Statistical Journal of the IAOS 33 (2017) 503–514
DOI 10.3233/SJI-150190
IOS Press

ICT tools for creating, expanding and exploiting statistical linked Open Data

Evangelos Kalampokis[a,b,*], Efthimios Tambouris[a,b] and Konstantinos Tarabanis[a,b]
[a]*University of Macedonia, Egnatia 156, 54006, Thessaloniki, Greece*
[b]*Information Technologies Institute, Centre for Research & Technology – Hellas, 6th km Xarilaou – Thermi, Thessaloniki 57001, Greece*

Abstract. A major part of Open Data concerns statistics such as financial and social indicators. Accurate and reliable statistics provide the solid ground for performing analyses that support businesses and governments in understanding the world and making better decisions. More importantly, the combination of statistical figures coming from disparate sources can unveil unexpected and unexplored insights. The adoption of the Linked Data principles and technologies has promised to facilitate data integration at a Web scale. In this paper, we describe the development of tools that support the whole lifecycle of linked statistical data including creation, expansion, and exploitation. Our approach is based on actively engaging organizations handling statistics as part of their everyday activities. The final technological outcome is the OpenCube Toolkit, a software platform that includes a set of relevant tools.

Keywords: Linked data, Open Data, statistics, data cube, data analytics

1. Introduction

Governments, organisations and companies are increasingly opening up their data for others to reuse. They launch data infrastructures (e.g. open data portals) to provide the data they produce or collect [9]. A major part of these open data concerns statistics such as financial and social indicators. For example, the vast majority of datasets published on the open data portal[1] of the European Commission is provided by Eurostat and thus is of statistical nature.

Statistical data are often structured in a multidimensional manner where a measured fact is described based on some dimensions, e.g. poverty rate could be described based on geographic area, time and age group. In this case, statistical data structure a *data cube*, where each cell is identified based on the values of the dimensions and contains a measure or a set of measures.

Linked Data has been introduced as a technological paradigm for opening up data because it facilitates data integration across the Web. The term Linked Data refers to "*data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external datasets, and can in turn be linked to from external datasets*" [2].

In the case of cubes, Linked Data could enable the easy discovery and integration of multiple cubes on the Web and thus performing analytics on top of integrated but previously isolated cubes [10]. A fundamental step towards this vision is the data cube (QB) vocabulary, which enables modeling cubes as graphs [4]. During the last couple of years, a few sparse endeavors have been developed aiming at supporting the process of modeling data cubes according to the QB vocabulary. These components and tools, however, present some limitations regarding (a) the functionalities they provide, (b) their licenses that hamper commercial exploitation, (c) their dependencies to specific platforms

*Corresponding author: Evangelos Kalampokis, University of Macedonia, Egnatia 156, 54006, Thessaloniki, Greece. E-mail: ekal@uom.gr.

[1]http://open-data.europa.eu.

and environments, and (d) the capability to be used in complex scenarios in an integrated manner [11,12].

In this paper, we present the OpenCube Toolkit comprising a number of tools that aim at overcoming these limitations and provide a solution for linked data cube management. The methodology followed to develop the Toolkit is based on actively engaging organizations that deal with data cubes in real-world settings.

The rest of this paper is organized as follows. Section 2 presents the background of our work regarding open data, linked data, and data cubes. Section 3 describes the approach that we followed to develop the OpenCube Toolkit, while Section 4 presents the results of each step of our approach. Finally, Section 5 draws conclusions.

2. Background

2.1. Open Data

The term "Open Data" originates from some of the same roots as "Open Source" or "Open Access". Although "Open" in software normally means libre (i.e. free in the sense of having no restrictions), often "Open Access" is used as meaning gratis (i.e. free in the sense of costing no money). The GNU project suggests that Open Source (or Free) software is a matter of liberty, not price, and means that "the users have the freedom to run, copy, distribute, study, change and improve the software".

The European Commission defines open data as referring to the idea that certain data should be freely available for re-use [5]. This includes the use of the data for purposes foreseen or not foreseen by the original creator.

The World Bank categorizes the conditions that open data have to satisfy in two broad categories:

- Technically open: Available in a machine-readable standard format, which means "it can be retrieved and meaningfully processed by a computer application"
- Legally open: Explicitly licensed in a way that permits commercial and non- commercial use and re-use without restrictions.

McKinsey Global Institute suggests that open data share the following characteristics [13]:

- Accessibility: A wide range of users is permitted to access the data.
- Machine readability: The data can be processed automatically.
- Cost: Data can be accessed free or at negligible cost.
- Rights: Limitations on the use, transformation, and distribution of data are minimal.

For the purposes of this paper, open data is as defined by the Open Knowledge Foundation:[2] *"Open data is data that can be freely used, re-used and redistributed by anyone – subject only, at most, to the requirement to attribute and sharealike"*.

2.2. Linked data

Linked data is based on Semantic Web philosophy and technologies but in contrast to the full-fledged Semantic Web vision, it is mainly about publishing structured data using the Resource Description Framework (RDF) data model and Unified Resource Identifiers (URIs) rather than focusing on the ontological level or inferencing [7]. It promises the creation of the "Web of data" as data from decentralized and heterogeneous sources can be interlinked through typed links. Web of data aims at replacing data silos with a giant distributed dataset built on top of the Web architecture [8].

Linked Data following a RESTful approach require the identification of resources with URI references that can be dereferenced over the Hypertext Transfer Protocol (HTTP) into RDF data that describes the identified resource. Moreover, Linked Data include the creation of typed links between URI references, so that one can discover more data. More specifically, the four Linked Data principles as described by Berners-Lee [1] are the following:

- All items should be identified using URIs;
- All URIs should be dereferenceable, that is, using HTTP URIs allows looking up the item identified through the URI;
- When looking up a URI it leads to more data, which is usually referred to as the follow your nose principle;
- Links to other URIs should be included in order to enable the discovery of more data.

Linked data distinguishes between information and non-information resources. The latter refers to real world thing such as people, buildings, and public agencies, while the former refers to all the resources we find on the traditional document Web such as documents and images. The adoption of identifiers ensures uniquely identifying information resources on the Web

[2] http://opendefinition.org.

but not the real world things the information resources refer to. Hence, an important issue in the Web of data is finding identifiers that refer to the same real world thing. The use of Linked Data technologies for publishing data on the Web provides the following advantages:

- Enables data to be integrated with the Web. This describes the ability to link together different pieces of information published on the Web and the ability to directly reference a specific piece of information.
- Reduces the challenge of integrating heterogeneous data and building large-scale, *ad hoc* mashups.

The specification of the Linked Data principles resulted in the emergence of the Web of Linked Data, which currently comprises more than 1000 datasets in various domains [17]. The Linking Open Data (LOD) cloud diagram depicts this Web of Linked Data (Fig. 1). In the LOD cloud diagram the different datasets are depicted as bubbles and the connections between datasets as arrows. The direction of the arrows indicate the dataset that contains the links, e.g., an arrow from A to B means that dataset A contains RDF triples that use identifiers from B. Bidirectional arrows usually indicate that the links are mirrored in both datasets.

2.3. Linked data cubes

The multidimensional data model, which is often compared to a data cube, was introduced to define the analytic requirements of *Online Analytical Processing (OLAP)* and *data warehouse (DW)* systems. The notion of OLAP that were introduced by Codd [3] refers to the technique of performing complex analysis over information stored in a DW. A DW is a large data repository with integrated historical data organized specifically for analytical purposes.

In general, as described in [16] *dimensional concepts* structure the multidimensional space where the fact is placed. Dimensional concepts can be used as a perspective of analysis and have been classified as *dimensions*, *levels* and *descriptors*. A dimension is considered to contain a *hierarchy of levels* representing different granularities (or levels of detail) to study data, and a level to contain *descriptors*. On the other hand, a *fact* contains *measures* of analysis. One fact and several dimensions to analyze it give rise to a *multidimensional schema*. Finally, *base* is a minimal set of levels functionally determining a fact. Thus, two different instances of data cannot be placed in the same point of the multidimensional space.

The RDF Data Cube (QB) vocabulary is a $W3C$ standard for modelling data cubes as graphs and thus adhering to the RDF model and Linked Data principles. Centric class in the vocabulary is *qb:DataSet* that defines a cube. A cube has a *qb:DataStructureDefinition* that defines the structure of the cube and multiple *qb:Observation* that describe each cell of the cube. The structure is specified by the abstract *qb:ComponentProperty* class, which has three sub-classes, namely *qb:DimensionProperty*, *qb:MeasureProperty*, and *qb:AttributeProperty*. The first one defines the dimensions of the cube, the second the measured variables, while the third structural metadata such as the unit of measurement.

At the moment, 11.24% of the datasets on the Web of Linked Data use the QB vocabulary and thus regard *linked data cubes* [17]. Moreover, a number of datasets that contain linked data cubes have been also created. For example, the European Commission's Digital Agenda provides its Scoreboard[3] as linked data cubes. The linked data transformation[4] of Eurostat's data, which was created in the course of a research project, includes more than 5,000 linked data cubes. Census data of 2011 from Ireland and Greece and historical censuses from the Netherlands have been published as linked data cubes [14,15].

3. Approach

The approach that we follow to develop the OpenCube Toolkit requires the active engagement of organizations that deal with linked open statistical data (LOSD) in their everyday activities. These organizations mainly participate in the requirements identification and the evaluation of the developed set of tools. The methodology comprises five steps, while the focus of this paper is on the first four.

3.1. Requirements analysis

The first step deals with the identification and documentation of the needs of organizations that either have the mandate to collect and disseminate statistics or use statistics in decision-making processes. This step comprises the following activities:

[3] http://digital-agenda-data.eu/data.
[4] http://eurostat.linked-statistics.org.

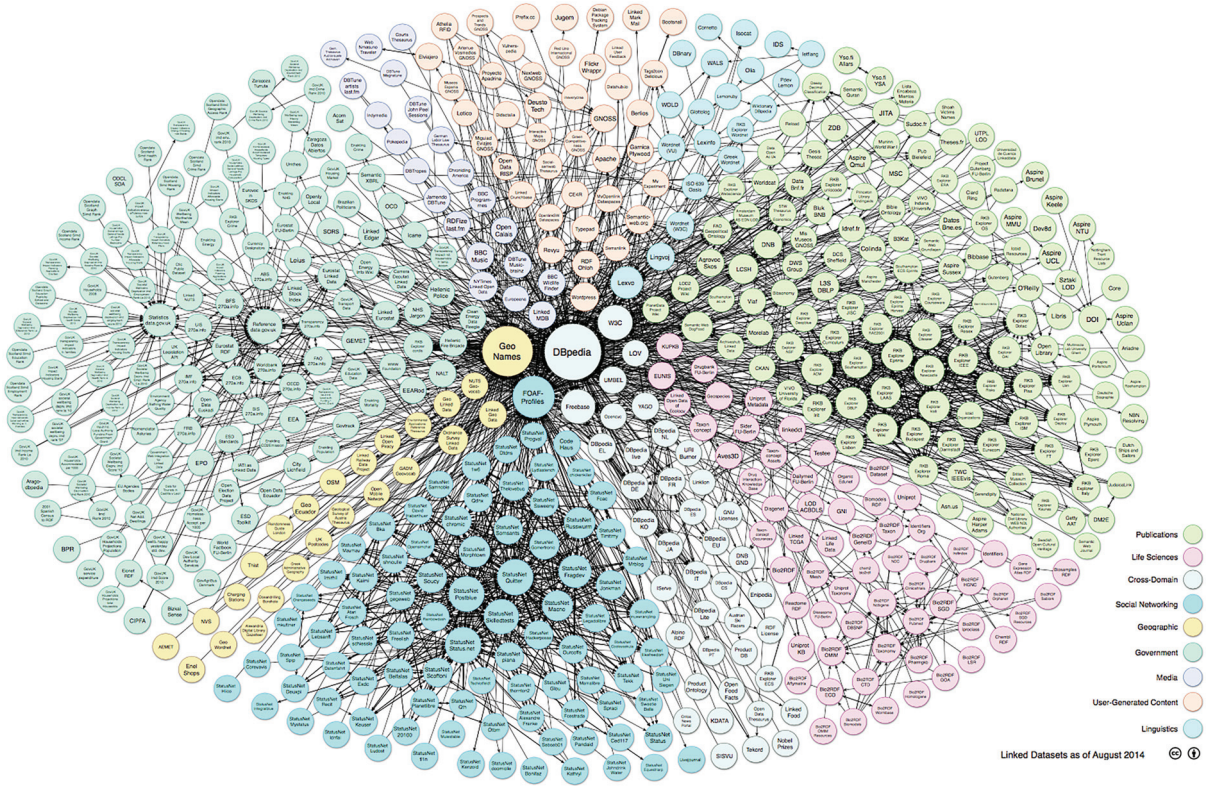

Fig. 1. The linked open data cloud (http://lod-cloud.net).

a. Review of existing linked data management tools and identification of their functionalities.
b. Literature review and analysis of cases that involve publishing and reusing of LOSD.
c. Interviewing employees from five organizations namely the UK Department for Communities and Local Government, the Research Centre of the Government of Flanders, the Open Data Team of the Flemish Government, the Irish Central Statistics Office, and a Swiss Bank.

3.2. First cycle of the OpenCube Toolkit development

This step deals with the actual development of the set of the tools and results in the first release of the Toolkit. The Information Workbench (IWB) platform [6] serves as a backbone for the OpenCube Toolkit. The components are integrated into a single architecture via standard interfaces provided by the IWB SDK: widgets (for UI controls) and data providers (for data importing and processing components). The overall UI design is based on the use of wiki-based templates providing dedicated views for RDF resources: an appropriate view template is applied to an RDF resource based on its type. All components of the architecture share the access to a common RDF repository (local or remote) and can retrieve data by means of SPARQL queries. Given the potentially large scale of data, which has to be processed, different data cubes can be stored in separate data repositories and queried using the SPARQL 1.1 federation capabilities.

The first release of the Toolkit includes the following tools [11]:

- *TARQL extension for data cubes*
- *D2RQ extension for data cubes*
- *Aggregator*
- *OpenCube Browser*
- *OpenCube MapView*
- *R Statistical Analysis Tool*

3.3. Evaluation of the first release of the OpenCube Toolkit

In this step, the first version of the Toolkit is tested and evaluated based on one of the most influential research models in information systems, namely the Technology Acceptance Model (TAM) [19] and its extensions. According to TAM, end-users' overall attitude

and intention toward using a system is a major determinant of whether they will actually use it. Towards this end, employees of the Research Unit of the Flemish Government were involved. We asked the evaluators to describe the system and/or its components according to the following criteria:

- Job Relevance (JR)
- Output Quality (OQ)
- Result Demonstrability (RD)
- Perceived Ease of Use (PEU)
- Perceived Usefulness (PU)
- Intention to Use (IU)

We should, however, note that we employ TAM to structure the interviews with the empoloyees and thus to receive a qualitative feedback. As a result, the final result was not a quantitative indication of the above criteria.

3.4. Second cycle of OpenCube Toolkit development

Based on the feedback received during the first cycle of evaluation the existing tools of the OpenCube Toolkit were improved while new tools were also created. This step resulted in the final version of the OpenCube toolkit.

3.5. Final evaluation of the OpenCube Toolkit

This step includes the evaluation of the final version of the OpenCube Toolkit. Because this is a very important step of our methodology with many details, it is not included in this paper.

4. Results

4.1. Requirements analysis

The requirements analysis resulted in a list of 56 functional and 13 non-functional requirements. Thereafter, the requirements were prioritized by the employees of the five organizations. This resulted in 35 functional and 3 non-functional requirements of high priority. Moreover, 15 functional and 6 non-functional requirements were characterized of medium priority while 6 functional and 4 non-functional of low priority.

This step also resulted in a lifecycle that describes the process that raw statistical data go through in order to create value based on linked data [18]. In particular, we consider that raw data go through a lifecycle that enables (a) creating, (b) expanding, and (c) exploiting LOSD. Figure 2 presents these three phases of the lifecycle and the respective steps that can be followed in each phase [18].

In particular, the first phase deals with transforming raw statistics into LOSD and addresses the following tasks:

- Discover & pre-process raw data in various data formats such as Comma Separated Values (CSV) files, spreadsheets, and Relational Databases (RDBMS).
- Create RDF data adhering to the QB vocabulary
- Manage and re-use controlled vocabularies (concept schemes, code lists etc.)
- Publish cubes through different interfaces i.e. Linked Data, SPARQL endpoint etc.
- Manage metadata.

The second phase deals with expanding LOSD by joining data cubes on the Web and addresses the following tasks:

- Discover compatible to join cubes on the Web of linked data.
- Establish typed links between compatible to join cubes.
- Create expanded cubes by increasing the size of one of the sets that define a cube i.e. measures, objects of a dimension's level, levels of a dimension, or dimensions.

The final phase deals with exploiting LOSD in data analytics and visualizations and considers the following tasks:

- Discover and explore LOSD.
- Perform OLAP operations on LOSD.
- Perform statistical analyses on LOSD e.g. compute descriptive statistics, calculate statistics such as correlation coefficient, and create learning models.
- Communicate results through visualizations.

4.2. First release of the OpenCube Toolkit

The tools included in the first version of the Toolkit are described below based on the three phases of the lifecycle.

4.2.1. Creating linked open statistical data

OpenCube tools that support the *creating phase* focus on enabling the user to transform legacy data into RDF data based on the QB vocabulary, to attach metadata allowing further search & discovery of relevant data, and to provide query access to data. These tools include:

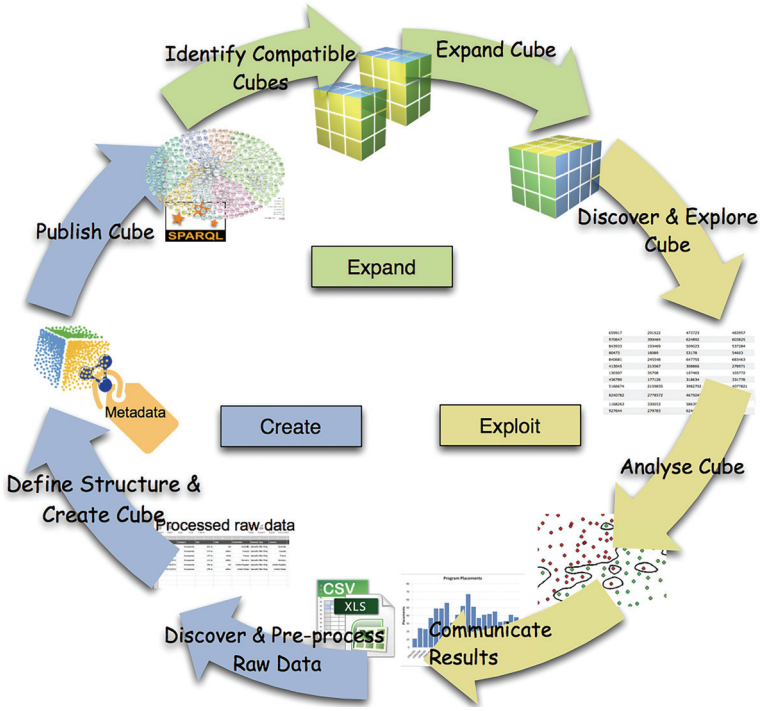

Fig. 2. The linked open statistical data lifecycle.

- *TARQL extension for data cubes*: data conversion to RDF according to QB vocabulary from legacy tabular data, such as CSV/TSV files.
- *D2RQ extension for data cubes*: data conversion to RDF according to QB vocabulary from relational databases.

4.2.2. *Expanding linked open statistical data*

In the first version of the Toolkit, one tool (termed Aggregator) was developed for linked data cube expansion. Its main role is to compute aggregations of existing cubes using an aggregate function. Three types of aggregate functions are distinguished in the literature: Σ, applicable to data that can be added together, φ, applicable to data that can be used for average calculations, and c, applicable to data that is constant, i.e., it can only be counted. Considering only the standard SQL aggregation functions, we have that $\Sigma = \{$SUM, COUNT, AVG, MIN, MAX$\}$, $\varphi = \{$COUNT, AVG, MIN, MAX$\}$ and $c = \{$COUNT$\}$. The aggregate function that can be applied to a cube depends on the following parameters:

- The dimensions and measures of a cube. For example, the SUM function can be applied to the sales measure over time, while it cannot be applied to the election results over time.
- The measure's unit of the cube. For example, if a cube's unit is "percentage" the SUM or AVG functions cannot be applied to the observations.

The aggregate functions described above can be applied to aggregate the cube observations. The OpenCube Aggregator distinguishes two categories of aggregation:

- Aggregation across a dimension. In this case, the observations are aggregated across one of the dimensions of the cube. For example, compute the SUM of the sales over time and thus ignore the time dimension of the cube. This type of aggregation enables the "Add Dimension" functionality of the OpenCube Browser (see below).
- Aggregation across a hierarchy. In this case the observations are aggregated across a hierarchy of a dimension. For example, if a cube contains the election results at municipality level, then the Aggregator can compute the results at region and at country level with the prerequisite that the corresponding hierarchy (municipality $\rightarrow$ region $\rightarrow$ country) exists. Note that the opposite is not possible, i.e. to go down from the country level to regions and municipality.

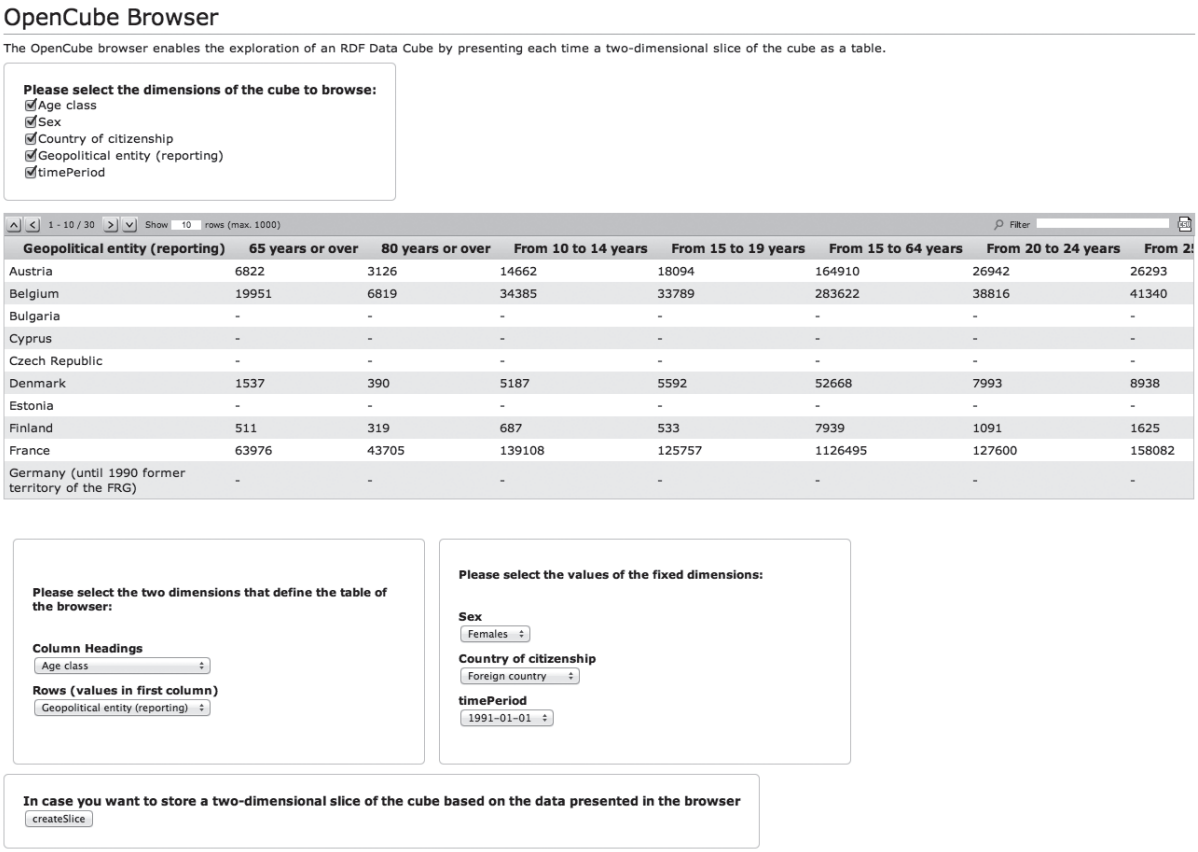

Fig. 3. The OpenCube browser.

4.3. Exploiting linked open statistical data

4.3.1. OpenCube browser

The OpenCube browser enables exploring LOSD and supports the following functionalities:

1. It presents in a table the values of a two-dimensional slice of a linked data cube. The user can change the number of rows of the table (by default the browser presents 20 rows per page).
2. The user can change the two dimensions that define the table of the browser.
3. The user can change the values of the fixed dimensions (i.e. the dimensions of the cube that are not shown in the table) and thus select a different slice to be presented.
4. The user can remove dimensions of the cube to browse. This functionality is supported only for cubes having at least one aggregatable measure.
5. The user can create and store a two-dimensional slice of the cube based on the data presented in the browser.

By default, the OpenCube Browser defines and presents a two-dimensional slice of the cube in the following way:

It assumes that all the dimensions of the cube will be included in the browser.

- It selects the largest dimension as rows dimension.
- It randomly selects the columns dimension.
- It sets a fixed value for each of the other dimensions (the first value as it appears).
- It randomly selects one measure (in the case of cubes having multiple measures).

In Fig. 3 the interface of the OpenCube Browser is depicted. On the top of the page the user can select the dimensions of the cube to browse. In particular, the check boxes enable the insertion or reduction of dimensions. Below the check boxes the actual table is presented while below the table the drop-down lists enable users to change the dimensions that are presented in the table and the values of the fixed dimensions. Finally, at the bottom of the page the use can create and store a slice as this is presented in the browser.

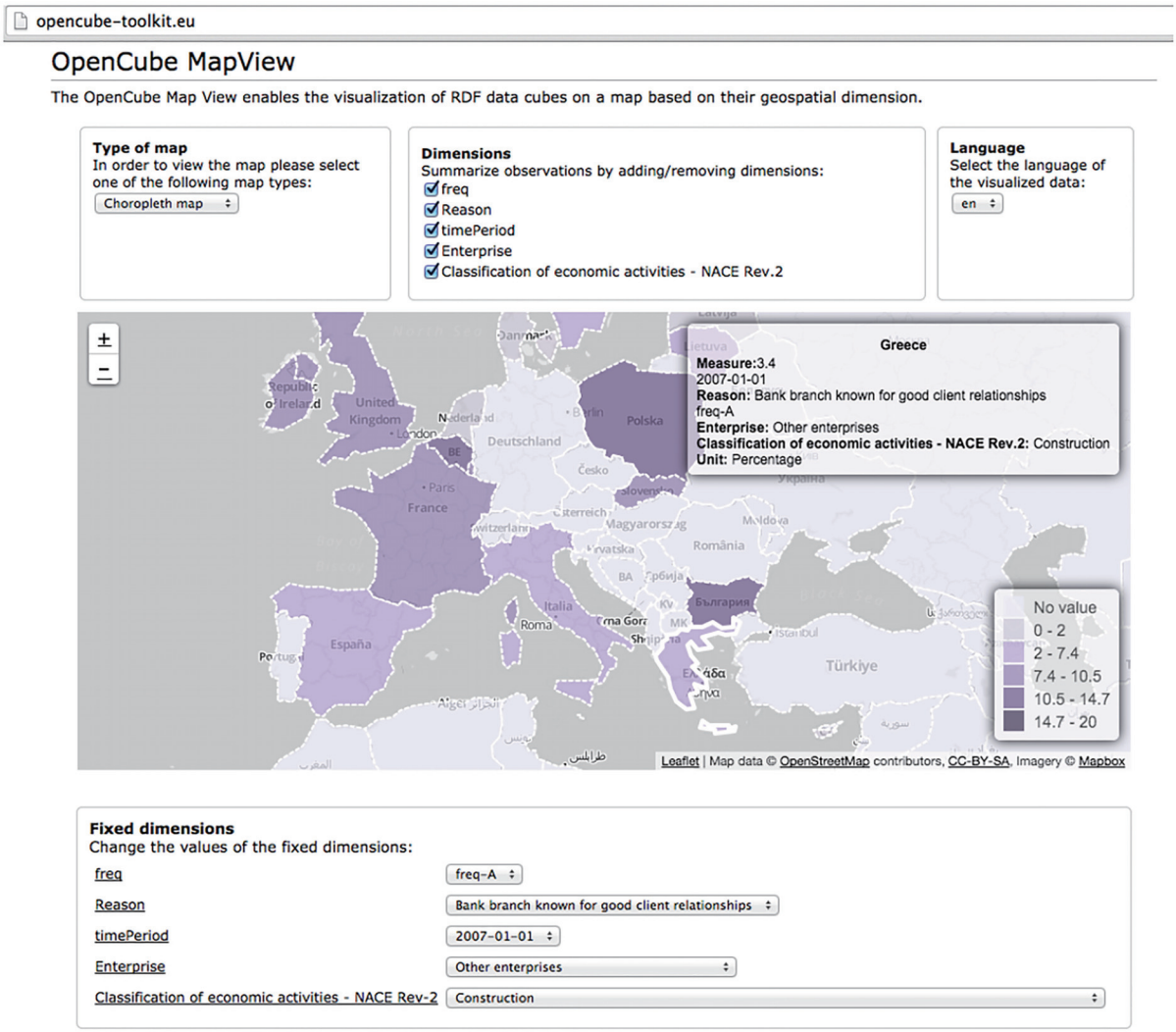

Fig. 4. OpenCube MapView.

4.3.2. *OpenCube MapView*

The OpenCube MapView enables the visualization of LOSD on a map based on their geospatial dimension. In the first release the MapView supports markers, bubbles and choropleth maps. In Fig. 4 a data cube is visualized on a map based on its geospatial dimension property using a choropleth heat map.

4.3.3. *R statistical analysis tool*

This tool enables implementing various statistical analysis methods on top of linked data cubes by integrating the R package in the underlying open source linked data management platform adopted by OpenCube. R is run as a web service (using Rserve package) and accessed via HTTP. Input data are retrieved using SPARQL queries and passed to R together with an R script provided by the user. Then, R capabilities can be exploited in two modes: (i) as a widget (the script generates a chart, which is then shown on the wiki page) and (ii) as a data source (the script produces a data frame, which is then converted to RDF using defined R2RML mappings and stored in the data repository).

4.4. *First cycle of evaluation*

In general, the feedback received by the employees of the Flemish Government should be understood in the context of a department seeking to replace an existing solution, which is expensive and not user friendly. Although the overall feedback was positive the follow-

ing remarks and comments for improvement were expressed:

- The multilinguality of the platform was considered as a very important feature.
- Although the performance of the platform was considered acceptable, some users requested better response time.
- The users requested to be able to perform drill-down and roll-up operations over hierarchical code lists (e.g. in geo-spatial dimensions to be able to move across different levels i.e. municipality → district → province → region). They asked for this feature in both OpenCube Browser and MapView.
- The users suggested that the interface of the OpenCube Browser and MapView is not clear and easy to use. They proposed to bring all configuration widgets above the table and a dynamically adapted title should describe what is shown.
- The users suggested that the dimension insertion and removal feature is not clear for an average user e.g. a citizen.
- The users requested a feature allowing combining measures in a table (showing more than 2 dimensions in the table).
- The users requested an additional export facility to MS-Excel next to csv.
- The users requested a feature enabling to define the legend of the choropleth map themselves including the ability to add explanations.

Moreover, we should note that the employees of the Flemish Government evaluated OpenCube toolkit in relation to several demos of relevant tools. In this context, their attitude towards OpenCube is best summarized with a quote from an evaluation form: *"We don't see added value compared to other tools"*. The rationale was that, for the moment, the promise of providing added value through LOSD integration across the Web was not visible yet.

Summarizing, the main points expressed in the first phase of evaluation are the following:

- The performance of the tools needs to be enhanced.
- Much more attention should be drawn to usability.
- OLAP operations should be enabled in the next phase of the Browser and MapView.
- LOSD integration should be available in a transparent to the user manner.

4.5. Second release of the OpenCube Toolkit

4.5.1. Creating linked open statistical data

During the second release and based on users feedback two new tools were developed. These tools support (a) the JSON-stat data format, and (b) the R2RML mapping language. In particular:

- JSON-stat to QB tool: data conversion to RDF according to QB vocabulary from JSON-stat files. The JSON-stat[5] format is a simple lightweight JSON format and it is based on a cube model that arises from the evidence that the most common form of data dissemination is the tabular form.
- R2RML tool: transformation of relational data into RDF data cubes using the extended R2RML mappings language.[6] R2RML is a language for expressing customized mappings from relational databases to RDF datasets

4.5.2. Expanding linked open statistical data

The main criticism during the first evaluation was that integration of LOSD across the Web was not supported. Statistics integration justifies the need for exploiting linked data technologies. In this context, two new tools were developed that support the identification and integration of statistics in the form of linked data cubes on the Web.

OpenCube compatibility explorer

The main role of the OpenCube compatibility explorer is to (a) identify compatible to merge cubes and (b) establish typed links to facilitate discovery. The OpenCube Compatibility Explorer mainly deals with two merge-related operations:

- Add measure. An expansion cube is compatible to add a new measure to an original cube if: (i) both cubes have the same dimensions, (ii) the expansion cube has at least the same values at each dimension of the original cube (it may contain and more values than the original cube) and iii) the expansion cube has at least one measure that does not exist at the original cube.
- Add value to dimension. An expansion cube is compatible to add a new value to a dimension of an original cube if: (i) both cubes have the same dimensions, (ii) both cubes have the same measures and (iii) the expansion cube has at least one

[5] http://json-stat.org.
[6] http://www.w3.org/TR/r2rml/.

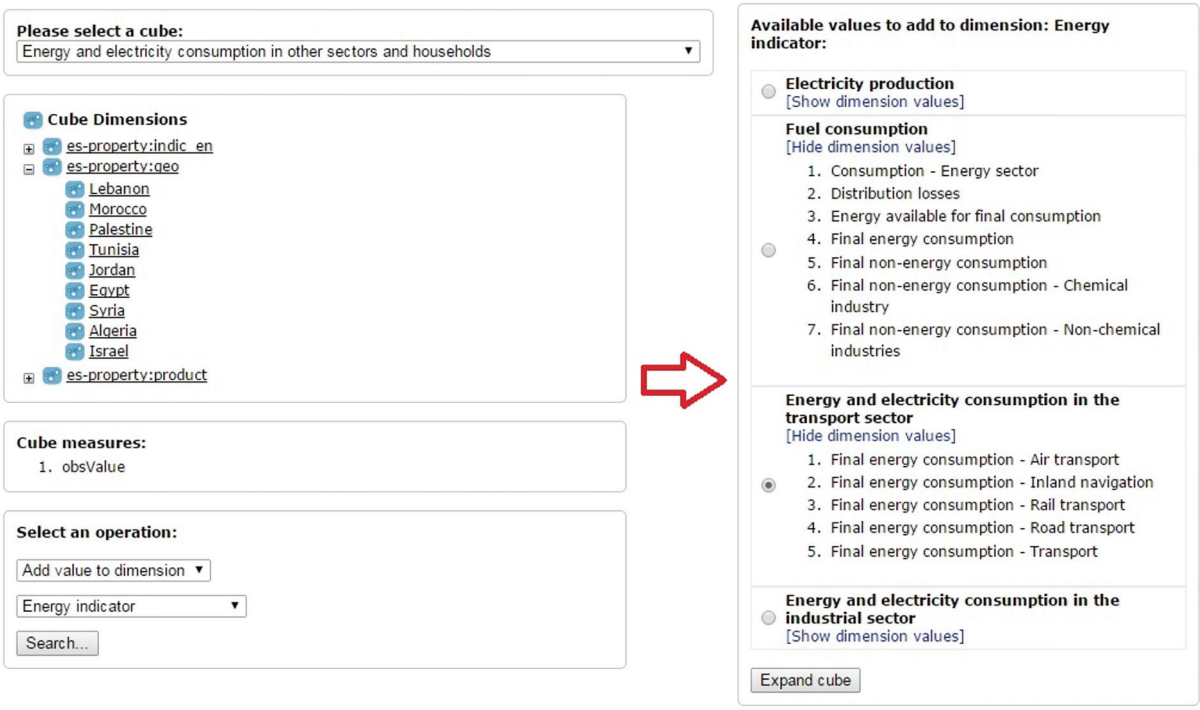

Fig. 5. OpenCube expander user interface: adding dimension values.

more value than the original cube at the expansion dimension and has the same values with the original cube at all remaining dimensions.

The OpenCube Compatibility Explorer after detecting compatible cubes based on the compatibility types presented above, creates links in order to be able to easily identify compatibility when requested (e.g. when browsing a cube).

OpenCube expander

The OpenCube Expander (a) searches for compatible cubes and (b) creates a new expanded cube by merging two compatible cubes. The functionality of the OpenCube Expander is based:

- On the links created by the OpenCube Compatibility Explorer in order to detect external compatible cubes.
- On the aggregations (across a dimension and across a hierarchy) to detect compatible precomputed aggregate cubes. The links enable the fast detection of the compatible cubes since no complex computations are made.

When launched, this tool starts by presenting the structure of the cube (Fig. 5), i.e.: (i) the cube dimensions, (ii) the values for each dimension, and (iii) the cube measures. Thereafter, the user can search for compatible cubes based on the following operations:

1. Add measure. This operation identifies and presents cubes that are compatible to add new measures to the original cube.
2. Add value to dimension. In this case the user selects an expansion dimension and the operation identifies and presents compatible cubes that can be used to add new values to the selected dimension.
3. Add hierarchy. This operation identifies and presents cubes that are compatible to add a hierarchy to the original cube i.e. pre-computed aggregations across a hierarchy created by the OpenCube Aggregator.
4. Add dimension. This operation identifies and presents cubes that are compatible to add a dimension to the original cube i.e. pre-computed aggregations across a dimension created by the OpenCube Aggregator.

The output of each of the above operations is a new merged cube that can then be used by other tools. For example the OpenCube OLAP Browser can be used to show the new merged cube. However, the creation of a new cube could require considerable time depending on the size of the compatible cubes to be merged. As

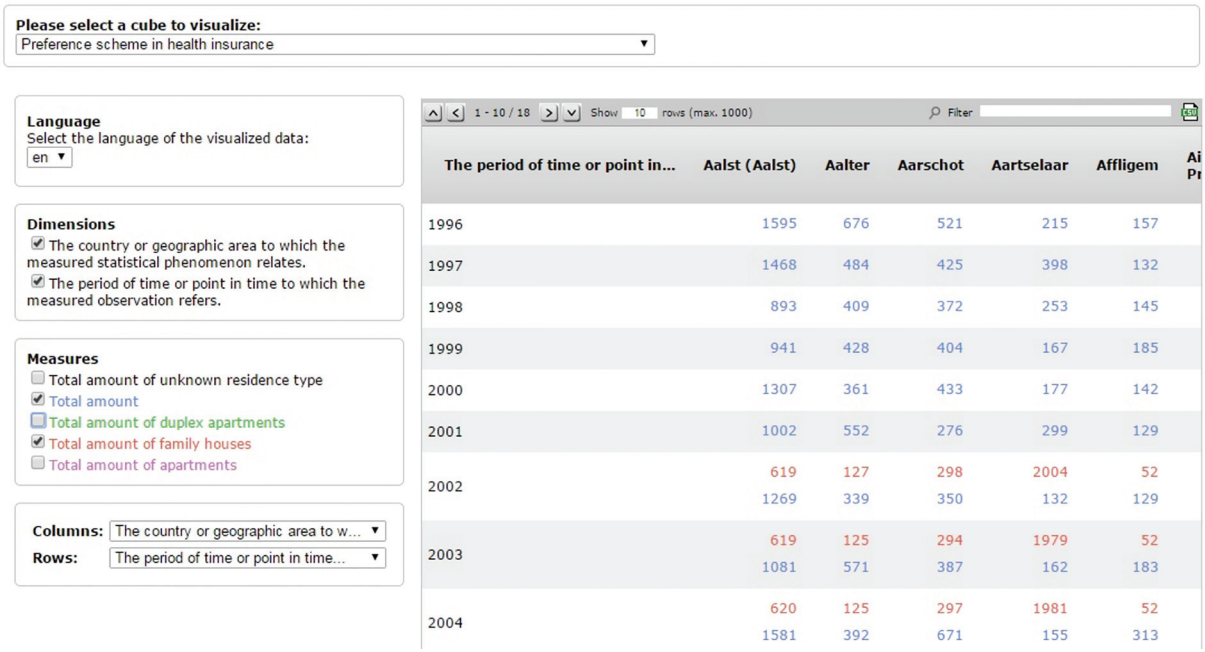

Fig. 6. The OpenCube OLAP browser.

a result, a part of the OpenCube Expander functionality is integrated to the OpenCube OLAP Browser. This enables viewing compatible cubes on the fly without the need to explicitly create new merged cube(s).

4.5.3. Exploiting linked open statistical data

Based on the feedback received during the first evaluation cycle the exploitation-related tools were improved and some new were developed. In this section we describe the OpenCube OLAP Browser, which is the second generation of the OpenCube Browser. We should note, however, that these tools are complementary and thus the former does not replace the latter.

OpenCube OLAP Browser

The OpenCube OLAP Browser introduces a more user-friendly, simple and intuitive interface. All the control operations (e.g. language select, selection on dimensions/measures) are presented together on the left, while the table view is presented on the right.

When launched, the OpenCube OLAP Browser (Fig. 6) presents only the structure of the cube (available dimensions and measures). Then, the user has to select at least one dimension and one measure to visualize. This visualization approach is more intuitive since it gives more control to the user. Moreover, the OpenCube OLAP Browser enables users to perform typical OLAP operations, such as drill-down and roll-up, on top of linked data cubes.

One of the main enhanced functionalities of the OpenCube OLAP Browser is the visualization of multiple cubes. This functionality enables the integrated view of compatible cubes on the fly without the need to create a new merged cube by the OpenCube Expander, thus saving execution time and improving the performance. In this case, the OpenCube Expander component passes as parameters to the OpenCube OLAP Browser the two compatible cubes to visualize together.

5. Conclusion

A major part of Open Data concerns statistics such as financial and social indicators. Accurate and reliable statistics provide the solid ground for performing analyses that support businesses and governments in understanding the world and making better decisions. The adoption of the Linked Data principles and technologies has promised to enhance the analysis of statistical data at a Web scale.

This article presented the OpenCube Toolkit developed to enable easy creating, expanding, and exploiting Linked Open Statistical Data formed as data cubes. The Toolkit integrates components dealing with dif-

ferent steps of the linked data cube lifecycle in order to provide the user with a rich set of functionalities for working with statistical semantic data. At the creating phase, the main focus is on supporting the user in transforming legacy data (such as CSV or relational databases) into RDF data cubes, attaching metadata allowing further search & discovery of relevant data, and providing query access to them. At the expanding phase, the toolkit enables the discovery of compatible to merge cubes and the creation of expanded cubes. At the exploiting phase of the lifecycle, the toolkit enables linked data cubes browsing and exploration as well as performing data analytics on top of them in an easy manner.

The tools were evaluated by organizations the employ data cubes in their everyday activities.

Acknowledgments

The work presented in this paper was partially carried out in the course of the OpenCube[7] project, which is funded by the European Commission within the 7th Framework Programme under grand agreement No. 611667. The authors would like to thank the whole OpenCube consortium that contributed to the development and evaluation of the OpenCube toolkit.

References

[1] T. Berners-Lee, Design issues: Linked data, 2006. URL http://www.w3.org/DesignIssues/LinkedData.html.

[2] C. Bizer, T. Heath and T. Berners-Lee, Linked Data – The Story So Far, Special Issue on Linked Data, *International Journal on Semantic Web and Information Systems* **5**(3) (2009), 1–22.

[3] E. Codd, S. Codd and C. Salley, Providing OLAP (Online Analytical Processing) to User-analysts: An IT Mandate. Codd & Associates, 1993.

[4] R. Cyganiak and D. Reynolds, The RDF Data Cube vocabulary, http://www.w3.org/TR/vocab-data-cube/ (2013).

[5] European Commission. Open data: An engine for innovation, growth and transparent governance. Communication from the Commission, COM (2011) 882 final, December 2011.

[6] P. Haase, M. Schmidt and A. Schwarte, The Information Workbench as a Self-Service platform for Linked Data Applications, in: *COLD 2011, ISWC 2011*, Shanghai, China (2011).

[7] M. Hausenblas, Exploiting linked data to build web applications *IEEE Internet Computing* **13**(4) (2009), 68–73.

[8] T. Heath, How will we interact with the web of data? *Internet Computing, IEEE* **12**(5) (2008), 88–91.

[9] E. Kalampokis, E. Tambouris and K. Tarabanis, A Classification Scheme for Open Government Data: Towards Linking Decentralized Data, *International Journal of Web Engineering and Technology* **6**(3) (2011), 266–285.

[10] E. Kalampokis, E. Tambouris and K. Tarabanis, Linked open government data analytics, in: *EGOV2013, LNCS, 8074*, M.A. Wimmer, M. Janssen and H.J. Scholl, eds., Springer, 2013, pp. 99–110.

[11] E. Kalampokis, A. Nikolov, P. Haase, R. Cyganiak, A. Stasiewicz, A. Karamanou, M. Zotou, D. Zeginis, E. Tambouris and K. Tarabanis, *Exploiting Linked Data Cubes with OpenCube Toolkit, Proc. of the ISWC 2014 Posters and Demos Track a track within 13th International Semantic Web Conference (ISWC2014), 19-23 October 2014, Riva del Garda, Italy, CEUR-WS Vol.* 1272 (2014).

[12] E. Kalampokis, A. Karamanou, A. Nikolov, P. Haase, R. Cyganiak, B. Roberts, P. Hermans, E. Tambouris and K. Tarabanis, *Creating and Utilizing Linked Open Statistical Data for the Development of Advanced Analytics Services, Proc. of the 2nd International Workshop on Semantic Statistics (SemStats2014) in conjunction with the 13th International Semantic Web Conference (ISWC2014), 19–23 October 2014, Riva del Garda, Italy, CEUR-WS proceedings*, 2014.

[13] J. Manyika, M. Chui, P. Groves, D. Farrell, S. van Kuiken and E.A. Doshi, Open data: Unlocking innovation and performance with liquid information. Technical report, McKinsey & Company, October 2013.

[14] A. Meron o-Penuela, A. Ashkpour, L. Rietveld and R. Hoekstra, Linked humanities data: The next frontier? a case-study in his-torical census data, in Proceedings of the 2nd International Workshop on Linked Science 2012, vol. 951, 2012.

[15] I. Petrou, G. Papastefanatos and T. Dalamagas, *Publishing census as linked open data: A case study, in Proceedings of the 2Nd International Workshop on Open Data, ser. WOD '13*. New York, NY, USA: ACM, 2013, pp. 4:1–4:3.

[16] O. Romero and A. Abello, A survey of multidimensional modeling methodologies, *International Journal of Data Warehousing and Mining (IJDWM)* **5**(2) (2009), 1–23.

[17] M. Schmachtenberg, C. Bizer and H. Paulheim, Adoption of the linked data best practices in different topical domains. In P. Mika et al., eds, *The Semantic Web – ISWC 2014, volume 8796 of Lecture Notes in Computer Science, Springer International Publishing*, 2014, pp. 245–260.

[18] E. Tambouris, E. Kalampokis and K. Tarabanis, Processing Linked Open Data Cubes, in: *EGOV2015, LNCS 9248*, E. Tambouris et al., eds, Springer, 2015, pp. 130–143.

[19] V. Venkatesh, M.G. Morris, G.B. Davis and F.D. Davis, User acceptance of information technology: Toward a unified view, *MIS Quarterly* (2003), 425–478.

[7] http://www.opencube-project.eu.