Exploiting Linked Data Cubes with OpenCube Toolkit

Evangelos Kalampokis[1,2], Andriy Nikolov[3], Peter Haase[3], Richard Cyganiak[4], Arkadiusz Stasiewicz[4], Areti Karamanou[1,2], Maria Zotou[1,2], Dimitris Zeginis[1,2], Efthimios Tambouris [1,2], Konstantinos Tarabanis[1,2]

[1] Centre for Research & Technology - Hellas, 6th km Xarilaou-Thermi, 57001, Greece
[2] University of Macedonia, Egnatia 156, 54006 Thessaloniki, Greece
{ekal, akarm, mzotou, zegin, tambouris, kat}@uom.gr
[3] fluid Operations AG, Altrottstraße 31, 69190 Walldorf, Germany
{andriy.nikolov, peter.haase}@fluidops.com
[4] Insight Centre for Data Analytics, Galway, Ireland
{richard.cyganiak, arkadiusz.stasiewicz}@insight-centre.org

Abstract. The adoption of the Linked Data principles and technologies has promised to enhance the analysis of statistics at a Web scale. Statistical data, however, is typically organized in data cubes where a numeric fact (aka measure) is categorized by dimensions. Both data cubes and linked data introduce complexity that raises the barrier for reusing the data. The majority of linked data tools are not able to support or do not facilitate the reuse of linked data cubes. In this demo we present the OpenCube Toolkit that enables the easy publishing and exploitation of linked data cubes using visualizations and data analytics.

Keywords: Linked data, statistics, data cubes, visualization, analytics.

1 Introduction

A major part of Open Data concerns statistics such as population figures, economic and social indicators. Analysis of statistical open data can provide value to both citizens and businesses in various areas such as business intelligence, epidemiological studies and evidence-based policy-making. Linked Data has emerged as a promising paradigm to enable the exploitation of the Web as a platform for data integration. As a result Linked Data has been proposed as the most appropriate way for publishing open data on the Web. Statistical data needs to be formulated as RDF data cubes [1] characterized by dimensions, slices and observations in order to unveil its full potential and value [2]. Processing of linked statistical data has only become a popular research topic in the recent years. Several practical solutions have been developed in this domain: for example, the LOD2 Statistical Workbench[1] brings together components developed in the LOD2 project by means of the OntoWiki[2] tool.

[1] http://wiki.lod2.eu/display/LOD2DOC/LOD2+Statistical+Workbench
[2] http://aksw.org/Projects/OntoWiki.html

In this demo paper we describe the OpenCube Toolkit that enable users to work with linked data cubes in an easy manner. In comparison with existing tools, our toolkit provides the following contributions:

- application development SDK allowing customized domain-specific applications to be built to support various use cases;
- new functionalities enabling users to better exploit linked data cubes;
- components supporting the whole linked data cube lifecycle.

2 OpenCube Toolkit

The OpenCube Toolkit[3] integrates a number of components which enable the user to work with semantic statistical data at different stages of the lifecycle: from importing legacy data and exposing it as linked open data to applying advanced visualization techniques and complex statistical methods to it.

The Information Workbench (IWB) platform [3] serves as a backbone for the toolkit components. 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.

The OpenCube Toolkit demo uses datasets from the Linked Data version of Eurostat[4] and can be currently accessed using the following link: http://data.fluidops.net.

2.1 Using the OpenCube Toolkit for data import, transformation, and publishing

Much of the relevant valuable statistical data are only available in various legacy formats, such CSV and Excel. To present these data in the form of linked RDF data cubes, they have to be imported, transformed into the RDF Data Cube format and made accessible for querying.

The **OpenCube TARQL**[5] component enables cubes construction from legacy data via TARQL (Transformation SPARQL): a SPARQL-based data mapping language that enables conversion of data from RDF, CSV, TSV and JSON (and potentially XML and relational databases) to RDF. TARQL is tool for converting CSV files to RDF using SPARQL 1.1 syntax. It is built on top of Apache ARQ[6]. The OpenCube TARQL component includes the new release of TARQL. It brings several improvements, such as: streaming capabilities, multiple query patterns in one

[3] http://opencube-toolkit.eu
[4] http://eurostat.linked-statistics.org
[5] https://github.com/cygri/tarql
[6] http://jena.apache.org/documentation/query/

mapping file, convenient functions for typical mapping activities, validation rules included in mapping file, increased flexibility (dealing with CSV variants like TSV).

The R2RML[7] language is a W3C standard for mappings from relational databases to RDF datasets. D2RQ[8] is a platform for accessing relational databases as virtual, read-only RDF graphs. **D2RQ Extensions for Data Cube** cover the functionality of importing of raw data as data cubes by mapping raw data to RDF. The process of mapping the data cube with a relational data source includes: (a) mapping the tables to classes of entities, (b) mapping selected columns into cube dimensions and cube measures, (c) mapping selected rows into observation values, and (d) generate triples with data structure definition. The user, by providing information about the dataset, such as the data dimensions and related measures, will receive an R2RML mapping file, which as a result will be used to generate and store the output.

2.2 Using the OpenCube Toolkit to utilize statistical data

To make use of available statistical data cubes, the user requires, as a minimum, to be able to explore and, visualize the data. The next step involves being able to apply to these data relevant statistical analysis methods.

The **OpenCube Browser** enables the exploration of an RDF data cube by presenting two-dimensional slices of the cube as a table. Currently browser enables users to change the two dimensions that define the table of the browser and also change the values of the fixed dimensions and thus select a different slice to be presented. Moreover, the browser supports roll-up and drill-down OLAP operations through dimensions reduction and insertion respectively. Finally, the user can create and store a two-dimensional slice of the cube based on the data presented in the browser. Initially, the browser selects two dimensions to present in the table and sets up a fixed value for all other dimensions. Based on these it creates and sends a SPARQL query to the store to retrieve the appropriate data. For the drill-down and roll-up operations the browser assumes that a set of data cubes has been created out of the initial cube by summarizing observations across one or more dimensions. We assume that these cubes define an Aggregation Set.

The **OpenCube Map View** enables the visualization of RDF data cubes on a map based on their geospatial dimension. Initially, Map View presents to the user the supported types of visualization (including markers, bubbles, choropleth and heat maps) along with all the dimensions and their values in drop-down lists.

The user selects the type of visualization and a map appears that actually visualizes a one-dimension slice of the cube where the geospatial dimension is free and the other dimensions are randomly “fixed”. In addition, the user can click on an area or marker or bubble and see the details of the specific observation. The maps are created using OpenStreetMap[9] and Leaflet[10] open-source library.

---

[7] http://www.w3.org/TR/r2rml/

[8] http://d2rq.org/

[9] http://wiki.openstreetmap.org/wiki/Develop

[10] http://leafletjs.com/

To allow the user explore the data in a data cube, it is important that the used visualization controls are (i) interactive and (ii) adapted to the cube data representation. In this way the user can easily switch between different slices of the cube and compare between them. To this end, we implemented our **Chart-based Visualization** functionality. The charts can be inserted into a wiki page of an RDF resource and configured to show data cube slices. When viewing the page, the user can change the selection of dimension values to change the visualised cube slices. The SPARQL query to retrieve the appropriate data is constructed based on the slice definition, and the data is downloaded from the SPARQL endpoint dynamically.

When working with statistical data, a crucial requirement is the possibility to apply specialized analysis methods. One of the most popular environments for statistical data analysis is R[11]. To use the capabilities of R inside the OpenCube Toolkit, we integrated it with our architecture through the **Statistical Analysis of RDF Data Cubes** component. R is run as a web service (using Rserve[12] package) and accessed via HTTP. Input data are retrieved using SPARQL queries and passed to R together with an R script. 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).

3 Conclusions

This demo paper presents the first release of the OpenCube Toolkit developed to enable easy publishing and reusing of linked data cubes. The toolkit smoothly integrates separate components dealing with different subtasks of the linked statistical data processing workflow to provide the user with a rich set of functionalities for working with statistical semantic data.

Acknowledgments. The work presented in this paper was partially carried out in OpenCube[13] project, which is funded by the EC within FP7 (No. 611667).

References

1. Cyganiak, R., Reynolds, D.: The RDF Data Cube vocabulary, http://www.w3.org/TR/vocab-data-cube/ (2013)
2. Kalampokis, E., Tambouris, E., Tarabanis, K.: Linked Open Government Data Analytics. In: Wimmer, M.A., Janssen, M., Scholl, H.J. (eds.) EGOV 2013. LNCS, vol. 8074, pp. 99-110. IFIP International Federation for Information Processing (2013)
3. Haase, P., Schmidt, M., Schwarte, A. Information Workbench as a Self-Service platform. COLD 2011, ISWC 2011, Shanghai, China (2011).

[11] http://www.r-project.org/
[12] http://www.rforge.net/Rserve/
[13] http://www.opencube-project.eu