In this paper, we present a generic methodology describing the basic steps and overall model to publish statistical data coming from tabular data sources or relational databases as Linked Open Data.

Keywords

Statistical data, Data Cube Vocabulary, Tabular data, Linked Data

1. INTRODUCTION

The number of open government initiatives and directives around the globe with focused interest on publishing large amounts of data on the Web as “open” is increasing rapidly in the recent years. Opening up data aims for citizens, scientists and organizations to easily access, discover and exploit the data and consequently to benefit out of them. Although a great number of datasets are already made available to the public, their published format is often non-machine readable and difficult to process, thus not allowing reusability and interoperability [1]. As a result, there has been an emerging need for integration and representation in transparent and reusable ways that facilitate interoperability, collaboration and information enrichment [2]. Linked Open Data is an emerging set of directives and technologies, commonly adopted to overcome the encountered problems of publishing these large-scale distributed data.

Public sector information (PSI) is comprised by a comprehensive variety of digital information produced, collected and processed by public bodies and includes data ranging from geospatial information, such as digital maps, meteorological and public transportation, to legal, financial and statistical data [3]. In this paper, we focus our interest on the publication of open statistical data. Statistical data, often published by government bodies and/or national statistical authorities provide insights and socioeconomic indicators that are often used for policy and decision making purposes, thus enabling a better understanding of both the qualitative and quantitative characteristics of societies, as well as quantifying the results and social impact of such decisions. Similarly, decisions in the strategic management of a business may be influenced by those indicators, in addition to more domain-specific statistical data, such as market trends and product sales. Scientists use large amounts of statistical data, which represent collections of observations, to observe phenomena or for other research purposes, such as economic research [4]. Therefore, managing, sharing, manipulating and publishing statistical data efficiently are critical aspects of society’s evolution.

The majority of statistical data are offered in the form of tabular data, such as CSV files and Excel sheets [1]. In this paper, we highlight the arising need of publishing statistical datasets in linked open data formats rather than being available only in tabular forms and files. This process is twofold; it first involves the conversion of already published datasets from tabular to RDF format but most notably the restructuring of the whole data-publishing lifecycle of statistical data towards linked open data formats.

There have been various attempts and tools to facilitate the conversion of tabular data, although most of them may require user knowledge on semantic technologies, not enabling users to share data, such as Tabels[1], Google Refine[2], Triplify/Sparqlify, Any23[3] [1] or do not use the RDF Data Cube Vocabulary[4], a W3C Recommendation, in the mapping process which is currently the most appropriate way to represent multi-dimensional data and specifically statistical data.

The underlining model behind the Data Cube Vocabulary is the multidimensional, or else *cube* model, comprised of three basic components: *dimensions*, *measures* and *attributes*. Dimensions indicate what an observation applies to, measures indicate the phenomenon being observed, such as the number of households, whereas attributes help to interpret the observed values, such as the frequency, decimals or unit of measurement. These components allow the user to define the structure of a statistical dataset, which is called the *data structure definition* (DSD).

[1] http://idi.fundacionctic.org/tabels/

[2] http://refine.deri.ie/

[3] https://any23.apache.org/

[4] http://www.w3.org/TR/vocab-data-cube/

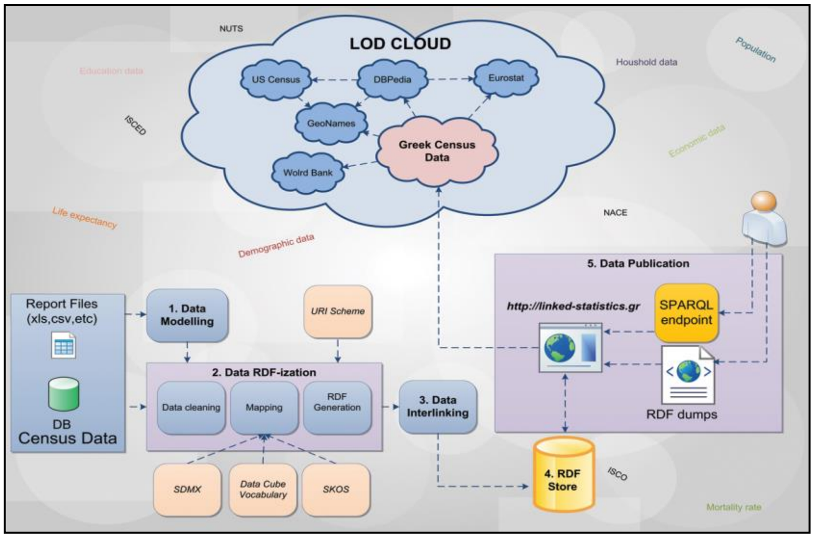

Fig. 1: Publishing Statistical Data as LOD

All together, the components and DSD, are used to define the actual measurements of the statistical datasets, or else called, the *observations*. It is important to note, that the Data Cube Vocabulary is built upon other vocabularies, SDMX[5](Statistical data and metadata exchange) and SKOS[6], to successfully describe statistical concepts; classifications, hierarchies or code lists[3,5].

The rest of the paper is structured as follows: In Section 2 we present a generic methodology for publishing statistical datasets as LOD. In Section 3 we present a list of common problems faced during the conversion raising the complexity of the process. Section 4 describes the current work of this research. Finally, in Section 5 concludes the paper, highlighting some future work.

2. METHODOLOGY

In this paper, we present a generic methodology describing the basic steps and overall model to publish statistical data coming from tabular data sources or relational databases as RDF Linked Open Data. The methodology builds on top of existing publishing tools, statistical vocabularies and LOD storage technologies to ease the process of publishing statistical data. The methodology is mainly comprised of five steps, as explained below [6].

2.1 Data modelling

The first step involves identifying and modelling custom ontologies for all domain-specific concepts and indices, which are not defined by other sources. A typical example concerns the structure of administrative divisions in Greece: in the 2001 Census Survey the divisions were defined according to the “KAPODISTRIAS” Plan containing six hierarchy levels of divisions, whereas in 2011, restructuring according to “KALLIKRATIS” Plan resulted in eight levels. This example, also reveals one of the problems discussed further in Section 3, involving the changing nature of resources over time which need to be addressed efficiently, maintaining the link between previous, current and future identical resources, as well as, tracking down the changed ones and how they changed. [6].

2.2 Data RDF-ization

The second step involves cleaning up the data, the identification of all the concepts within the dataset, the definition of each concept’s role as a dimension, measure or attribute and their mapping within the source file (eg to the appropriate column in the xls file), the identification of the actual data (observations), and all the dataset’s metadata. Concepts within the datasets may be mapped with existing concepts suggested and defined in the Content-Oriented Guidelines (COGs)[7] by SDMX, which contains a set of cross-domain concepts and code lists providing compatibility and interoperability across agencies [5]. In this step, the identification and definition of the code lists that are used to give a value to any of the components is also done. The components identified within the dataset are then mapped to the appropriate terms of the Data Cube Vocabulary to create the dataset’s structure, the dataset itself and the including observations, using the appropriate URI Scheme for each type of resource. The data are then exported as RDF in an RDF compliant serialization, such as RDF/XML. The mapping along with the RDF generation is done within the custom platform developed for data transformation in real-time.

2.3 Data interlinking

The different versions of codelists coming from the same resource are interlinked with each other using the appropriate linking property, eg skos:exactMatch for concepts. The transformed data are, also, linked with other resources. For example, indices are linked with datasets from the World Bank[8] and economic activities, occupational and educational data are linked with Eurostat’s datasets via the NACE, ISCO and ISCED classifications, respectively [6]. Various existing tools to discover links between data coming from different Linked data

[5] http://sdmx.org/

[6] http://www.w3.org/2009/08/skos-reference/skos.html

[7] http://sdmx.org/?page_id=11

[8] http://worldbank.270a.info/

sources, such as SILK Link Discovery Framework[9] and RelFinder[10] can then be used to enrich the data with meaningful links to external datasets or intra-dataset resources.

2.4 Data storage

The produced RDF data are uploaded, stored and maintained in a dedicated RDF store [6]. We use OpenLink Virtuoso[11], which is open source, and can be used for storing and dereferencing the data. Users may use other triple stores for hosting their data such as Fuseki[12], Neo4j[13] and H_+2RDF[14].

2.5 Data publication

According to the principles of Linked Data, the way to access Linked Data is via dereferencable URIs, which, when accessed, provide meaningful descriptions of the concept they represent in a variety of formats. Moreover, access to the data can be provided as RDF dumps or via SPARQL endpoints. For dereferencing the data OpenLink Virtuoso is used and all the data are accessible via the built-in SPARQL endpoint service or through the faceted browsing facility, where users can search resources and navigate from one resource to another.

3. PROBLEMS

During the conversion process, several problems arise that increase the complexity of converting tabular data to RDF. One of the problems we need to tackle through the conversion process relates to the evolution of the concepts (both in terms of structure and data values) over time. For example, in the case of census data, code lists used in the census in 2011 may vary from a previous or future census survey. The variations need to be tracked, versioned and recorded and unchanged members of the code lists have to be interlinked. Datasets, or parts of datasets, may also be corrected or revised. Inconsistency in the labeling of concepts is another common phenomenon, which makes the procedure of automatic matching more complicated. Inconsistency is, also, present in the structure of files of proprietary formats, such as MS Excel. The variations between excel formats make it harder to cover all the possible formats available. Another problem is that headers may be repeated within the datasets. Moreover, components, and specifically dimensions, may be nested in one another. This is challenging to overcome with an automatic way without any user supervision. Other problems of tabular data are mentioned in [1].

4. CURRENT WORK

Currently, our efforts focus on census data collected during Greece's 2011 Census Survey and provided by the Hellenic Statistical Authority (EL.STAT.) [2, 5]. We develop a platform through which the Greek Census Data are converted, interlinked and published. The methodology is applied to the census MS Excel files downloaded from the website of EL.STAT.[15], in order to optimize and accurately define all the sub-steps needed to ease the process of transformation. Figure 1 shows the overall model of the methodology. The datasets converted can be accessed at http://linked-statistics.gr in three ways: (a) download the data as RDF dumps for local processing, (b) query and browse the data using the SPARQL endpoint service and SPARQL query form and (c) link to the data by referencing to their unique identifier (URI).

5. CONCLUSIONS AND FUTURE WORK

In this paper we present a methodology on publishing statistical data as Linked Open Data. Statistical results usually become available for open access as .csv or .xls files in tabular form [3]. The methodology has been tested to a subset of the results of Greece's resident population census, conducted in 2011. This is an ongoing project, and more datasets will be published using the methodology to optimize and accurately define all the sub-steps needed to ease the process of transformation. A semi-automated tool will be developed to support the methodology, which will integrate the semi-automatic mapping of a statistical dataset to RDF using Data Cube Vocabulary without the requirement of LOD expertise. For example, a default URI minting scheme will be available for non-expert users. The user will be guided step-by-step through the process of conversion to easily identify all the appropriate components. More complex formats of Excel files will be supported. Future work, also, includes integrating generic visualization techniques of domain-specific statistical data.

6. ACKNOWLEDGMENT

This research is co-financed by the European Union (European Regional Development Fund - ERDF) and Greek national funds through the Operational Program "Competitiveness and Entrepreneurship" (OPCE II) of the National Strategic Reference Framework (NSRF) - Research Funding Program: KRIPIS.

7. REFERENCES

[1] Lebo T., Erickson J. S., Ding L., Graves A., Williams G. T., DiFranzo D., Li X., Michaelis J., Zheng J. G., Flores J., Shangguan Z., McGuinness D. L., and Hendler J. : "*Producing and using linked open government data in the TWC LOGD Portal*" in Linking Government Data (D. Wood, ed.), New York, NY: Springer, 2011.

[2] Ermilov I., Auer S., Stadler C.,: "*User-driven Semantic Mapping of Tabular Data",* in Proceedings of the 9th *International Conference on Semantic Systems (I-SEMANTICS'13),*Graz, Austria, 2013

[3] Petrou I., Papastefanatos G., Dalamagas T.: "*Publishing Census as Linked Open Data. A Case Study",* in Proceedings of the *2nd Int. Workshop on Open Data (WOD'13),* Paris, France, 2013

[4] Salas P. E.R. , M.M., Mota F.M.D. , Auer S., Breitman K., Casanova M.A.,L: "*Publishing Statistical Data on the Web",* in *Semantic Computing (ICSC), 2012 IEEE Sixth International Conference*2012: Palermo. p. 285 - 292

[5] Cyganiak, R., Reynolds, D., Tennison, J.: *The RDF Data Cube Vocabulary*, World Wide Web Consortium, 2013, http://www.w3.org/TR/vocab-data-cube/.

[6] Petrou I., Papastefanatos, G. : *"Publishing Greek Census Data as Linked Open Data",* in ERCIM News 96, Special theme: Linked Open Data, January, 2014.

[9] http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/
[10] http://www.visualdataweb.org/relfinder.php
[11] http://virtuoso.openlinksw.com/download/
[12] http://jena.apache.org/documentation/serving_data/index.html
[13] http://www.neo4j.org/
[14] https://code.google.com/p/h2rdf/
[15] http://www.statistics.gr/portal/page/portal/ESYE