Creating and Utilizing Linked Open Statistical Data for the Development of Advanced Analytics Services

E. Kalampokis, A. Karamanou, A. Nikolov, P. Haase, R. Cyganiak, B. Roberts, P. Hermans, E. Tambouris, K. Tarabanis
Objective

- A major part of Open Data concerns statistics that can be formulated as data cubes.
- The objective of this paper is to present the OpenCube approach for working with linked data cubes.
- The ultimate goal of OpenCube is to facilitate
  - Publishing of high-quality linked statistical data
  - Reusing linked statistical datasets in visualizations and analytics
OpenCube develops **components** to support the whole lifecycle of linked statistical data.

The lifecycle describes **steps** that raw data cubes should go through in order to create value.
Different steps of the lifecycle are realized by separate components.

Two different implementation approaches are considered based on the underlying platform.

- fluidOps’ Information Workbench
- Swirrl’s PublishMyData

Extensions for the commercial platforms and an Open-Source toolkit.
Components

- Publishing components
  - TARQL extension
  - D2RQ /R2RML-QB extension
  - JSON-stat
  - Grafter

- Consuming components
  - Data catalogue
  - OpenCube Browser
  - OpenCube MapView
  - R Analysis Chart
  - Aggregation component
TARQL OpenCube Extension

- TARQL is a command-line tool for converting CSV files to RDF using SPARQL 1.1 syntax
  - https://github.com/cygri/tarql
- TARQL is a SPARQL based data mapping language.
- The OpenCube TARQL extension enables RDF data cubes construction from CSV files.
  - Redesigned TARQL API
  - Added streaming evaluation mode
- It will be integrated to the IWB platform very soon.
The D2RQ OpenCube component enables the generation of **RDF data cubes** from relational tables.

It builds upon the **D2RQ** open source platform and it leverages **R2RML** language.

The component will be integrated into the **IWB** platform and it will provide an easy to use interface to adjust output mapping.
JSON-stat

- The JSON-stat format is a simple lightweight **JSON format for multidimensional data.**
  - [http://json-stat.org/format/](http://json-stat.org/format/)
- A JSON-stat file can contain one or more datasets.
- Multiple datasets responses allow a provider to disseminate information with few common dimensions in a single response.
Grafter

- Open source software framework for transforming tabular data (CSV or XLS) to RDF
  - [http://grafter.org](http://grafter.org)
- Automatable/works with API
- Designed to support a graphical user interface (work in progress)
- Performs well with large datasets
Managing metadata over data cubes

- Data collection
  - Integration of major open data catalogs
- UI for search and exploration of data sets
  - Rich metadata based on open standards
  - Both descriptive and structural metadata
- Self-service UI
  - Custom queries and visualizations
  - Widgets, dashboarding, etc.
Data catalogue management

- Managing catalogues of datasets
  - Search & discovery of relevant data
- **Goal**: on-demand provisioning
OpenCube browser

- It enables the **exploration** of an RDF data cube by presenting a **two-dimensional slice** of the cube as a **table**.
- The slice is created by setting a **fixed values for each dimension** that is not presented in the table.
- The browser is integrated in both **IWB** and **PublishMyData** platform.
OpenCube browser (IWB extension)

Summarize observations across a dimension (dimension reduction)

Change the axes of the table

Change the language

Change the fixed values
Data cube grid view (PublishMyData extension)

- See [http://opendatacommunities.org](http://opendatacommunities.org) for live examples

### Domestic Energy Performance Certificates Lodged on Register - By Energy Efficiency Rating (2014 Q2)

<table>
<thead>
<tr>
<th>Reference area</th>
<th>Not recorded</th>
<th>Rating A</th>
<th>Rating B</th>
<th>Rating C</th>
<th>Rating D</th>
<th>Rating E</th>
<th>Rating F</th>
</tr>
</thead>
<tbody>
<tr>
<td>E06000001 Hartlepool</td>
<td>0</td>
<td>12</td>
<td>172</td>
<td>173</td>
<td>74</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>E06000002 Middlesbrough</td>
<td>0</td>
<td>17</td>
<td>57</td>
<td>271</td>
<td>177</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>E06000003 Redcar and Cleveland</td>
<td>0</td>
<td>17</td>
<td>59</td>
<td>221</td>
<td>76</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>E06000004 Stockton-on-Tees</td>
<td>2</td>
<td>47</td>
<td>103</td>
<td>266</td>
<td>135</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>E06000005 Darlington</td>
<td>0</td>
<td>21</td>
<td>96</td>
<td>153</td>
<td>81</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>E06000006 Halton</td>
<td>0</td>
<td>5</td>
<td>93</td>
<td>145</td>
<td>46</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>E06000007 Warrington</td>
<td>0</td>
<td>15</td>
<td>111</td>
<td>212</td>
<td>60</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>E06000008 Blackburn with Darwen</td>
<td>0</td>
<td>12</td>
<td>104</td>
<td>196</td>
<td>97</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>E06000009 Blackpool</td>
<td>0</td>
<td>3</td>
<td>68</td>
<td>272</td>
<td>167</td>
<td>57</td>
<td></td>
</tr>
</tbody>
</table>

Download results as CSV

---

**Row and Column Headings**

- Column Headings: By Energy Efficiency Rating

- Rows (values in first column): Reference area

---

**Other Options**

- Reference period: 2014 Q2
Data cube grid view

- Shows two dimensional slice of data
- Controls to set values of other dimensions
- Download chosen slice as CSV
- Performs well with large datasets by loading data asynchronously as users scrolls through
- See http://opendatacommunities.org for live examples
OpenCube MapView

- It enables the visualization of RDF data cubes on a map based on their geospatial dimension.
- It supports:
  - Markers
  - Bubble
  - Choropleth maps (need for polygons)
- It is integrated in both
  - IWB and
  - PublishMyData
Choropleth map in PublishMyData

This data set contains unrounded figures, rounded figures are available in Table 253, available for download as an Excel spreadsheet.

Mapper

Spreadsheet view

This dataset contains multidimensional data (a data cube) which can be displayed as a grid to compare two dimensions at a time. Use the drop-down menus below the grid to choose which dimensions to show as rows and columns (and, optionally, to filter the other dimensions by value).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>E06000001 Hartlepool</td>
<td>230</td>
<td>150</td>
<td>190</td>
<td>170</td>
<td>170</td>
</tr>
<tr>
<td>E06000002 Middlesbrough</td>
<td>320</td>
<td></td>
<td></td>
<td>220</td>
<td></td>
</tr>
<tr>
<td>E06000003 Redcar and Cleveland</td>
<td>210</td>
<td>250</td>
<td>260</td>
<td>230</td>
<td>270</td>
</tr>
<tr>
<td>E06000004 Teesside</td>
<td>420</td>
<td>400</td>
<td>420</td>
<td>430</td>
<td>440</td>
</tr>
</tbody>
</table>
Support for advanced analytic tasks

- Reuse of existing established tools to support advanced analytic tasks
- Loose coupling integration with R
  - R is accessed as a web service
- Rich analysis capabilities (all packages developed by the R community)
Integration with R

- Visualisation of analysis results (charts & tables)
- Reuse of analysis results: preserving R output as linked data
- Managing a catalogue of the analytics experiments („recipes“)
Data Cube Visualization

Visualization and Exploration

- **Widget-based visualization of data**
  - **Pre-existing**: Configuration using explicit SPARQL queries
    - More appropriate for engineers building custom solutions than for end users
  - **Goal**: Intuitive configuration of visualization views exploiting the Data Cube structure

Analytics and Reporting
Stock chart visualization

- Adaptation of the stock chart view to the RDF data cube datasets
- Improved configuration UI
  - specifying dimension restrictions instead of the complete SPARQL query
- Additional features (e.g., comparison between slices)
Initial Evaluation Results

- We currently perform evaluations of the components in four pilots
  - Department for Communities and Local Government (UK)
  - Central Statistics Office (Ireland)
  - Flemish Government (Belgium)
  - Swiss Banks

- Some interesting findings
  - Why to use linked data
  - Performance issues with large data sets
  - Noisy data
OpenCube toolkit

- For more information
  - [http://opencube-project.eu](http://opencube-project.eu)
  - [http://opencube-toolkit.eu](http://opencube-toolkit.eu)