Publishing Census as Linked Open Data. A Case Study*

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

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

Theodore Dalamagas
IMIS / RC "Athena"
dalamag@imis.athena-innovation.gr

ABSTRACT
In this paper we present a case study on publishing statistical data as Linked Open Data. Statistical or fact-based data are maintained by statistical agencies and organizations, harvested via surveys or aggregated from other sources and mainly concern to observations of socioeconomic indicators. In this case study, we present the publishing as LOD of the preliminary results of Greece's resident population census, conducted in 2011. We have employed the Data Cube vocabulary and the Google Refine tool for modelling and publishing the census results.

Keywords
Statistical data, Data Cube Vocabulary, Google Refine

1. INTRODUCTION
Public sector information (PSI) characterizes a wide variety of digital information produced and collected by public bodies and includes digital maps, meteorological, legal, traffic or finally statistical data. Statistical data are maintained by statistical agencies and organizations, harvested via surveys or aggregated from other sources and mainly concern to observations of socioeconomic indicators, such as a country's population, the GDP, the CPI, etc. Statistical data are consumed by several stakeholders, such as governmental institutions, private organizations, journalists and scientists [1, 2]. Therefore, there is an increased interest on how statistical data can be published and reused on the Data Web. In the current paper, the principles[1] of Linked Data are followed, using the RDF data model. Our main objective for publishing the census data as LOD, rather than maintaining multiple .xls files, is these datasets to be available in an easier to process format (they can be crawled or queried via SPARQL), to be identifiable at the record level through their assignment with URIs and finally to be linkable, i.e., offer the ability to other sources to link and connect with them. In this way, census data will be available for consumption and exploitation by third parties enabling the data exploration and the development of novel applications. Moreover, publishing Greek census data as LOD will facilitate their comparison and linkage with datasets coming from other administrative resources, and deliver consistency and uniformity between the current and future datasets.

Various statistical vocabularies and interoperability standards have been proposed, such as the SDMX (Statistical data and metadata exchange) standard [3]. SDMX key elements are the Content-Oriented Guidelines (COGs), which in turn define a set of cross-domain concepts, code lists and categories in order to provide compatibility and interoperability across agencies. SCOVO [4] (Statistical Core Vocabulary) has been the first vocabulary for expressing statistical data in RDF. Although SCOVO is a relatively simple and easy to use vocabulary, it lacks of supporting important features when it comes on publishing statistical data, such as organizing the data within a dataset into slices, or distinguishing different concepts for data attributes, dimensions and measures, or finally describing only the contents of the dataset, rather than its structure. Therefore, the authors of SCOVO highly recommend the use of Data Cube Vocabulary [5] for publishing multidimensional statistical data.

In this paper we present a case study on representing and publishing statistical data concerning Greece's resident population census. Section 2 describes an overview of the census survey. Section 3 covers the LOD technology adopted for the specific case study, vocabularies and tools, and the detailed steps for publishing the census data. Finally, Section 4 concludes the paper, highlighting some future work.

2. CENSUS OVERVIEW
Census 2011 survey was conducted via questionnaires and concerned two types of information; data about Greece population (inhabitants), and data about households and dwellings. The data recorded for a person included: sex, birth date, marital status (Single, Married, Divorced, etc.), address of residence (in the form of geographical code which will be used in our case study), nationality, municipality, education level (Phd, MSc, School Leaving Certificate, etc.), job description (type, position, working hours, address), income (pensions, investments, work, etc.) and number of children. The data for households and dwellings included: address (levels in NUTS[2], Nomenclature of territorial units for statistics, classification and Greece's geographical code), type of residence (normal, mobile homes, etc.), characteristics of the dwelling (status, type of building, date of construction), total area (in m^2), facilities within the dwelling (number of bedrooms, heating, insulation), household comforts (electricity, gas, internet access, car parking, etc.), residents' information and relationships among the residents. Census questionnaires were processed via OCR for populating a relational database. Census results are then published as tabular data in .xls format, after all quality controls and anonymization methods have been applied.

[1] http://www.w3.org/DesignIssues/LinkedData.html

* This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thalis. Investing in knowledge society through the European Social Fund.

[2] http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction

3. CENSUS DATA AS LOD

3.1 LOD Technology adopted

There are various tools for mapping and publishing relational data or .xls files to RDF. In this case study, we have used Google Refine[3], which is a simple, though powerful tool, that offers the user the ability to clean up messy data, easily process the data using its own expression language (GREL), and transform data between different formats, such as TSV, CSV, *SV, Excel, JSON, XML, and Google Data documents. More formats can be supported using Google Refine extensions, such as the RDF Refine plugin[4] used for defining and mapping tabular data to RDF concepts. Another feature of this tool is the ability to reconcile and link the data against SPARQL endpoints and RDF dumps.

Regarding the vocabularies adopted for publishing the statistical information, we have used the Data Cube Vocabulary. Data Cube Vocabulary relies on the cube model [5]. A cube is organized according to a set of dimensions, attributes and measures, which all together are called components and are defined as RDF properties. A dimension component (defined as a *qb:DimensionProperty*) indicates what the observation applies to, such as the location or time of the observation. An attribute component (*qb:AttributeProperty*) is used for attributes of the observed value(s), such as the unit of measure. Finally, a measure component (*qb:MeasureProperty*) represents the phenomenon being observed, such as the number of inhabitants. Data Cube Vocabulary uses the SKOS [6] concepts for defining classifications, levels and hierarchies. Moreover, each dataset (represented as a *qb:DataSet*) conforms to a data structure definition (DSD) via a *qb:structure* property. The DSD concept (*qb:DataStructureDefinition*) refers to all dimension, attribute and measure components comprising the dataset, via *qb:dimension*, *qb:attribute* and *qb:measure* properties respectively. An observation is an instance of the class *qb:Observation* and is explicitly assigned with the observed values for the dimensions, attributes and measures via the defined *DimensionProperty*, *AttributeProperty*, and *MeasureProperty* respectively. Still, we may omit repeating the value for a component by attaching it at a higher level, for example at the dataset level. In this case, this value is applicable to all observations. Furthermore, Data Cube Vocabulary allows us to group subsets of observations together by keeping constant some of the dimension values. It uses the slice concept for declaring such subsets. Like datasets, slices have a structure (*qb:SliceKey*), in which the constant values for dimensions are defined. Each slice is an instance of the *qb:Slice* class, linked to the entire dataset through a *qb:slice* property, and to all comprising observations via *qb:observation* properties. SKOS Vocabulary was also used mainly for modelling the classes of our dataset in the form of concept schemes.

3.2 Case Study: Publishing Population Data

The first step in our case study was to choose the base domain (*http://linked-statistics.gr*) and the URI Scheme for hosting the produced linked datasets. We have followed the LATC Project[5] URI scheme that has been used for publishing Eurostat's data. Our URI scheme distinguishes between: (a) Schema components, such as the dataset structure, specifications for dimensions, measures, and attributes, which are located in the *schema* path and identified by a URI of the form *{BASE_URI}/schema#{ComponentName},* (b) Dataset and observations are located in the *data* path and identified by *{BASE_URI}/data/{DatasetName}* and *{BASE_URI}/data/{DatasetName}#{DatasetKey}* respectively, (c) Concepts and their values reused across multiple datasets are located in *dic* (dictionary) path and identified by *{BASE_URI}/dic/{ConceptName}* and *{BASE_URI}/dic/{Concept Name}#{value}*.

Next, we downloaded and imported the census results (available as .xls file[6]) in Google Refine. The results contain the permanent population in Greece for 2011 based on the place of residence for all levels of administrative divisions of the country (i.e., region, regional unit, municipality, municipal unit, municipal/local community, and village/neighbourhood). Each of these divisions are identified by a unique hierarchical geographical code, called geocode, in which the first two digits identify the region, the next two identify the regional unit, etc. Following, we defined our schema and built up the RDF skeleton, i.e. the mappings between the columns of the input file and the concepts used in our schema. We have considered the following schema components: (a) the column with the region divisions is our dimension. It was assigned with the *qb:DimensionProperty* and *qb:CodedProperty* properties and the *schema:geocodeDim* URI, (b) the population column is the measure. It was assigned with the *qb:MeasureProperty* and the *schema:population* URI, and (c) the unit of measurement for the population is an attribute, denoting that the observation is measured by number of habitants. It was assigned with *qb:AttributeProperty* and the *schema:UnitOfMeasure* URI. We linked this concept with *<http://eurostat.linked-statistics.org/dic/unit#HAB>* URI used by Eurostat for the number of habitants. Finally, the dataset structure was defined as *qb:DataStructureDefinition*, given the *schema:PopulationPerGeocodeCensus2011* URI and all aforementioned components were attached to it.

For modelling administrative divisions of residence, we have defined a new *skos:ConceptScheme* with the *dic:geocode* URI. The concept scheme *dic:geocode* was modelled as a subclass of the abstract *dic:codedList* class, that was introduced for modelling all attributes in census that contain data in the form of coded lists. Each division is uniquely identified by its geocode value, and characterized by its description and level. Thus, it was assigned with a unique *dic:geocode#{geocodeValue}* URI, e.g. *dic/geocode#01020204.* The description of the division was provided by *skos:prefLabel* property. Furthermore, each division was connected via a property *dic:haslevel* with the respective level that belongs to. For the division levels, we have defined a new *skos:ConceptScheme* with the *dic:geolevel* URI and we have identified each different level by the *dic:geolevel#{levelValue}* URI, e.g. *dic:geolevel#5.* Again, the concept scheme *dic:geolevel* is a subclass of the abstract *dic:levelList* class, that models lists of levels for hierarchical attributes, such as the geocode. Finally, we employed the *skos:broader* property for modelling the part-of relationships between divisions belonging to different levels in the hierarchy, as shown by the following example:

<dic:geocode#0102><skos:broader><dic:geocode#01>

[3] http://code.google.com/p/google-refine/wiki/Downloads?tm=2

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

[5] http://eurostat.linked-statistics.org

[6] Available only in Greek: www.statistics.gr/portal/page/portal/ESYE/BUCKET/General/resident_population_census2011.xls

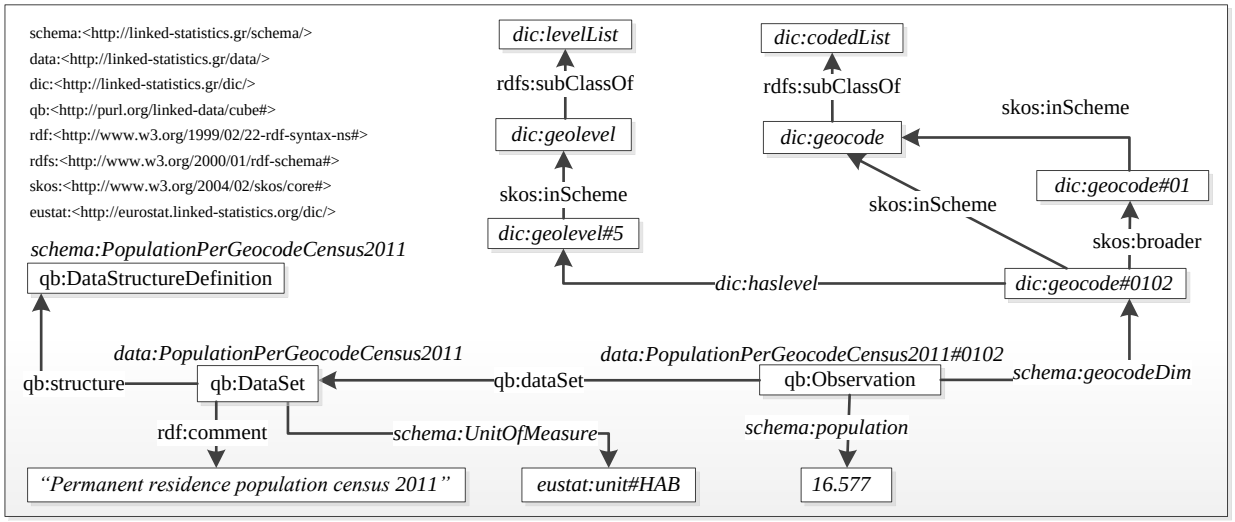

Figure 1: An RDF representation for the population of the division with the geocode 0102

An alternative design was also taken into consideration for modelling geocode concept. Instead of having one class, we could model all different levels of geocodes as separate classes (e.g., a different class for a region, a regional unit, a municipality, etc.) and connect them via part-of relationships. Although, this approach handles explicitly the different levels of divisions, it was not chosen, for reasons of future extensibility and maintenance of the model, especially for dealing with the case of changes in the structure of the hierarchies. For example, consider the case of a new division level added in the structure of the hierarchies; in our approach this requires inserting the triples describing the geocodes of this new level in the published dataset, without changing the overall model. Alternatively, we would have to insert a new class for the new level, and rewrite all part-of relationships to the reorganized hierarchy.

Finally, we have employed the *data:PopulationPer GeocodeCensus2011* URI for representing the dataset, assigned the *qb:DataSet* type and connected it via a *qb:structure* with the DSD. The *schema:UnifOfMeasure* was attached at the dataset level, instead of repeating it in all observations, via a *qb:componentAttachment* property. Each record in the file corresponds to an observation and is uniquely identified by the geocode value. Therefore, we have defined a *qb:Observation* for each record with the *data:PopulationPerGeocodeCensus2011 #{geocode}* unique URI. Each observation was connected via the *schema:geocodeDim* property with the appropriate division (geocode) and via the *schema:population* property with the value denoting the population for this geocode. Figure 1 shows an example for the population of the division with geocode 0102.

4. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a case study on publishing census data as Linked Open Data. Statistical results are usually become available for open access as .csv or .xls files in tabular form. In this case study, we have presented the publishing as LOD of the preliminary results of Greece's resident population census, conducted in 2011. We have used the Data Cube vocabulary and the Google Refine tool for modelling and publishing the census results. This is an ongoing project, in close collaboration with the Hellenic Statistical Authority, and the current case study is the first attempt for publishing statistical data coming from the census of 2011 as Linked Data. Future work aims at extending the proposed data model (a new release of the Data Cube Vocabulary supporting hierarchies will be further studied) for representing most statistical indexes and more complex census datasets. In addition, a SPARQL endpoint service will be configured to offer querying capabilities over the published data. Publishing the census data, and more statistical datasets thereafter, will offer the opportunity to other information sources to have direct access and link their datasets with the primary source of statistical information in Greece. It will enable users to perform complex queries on statistical data and develop innovative applications for exploration, integration and visualization of statistical datasets. All published datasets for this case study can be found in *http://linked-statistics.gr*.

5. REFERENCES

[1] Hausenblas M., H.W., Raimond Y., Feigenbaum L., and Ayers D., *Scovo: Using statistics on the web of data* ESWC, Springer, 2009. 5554.

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

[3] ISO. Statistical data and metadata exchange (SDMX), 2005.

[4] The Statistical Core Vocabulary (SCOVO), 2012, http://vocab.deri.ie/scovo

[5] Tennison J., T., *The RDF Data Cube Vocabulary*, 2012, http://www.w3.org/TR/vocab-data-cube/.

[6] A. Miles, S. Bechhofer: SKOS Simple Knowledge Organization System, August 2009, http://www.w3.org/TR/skos-reference/