1 Publishable Summary

1.1 Project context and objectives

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 and evidence-based policy-making.

Recently, Linked Data 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 cubes characterized by dimensions, slices and observations in order to unveil its full potential and value. Linked data cubes could open up new possibilities in performing data analytics at a Web scale (e.g. by integrating disparate datasets and extracting of interesting and previously hidden insights).

However, both Linked Data and data cubes introduce complexity that raises the barrier for opening up and exploiting statistical data. Here, the RDF Data Cube (QB)[1] vocabulary provides the fundamental background for modelling the data. As regards software components and tools, it was only recently that components and tools for publishing and reusing linked data cubes were developed. 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 and environments, and (d) the capability to be used in complex scenarios in an integrated manner.

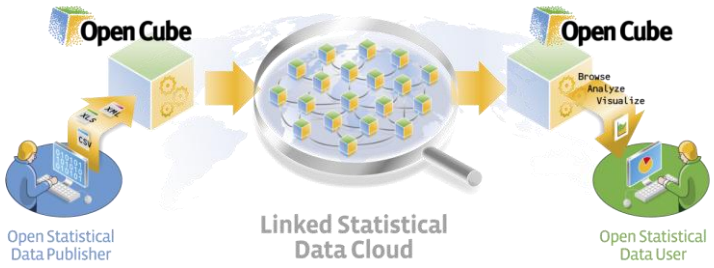

Figure 1 The OpenCube approach

OpenCube project aims at overcoming these limitations by developing software tools that:

- Provide advanced functionalities to users such as data cubes expansion, OLAP analysis, statistical analysis, and map visualizations.
- Are based on open-source licenses that allow for reuse.
- Are not dependent to specific software environments.
- Can be easily integrated together through two linked data management platforms i.e. PublishMyData and Information Workbench.

[1] http://www.w3.org/TR/vocab-data-cube/

1.2 Expected results and impact

OpenCube expected results are:

- The OpenCube Toolkit[2] comprising open-source tools.
- The OpenCube extension of the **fluidOps' Information Workbench** platform[3].
- The OpenCube extension of the **Swirrl's PublishMyData** platform[4].
- The OpenCube Pilots in three public authorities and businesses.

The outcome of OpenCube will offer a competitive advantage to the portfolio of the commercial project partners, as it will be the first comprehensive software solution for data analytics based on Linked Statistical Data. It will also enable public authorities to publish and exploit open statistical data of high quality using the OpenCube toolkit.

1.3 Current progress

During the first year of the project the requirements regarding creating and exploiting linked data cubes were identified and categorized by means of the OpenCube lifecycle. This lifecycle describes the steps that raw data cubes should go through in order to create value by means of Linked Data technologies. The steps are categorized into two phases (a) the publish phase that includes creating linked data cubes out of raw data, and (b) the reuse phase that includes utilizing linked data cubes in advanced analytics and visualizations.

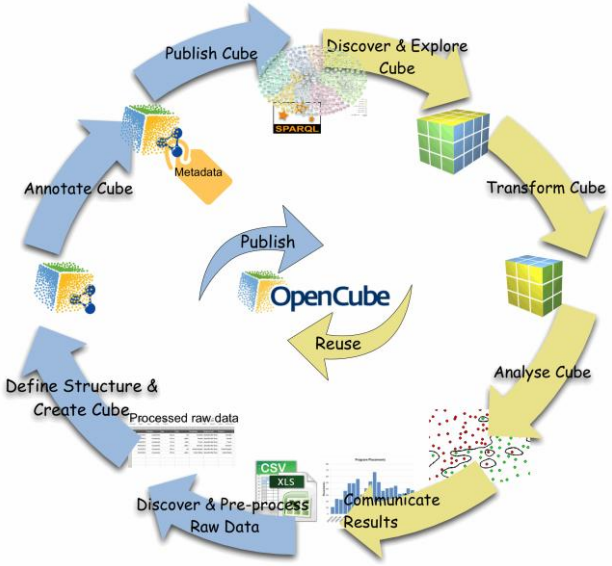

Figure 2 The OpenCube lifecycle

Based on the OpenCube lifecycle a documented definition of a common software architecture was created for the OpenCube project and its three software deliverables: the OpenCube extensions for Swirrl's PublishMyData product, the OpenCube extensions for fluidOps' Information Workbench product, and the stand-alone open source OpenCube Toolkit. For each step of the lifecycle five architecture layers were defined: (a) user interface, (b) data management, (c) infrastructure, (d) storage, and (e) model.

Different steps of the lifecycle are realized by separate components. These components are integrated together by means of a common platform constituting a toolkit providing a single work environment to the user. Two different implementation

[2] http://opencube-toolkit.eu/
[3] http://www.fluidops.com/en/portfolio/information_workbench/
[4] http://www.swirrl.com/publishmydata

approaches of this toolkit are considered based on the underlying platform. In particular, OpenCube components have been included in two platforms i.e. Information Workbench and PublishMyData.

The main outcome of the development work in the first year was the release of the first implementation prototypes of the software components.

The Information Workbench platform serves as a backbone for the open source OpenCube Toolkit. In the first release the **OpenCube Toolkit**[5] comprises the following components:

- Data Publishing
    - **TARQL extension for data cubes**: data conversion to RDF from legacy tabular data, such as CSV/TSV files.
    - **D2RQ extension for data cubes**: data conversion from relational databases into RDF data cubes.
    - **JSON-Stat to QB:** data conversion from JSON-Stat files to RDF data cubes.
- Data Reuse
    - **Data catalogue management**: user interface (UI) templates for managing metadata on RDF data cubes and supporting search and discovery.
    - **OpenCube Browser**: table-based visualizations of RDF data cubes.
    - **OpenCube Aggregation component**: it pre-computes aggregations to enable OLAP operations.
    - **R statistical analysis**: R integration module in the Information Workbench.
    - **Interactive chart visualization widgets**: visualization of the RDF data cube slices with charts.
    - **OpenCube MapView**: map-based visualizations for geographically located data.

Extensions to **PublishMyData** intended for the commercial exploitation include the following tools:

- **Grafter**[6] data transformation pipeline toolkit.
- Data cube grid view component
- Choropleth map comparison component

These prototypes were used to perform initial user studies with four use case partners: the Department for Communities and Local Government UK (DCLG), the Central Statistics Office Ireland (CSO), the Studiedienst Vlaamse Regering (SVR) in Belgium, and a Swiss Bank. The three first are committed trial partners, while the last one is an associated organization. Although these use case

[5] http://opencube-toolkit.eu
[6] http://grafter.org

partners are not formal members of the consortium, they are committed to participate in the activities of the project.

The pilot at the DCLG, which was supported by the consortium partner Swirrl, involved the use of Grafter for the conversion of legacy data relating to council tax and Energy Performance Certificates and the use of the reuse components (e.g. Table Viewer, Map Viewer, Sparql console) in their PublishMyData platform already in use at DCLG.

The pilot at CSO involved the conversion of large sets of existing data (Irish 2011 Census and Statbank) in a timely manner by NUIG with the publishing components TARQL and 'JSON-Stat to QB'.

ProXML used TARQL to convert legacy data on employment and nationalities from the SVR. SVR used the open-source OpenCube Toolkit (using OpenCube Browser, Interactive Chart Visualisation, OpenCube MapView, R Chart Widget) to develop wiki pages that enable end-users to visualize and analyse the converted data.

fluidOps carried out a scenario-driven workshop at a Swiss Bank, a scenario to find relevant statistical datasets, and once found to visualize and analyse them using the commercial version of Information Workbench.

These initial studies have shown promising results as well as identified priority directions for improvements and further developments of software components. In particular, the evaluation revealed that the tools for converting legacy data to the RDF Data Cube format (TARQL, Grafter) were up to the task, but need a layer above, so that less technically skilled people will be able to do the conversions themselves. Although the visualising and analysing components show great promise, stakeholders suggested that the user friendliness of the components should be improved.

1.4 Project website and other information

Project title:

OpenCube - Publishing and Enriching Linked Open Statistical Data for the Development of Data Analytics and Enhanced Visualisation Services

Project Coordinators:

Prof. Konstantinos Tarabanis, CERTH, e-mail: kat@uom.gr

Deputy Project Coordinator:

Ass. Prof. Efthimios Tambouris, CERTH, e-mail: tambouris@uom.gr.

Partners:

Centre for Research and Technology Hellas (GR), National University of Ireland, Galway (IE), Fluid Operations AG (DE), Swirrl IT Limited (UK), ProXML BVBA (BE)

Duration: November 2013 - October 2015

Further information: http://opencube-project.eu and http://opencube-toolkit.eu