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. A remarkable example is the Open Government Partnership[1] launched in 2011 with 8 founding countries, and growing to 64 participating countries by today. In the same year, datacatalogs.org[2] was launched too, which currently holds an inclusive list of 384 open data catalogs around the world. Opening up data aims for citizens, scientists and organizations to easily access, discover and exploit the data and consequently to benefit out of them, in line with governments providing transparency, accountability and empowerment to the citizens, while reducing corruption incidences.

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 [1]. 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. Eurostat highlights the importance and the role of statistics in policy making through the "Statistics for policymaking: Europe 2020" document in the context of its Europe 2020 strategy [2]. A most known example is the Greek crisis and the enactment of austerity measures triggered by the data published by the Hellenic Statistical Authority (EL.STAT)[3] in 2009 regarding the deficit, the debt and the GDP in Greece. Similarly, decisions in the strategic management of a business may be influenced by various socioeconomic indicators, 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 [3]. Therefore, managing, sharing, manipulating and publishing statistical data efficiently are critical aspects of society's evolution.

In Greece, although a great number of datasets are already made available to the public, their published format is often non-machine readable (e.g., pdf) and difficult to process and query (e.g., Excel files), thus not allowing reusability and interoperability. 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 [4]. [5] has characterized the current organization and publication of public sector data as "chaotic", pointing out the need for providing public sector data as Linked Open Data (LOD), for achieving greater data linkage and reusability. 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. Recently, several Greek LOD initiatives emerge; an example is the work of [6] where they generate, curate, interlink and distribute daily updated public spending data in Greece in LOD formats. Therefore, our current work is motivated by the following needs for publishing statistical data in LOD format:

- *Public sector accountability and transparency*: Open data enables citizens to have access to PSI, thus enabling transparency and accountability in the public sector. Moreover, linked data enables the retrieval, cross-reference and validation of data published by multiple public agencies' sites.
- *Interoperability and uniformity across public sector organizations:* LOD provides a means for publishing PSI in a

[1] http://www.opengovpartnership.org/

[2] http://datacatalogs.org/

[3] http://www.statistics.gr/

uniform and machine-readable format (RDF) with commonly agreed semantics, designed by global organizations and initiatives. This promotes better structure and understanding of available open data across different governmental agencies, as LOD help to have enhanced datasets, where cohesive information is available for the units that are common across these agencies.

- *Cost efficiency through Data reusability*: LOD reduces costs and manages time and resources more efficiently for recollecting, maintaining and producing PSI. For example, there is no need to conduct annual statistical surveys to collect data about imports-exports of organizations when such data are already maintained by the Customs agency.
- *Creation of added-value statistics, new insights for policy making procedures:* Data mash ups and correlation analysis across linked datasets allows investigating and answering more complex questions, which are required in the policy making process. There is no single agency that holds all the necessary information for different public issues.
- *Adoption of LOD practices and technologies in the Greek public sector*: The use of LOD technologies for making statistical data available can promote the wider adoption of LOD practices in the Greek public sector. This can help public servants and policy makers understand the opportunities and benefits gained by publishing the data in LOD format.

Currently, the majority of statistical data are offered in the form of tabular data, such as CSV files and Excel sheets [7]. 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[4], RDF Refine[5], Triplify, as supported by [7] or do not use the RDF Data Cube Vocabulary [8], a W3C Recommendation, which is currently the most appropriate way to represent multi-dimensional 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).

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[6](Statistical data and metadata exchange) and SKOS[7], to successfully describe statistical concepts; classifications, hierarchies or code lists[1,8].

In this work, we aim at providing a step-by-step guide to the user, in order to encourage publishing statistical data to LOD without prior knowledge or skills. Focusing on statistical, multi-dimensional data, the user will be provided a methodology to recognize the necessary components from the datasets which are required in the conversion process. The methodology may be applied and used with any combination of tools that support publishing data as LOD and are more appropriate to each user, such as the LOD2 Statistical Workbench [9]. Also, the methodology is closely related, though adjusted for statistical and multi-dimensional data, with other life cycle models suggested for publishing linked government data on the Web [10, 11, 12]. We further aim at highlighting parts of the publishing process that could be improved or automated to reduce the effort required by the user, as most of the available mechanisms, tools and practices are generic.

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 use case evaluation scenario to test our methodology. Section 4 describes a list of common problems faced during the publishing raising the complexity of the process. Finally, in Section 5 concludes the paper, highlighting some future work.

2. METHODOLOGY

Our goal is to 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. It follows the Best Practices for Publishing Linked Data [13], such as providing URI construction suggestions following URI policy rules, reuse of standard vocabularies, such as the Data Cube Vocabulary, SKOS, etc., converting data to RDF and providing access to the converted data. Figure 1 shows the overall model of the methodology. The methodology is mainly comprised of five steps, as explained below.

2.1 Data modelling

The first step involves identifying all the concepts within the dataset and modelling custom ontologies for all domain-specific concepts and indices, which are not defined by other sources. That involves extracting appropriate and useful datasets, either in CSV or .xls format from remote sites, analyzing the datasets and identifying the different concepts contained. Each concept is mapped with the corresponding concept of the multi-dimensional model, such as dimension, measure, code list, etc. The identification of common concepts that are already modeled from other sources can be referenced in this context. 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.

[4] http://idi.fundacionctic.org/tabels/
[5] http://refine.deri.ie/
[6] http://sdmx.org/
[7] http://www.w3.org/2009/08/skos-reference/skos.html

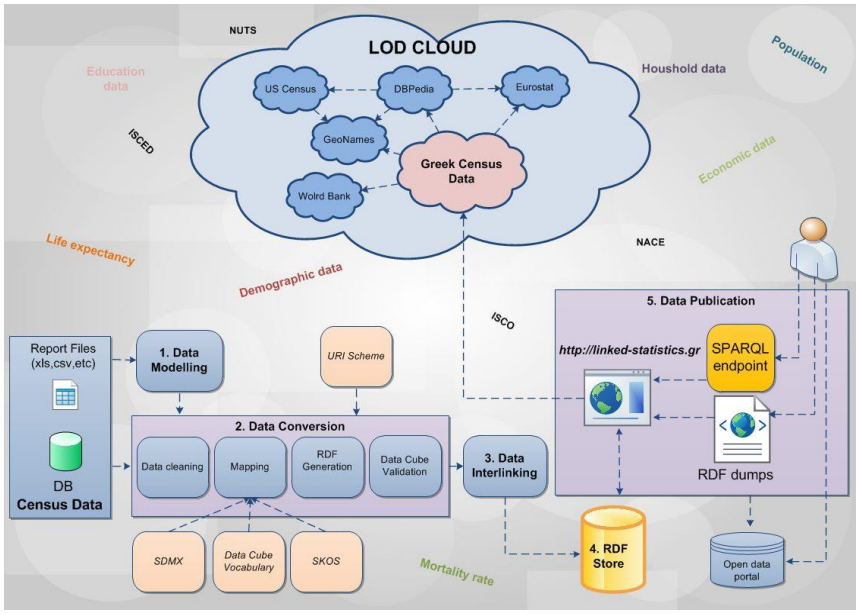

Fig. 1: Publishing Statistical Data as LOD

This example also reveals one of the problems discussed further in Section 4, involving the changing nature of resources over time which needs 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 [14].

2.2 Data Conversion

The second step involves cleaning up the data, mapping of dataset's concepts to the data cube elements, i.e., dimensions as qb:DimensionProperty, measures as qb:MeasureProperty or attributes as qb:AttributeProperty and furthermore to columns of the source file, the identification of the actual data (observations) as a set of qb:Observation instances, 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)[8] by SDMX, which contains a set of cross-domain concepts and code lists providing compatibility and interoperability across agencies [8]. The mappings are used to create the dataset's structure, the dataset itself and the including observations, using the appropriate URI Scheme for each type of resource. A default URI minting scheme, which has been designed using URI policy rules, is an input to this step to easily map the instances of the Data Cube Vocabulary and the resources. The code lists that are used to give a value to any of the components are also defined using SKOS vocabulary. The data are then exported as RDF in an RDF compliant serialization, such as RDF/XML and validated.

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 and economic activities, occupational and educational data are linked with Eurostat's datasets via the NACE, ISCO and ISCED classifications, respectively [14]. Various existing tools to discover links between data coming from different Linked data 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, such as OpenLink Virtuoso[11] or Fuseki[12].

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 and the data should be accessible to the public. Therefore, access to the data can be provided as RDF dumps or via SPARQL endpoints. Dereferencing of the data is achieved through the dedicated RDF store and all the data are accessible via a SPARQL endpoint service or through the faceted browsing facility, where users can search resources and navigate from one resource to another. Datasets may be further "announced" to the public, to be more discoverable, by publishing the data to international or national open data portals, such as *geodata.gov.gr* and *data.gov.gr*.

3. USE CASE EVALUATION

To evaluate, improve and complete our methodology we focused on publishing census data collected during Greece's 2011 Census Survey and provided by the Hellenic Statistical Authority (EL.STAT.) [1,14].

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

[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

Prefixes		
sdmx-dimension: <http://purl.org/linked-data/sdmx/2009/dimension#>		
sdmx-code : <http://purl.org/linked-data/sdmx/2009/code#>		
codelist: <http://linked-statistics.gr/ontology/code/2011/>		
age-groups: < http://linked-statistics.gr/ontology/code/2011/age/>		
measure: <http://linked-statistics.gr/ontology/measure/>		
skos: <http://www.w3.org/2004/02/skos/core#>		
qb: <http://purl.org/linked-data/cube#>		
Concept	**Vocabulary Mapping**	**URI**
Geographical code	qb:DimensionProperty	sdmx-dimension:refArea
Sex	qb:DimensionProperty	sdmx-dimension:sex
Sex code list Male Female Total	skos:ConceptScheme skos:concept skos:concept skos:concept	sdmx-code:sex sdmx-code:sex-M sdmx-code:sex-F sdmx-code:sex-T
Age	qb:DimensionProperty	sdmx-dimension:age
Age code list Age groups	skos:ConceptScheme skos:concept	codelist:age age:{*age-group}*
Population	qb:MeasureProperty	measure:population

Table 1: Concepts and Vocabulary mapping

Data modelling. Excel files were downloaded from the EL.STAT portal [15]. One sample dataset used is a dataset measuring "*Resident population by sex and age groups"* [16]. The concepts identified are administrative divisions of Greece (consisted of three other concepts; level of the administrative division, geographical code and a description) sex, age, population and age groups. Then, the type of each concept and its mapping to the appropriate cube construct was defined as follows: *geographical code*, *age* and *sex* as dimensions, and the *population* as a measure. The age groups are the values used to define the *age* dimension (0-4, 5-9, etc.) They were identified as a controlled code list. Looking at the concepts of administrative division it was decided that further modelling was required, as each administrative division is comprised by three concepts and hierarchy exists between each unit. Therefore, an administrative ontology was designed.

Data conversion. Unnecessary information was removed from the input file; two columns were removed (the level and description) as the geographical code was adequate to uniquely define the observations. The mapping between the concepts and the Data Cube Vocabulary and SDMX and the code lists with SKOS Vocabulary, and the URIs used are shown in Table 1. Given all the above modelling components (administrative ontology, dimensions, measures, code list) along with the original excel file and our URI minting scheme the dataset is converted to RDF and validated.

Data interlinking. The age groups were, then, linked with the corresponding age groups from e*urostat.linked-statistics.org.* The interlinking was automatically performed based on the age group description and it was manually curated for cases that no or multiple mappings occurred.

Data Storage. For storing the extracted triples and the triples from the interlinking, we chose to use OpenLink Virtuoso, which is open source, and is responsible for dereferencing the data and has a built-in SPARQL endpoint service, as well as faceted browsing.

Data publication. Finally, machine access is provided as it is one of the best practices of publishing Linked Data. 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).

4. LESSONS LEARNED

In the implementation of the methodology, several technical and modelling issues may arise, increasing the complexity of the publishing process. In following we present the most important aspects and discuss our insights based on our experience.

- *Evolution Modelling*: Evolution of the concepts (both in terms of structure and data values) over time may exist in the various datasets. 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. Datasets, or parts of the datasets, may face issues of evolution, as they may be corrected or revised. These issues must be considered when designing the URI scheme, in order to avoid conflicts, by adding *version* in the URI path. The variations need to be identified, versioned and recorded and unchanged members of the code lists have to be interlinked. Handling of evolving concepts and data should not be ignored as links to the published datasets need to have up-to-date information. This helps avoid changing URIs that end up with broken links.
- *Code List Identification and Reuse*. In many cases when converting a dataset from a file retrieved on the web, there is only a partial knowledge of all possible values and hierarchies comprising the code lists referenced in the dataset. This is due to the fact that the dimensions are identified and converted on the fly based on the source file. In addition, different values may refer to the same dimension value across different files, due to abbreviation reasons, errors in the production of the source files, etc. For example, one dataset may refer to individual ages ranging from 20-40 and a second dataset may refer to ages from 20-60. Before designing or versioning of a new code list, it is necessary to check if a code list already exists and can be enriched with new values rather than create a new one.
- *Conversion*: Automatic conversion of datasets may be problematic as there is lack of uniformity in the structure of the input source files. The structural heterogeneity of the tables within the input files makes it harder to cover all the possible formats available. Headers may be repeated within the datasets. Other similar problems of tabular data are mentioned in [7]. Moreover, components, and specifically dimensions, may be nested in one another, increasing the complexity of recognizing the individual values of each dimension. This is challenging to overcome with an automatic way without any user supervision.
- *Data Interlinking:* Regarding the interlinking of multidimensional data with other sources, there is a great need for creating links between observations from different datasets (especially those that come from different sites) and not only links between individual code lists or code list values. This enables merging and combining measurements coming from

different sources. However, linking of individual statistical measurements may be very difficult to achieve for datasets that are specific enough, i.e., they have a large number of dimensions whose values are broken down in low level details. For example, in the case of Census data, most national statistical datasets are bound to the reference area, which follow different administrative divisions across country and are broken down in different hierarchy details (e.g., neighbourhood). Individual observations can only be interlinked on the country level or on a commonly agreed area code list (e.g., NUTS II). This makes a large number of produced LOD statistics unexploited and reduces the available individual observations that can be interlinked across national statistical agencies. A possible solution would be the loose interlinking of observations based on similarity measures rather than exact matches, e.g., they may share the same DSD but refer to different dimension values or values from to different dimension hierarchy levels or different measure attributes, etc.

5. CONCLUSIONS AND FUTURE WORK

This work aims at highlighting the need and encourage government sectors/bodies, as well as other users, to publish their statistical data as LOD for public consumption in order to simplify the data linking process, promote data mash ups and reuse and perform correlation analysis combining relevant but different open government datasets or other external datasets. Nevertheless, using the Semantic Web technologies for data integration will reduce the costs of conducting new surveys and maximize the leveraging of existing data. 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 [1]. 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, that are not yet covered by the existing tools. 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] Petrou, Irene, George Papastefanatos, and Theodore Dalamagas. "Publishing census as linked open data: a case study." In *Proceedings of the 2nd International Workshop on Open Data*, p. 4. ACM, 2013.

[2] Eurostat, European Commision: "Statistics for policymaking: Europe 2020", Brussels, 2011, Available from:http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics_policymaking_europe_2020/documents/All.pdf [Accessed 10 June 2014]

[3] Salas, Percy E. Rivera, Michael Martin, Fernando Maia Da Mota, Sören Auer, Karin Breitman, and Marco A. Casanova. "Publishing statistical data on the web." In *Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on*, pp. 285-292. IEEE, 2012.

[4] Lebo, Timothy, John S. Erickson, Li Ding, Alvaro Graves, Gregory Todd Williams, Dominic DiFranzo, Xian Li et al. "Producing and using linked open government data in the twc logd portal." In *Linking Government Data*, pp. 51-72. Springer New York, 2011.

[5] Galiotou, Eleni, and Pavlina Fragkou. "Applying Linked Data Technologies to Greek Open Government Data: A Case Study." *Procedia-Social and Behavioral Sciences* 73 (2013): 479-486.

[6] Vafopoulos, Michalis, Marios Meimaris, Ioannis Anagnostopoulos, Agis Papantoniou, Ioannis Xidias, Giorgos Alexiou, Giorgos Vafeiadis, Michalis Klonaras, and Vasilis Loumos. "Public spending as LOD: the case of Greece."*Semantic Web Journal* (2013).

[7] Ermilov, Ivan, Sören Auer, and Claus Stadler. "User-driven semantic mapping of tabular data." In *Proceedings of the 9th International Conference on Semantic Systems*, pp. 105-112. ACM, 2013.

[8] Cyganiak, R., Reynolds, D., Tennison, J.: *The RDF Data Cube Vocabulary*, World Wide Web Consortium. Available from: http://www.w3.org/TR/vocab-data-cube/ (2014).

[9] LOD2 Statistical Workbench – LOD2 documentation – Confluence. [online] Available at: http://wiki.lod2.eu/display/LOD2DOC/LOD2+Statistical+Workbench [Accessed 04 June 2014]

[10] Auer, Sören, Lorenz Bühmann, Christian Dirschl, Orri Erling, Michael Hausenblas, Robert Isele, Jens Lehmann et al. "Managing the life-cycle of linked data with the LOD2 stack." In *The Semantic Web–ISWC 2012*, pp. 1-16. Springer Berlin Heidelberg, 2012.

[11] Hyland, Bernadette, and David Wood. "The Joy of Data – Cookbook for Publishing Linked Government Data." In *Linking Government Data*, pp. 3-26. Springer New York, 2011.

[12] Villazón-Terrazas, Boris, Luis M. Vilches-Blázquez, Oscar Corcho, and Asunción Gómez-Pérez. "Methodological guidelines for publishing government linked data." In *Linking Government Data*, pp. 27-49. Springer New York, 2011.

[13] World Wide Web Consortium. "Best practices for publishing linked data." (2014).

[14] Petrou, Irene, George Papastefanatos. "Publishing Greek Census Data as Linked Open Data." *ERCIM News* 2014, no. 96 (2014).

[15] Hellenic Statistical Authority. Available from: http://www.statistics.gr/portal/page/portal/ESYE

[16] Hellenic Statistical Authority. Available from: http://www.statistics.gr/portal/page/portal/ESYE/BUCKET/General/tab_04a_sex_age_de.xls