ICT Tools for statistical linked open data:

The OpenCube toolkit

Efthimios Tambouris (tambouris@uom.gr)[1], Evangelos Kalampokis[1] and Konstantinos Tarabanis[1]

Keywords: Linked Data, multi-dimensional data, data cube, data analytics, visualization.

1. INTRODUCTION

The recent Open Data movement results in an increasing number of data offered via the Web. A significant part of these data concerns statistics. The ability to manage statistical data at a Web scale provides unprecedented analysis opportunities. Imagine if analysts could easily find statistical data coming from different sources all over the world, integrate, analyse and visualise them in any way they want.

Yet, we are currently lacking an overall understanding, including ICT tools, to enable us reaping the benefits of open statistical data. Nevertheless, at the technological level, linked data is an emerging technology with potential to overcome some of the current limitations.

In this abstract, we present the OpenCube toolkit [1] that aims at overcoming current limitations by (a) supporting the full life-cycle of statistical linked open data management from production to exploitation, and (b) providing open-source ICT tools to support advanced functionalities such as OLAP analysis, statistical analysis, data integration, and map visualizations based on linked data technologies.

2. STATISTICAL DATA AND LINKED DATA

Governments, organisations and companies are increasingly launching data portals that operate as single points of access for data they produce or collect [2]. A major part of these open data concerns statistics, such as population figures, economic and social indicators. For example, the vast majority of datasets published on the European Commission open data portal[2] are of statistical nature. In addition, many international organizations such as Eurostat[3], World Bank[4], OECD[5] and CIA's World Factbook[6] open up statistical data on the Web.

Statistical data is often organized in a multidimensional manner where a measured fact is described based on a number of dimensions, e.g. the unemployment rate can be described based on geographic area, time and gender. In this case, statistical data is compared to a data cube, where each cell contains a measure or a set of measures, and thus we onwards refer to statistical multidimensional data as *data cubes* or just *cubes*.

Linked data has been introduced as a promising paradigm for opening up data because it facilitates the integration of data 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*

[1] University of Macedonia and ITI-CERTH, Thessaloniki, Greece
[2] http://open-data.europa.eu
[3] http://ec.europa.eu/eurostat/data/database
[4] http://data.worldbank.org
[5] http://www.oecd.org/statistics/
[6] https://www.cia.gov/library/publications/the-world-factbook/index.html

datasets" [3]. 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 in RDF using URIs rather than focusing on the ontological level or inference.

In the case of cubes, linked data has the potential to realize the vision of integrating disparate and previously isolated data to perform analytics. A fundamental step towards this vision is the RDF data cube (QB) vocabulary, which enables modeling cubes as RDF graphs [4]. Recently, a few endeavors aimed at supporting data modeled according to the QB vocabulary. The resulting 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 and environments, and (d) their ability to be used in complex scenarios in an integrated manner.

3. THE OPENCUBE LIFECYCLE

Exploiting statistical open data using linked data technologies calls for advances in relevant business processes. Figure 1 depicts a proposed general lifecycle that illustrates how linked data technologies can be integrated in organisations' data management business processes [5]. The lifecycle steps are categorized in two broad phases (a) the *publish phase* that includes creating linked data cubes out of raw data, and (b) the *reuse phase* that includes exploiting linked data cubes in advanced analytics and visualizations.

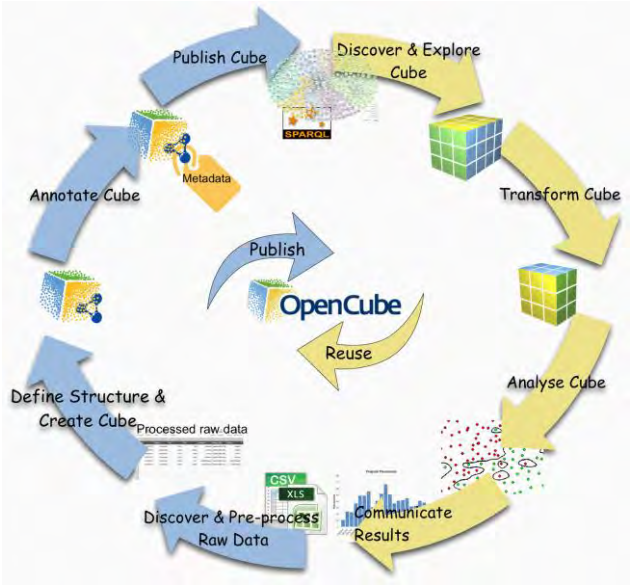

Figure 1. The OpenCube lifecycle

4. THE OPENCUBE TOOLKIT

The OpenCube toolkit includes ICT tools that support all steps in the relevant lifecycle. OpenCube tools that support the *publish phase* include:

- *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.
- JSON-Stat to QB: data conversion to RDF according to QB vocabulary from JSON-Stat files.

OpenCube tools that support the *reuse phase* include:

- *Data catalogue management* that provides user interface templates for managing metadata on linked data cubes and supporting search and discovery of cubes from various sources.
- *OpenCube Browser* that enables the exploration of a linked data cube by presenting two-dimensional slices of the cube in tabular form. This tool also supports OLAP-like operations such as dimension reduction.
- *OpenCube MapView* that enables the visualization of linked data cubes on a map based on their geospatial dimension. Currently, the MapView supports markers, bubbles and choropleth maps. It also supports OLAP-like operations to enable visualizing various views of a cube. For example, in Figure 2 a data cube is visualized using a choropleth heat map based on its geospatial dimension property.
- *Interactive chart visualization tool* that enables linked data cubes visualizations by summarizing data and creating charts.
- *OpenCube Aggregation tool* that pre-computes aggregations across dimensions in order to enable OLAP operations in the Browser and the MapView.
- *R statistical analysis tool* that 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.

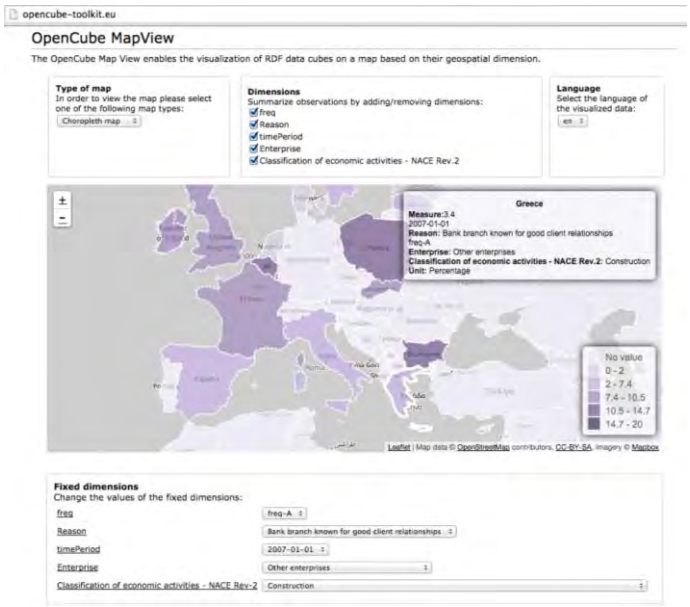

Figure 2. Visualization of a data cube that includes a geospatial dimension with the OpenCube MapView

An important benefit of the OpenCube toolkit when compared to more traditional approaches is cubes integration. This enables expanding a cube by integrating it with a second (compatible) cube. We suggest a cube can be expanded if it is possible to increase the size of one of the sets that define a cube i.e. the set of measures, the set of objects of an attribute (level) of a dimension, the set of attributes of a dimension, or the set of dimensions. In particular, the toolkit enables (a) finding cubes on the Web of linked data that are compatible to expand an initial cube, and (b) creating and storing a new expanded linked data cube from the initial one. The expanded cube can be thereafter analyzed and visualized using the rest of the OpenCube tools.

From a technical point of view, the OpenCube toolkit is based on the Information Workbench community edition platform[7], which is a linked data management platform. All the tools in the toolkit share access to a common RDF repository and can retrieve data by means of SPARQL queries. The user interface design is based on the use of wiki-based templates providing dedicated views for RDF resources.

5. CONCLUSIONS

A major part of open data concerns statistics that can be structured as multi-dimensional data cubes. Linked data technologies have the potential to realize the vision of finding, combining, analysing and visualizing previously isolated cubes at a Web scale. A fundamental step towards this vision is the RDF data cube (QB) vocabulary, which enables modeling cubes as RDF graphs. Current tools fall short to support the whole linked data cube lifecycle including the integration and analysis of multiple cubes. In this abstract, we presented the OpenCube Toolkit, a set of open source tools that cover the whole linked data cubes lifecycle in an integrated manner. The work reported here will continue by further improving existing tools, developing new ones and piloting them at various organizations across Europe.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 611667

REFERENCES

[1] E. Kalampokis, A. Nikolov, P. Haase, R. Cyganiak, A. Stasiewicz, A. Karamanou, M. Zotou, D. Zeginis, E. Tambouris, 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).

[2] E. Kalampokis, E. Tambouris, 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.

[3] 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 (IJSWIS), 5(*3), (2009), 1-22.

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

[5] E. Kalampokis, A. Karamanou, A. Nikolov, P. Haase, R. Cyganiak, B. Roberts, P. Hermans, E. Tambouris, K. Tarabanis (2014) 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.

[7] http://www.fluidops.com/en/company/training/open_source