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

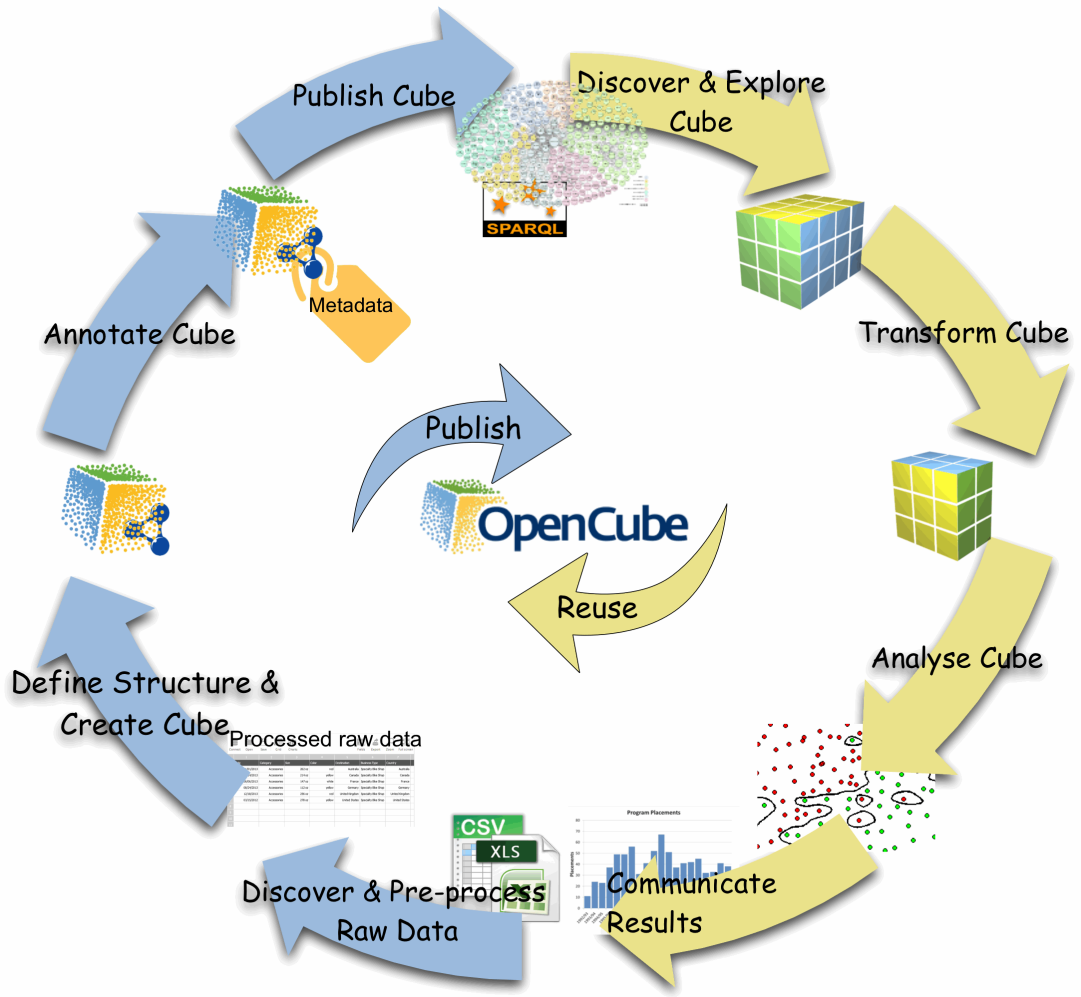
E. Kalampokis, A. Karamanou, A. Nikolov, P. Haase, R. Cyganiak, B. Roberts, P. Hermans, E. Tambouris, K. Tarabanis

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

Linked Statistical Data Lifecycle

- 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.



Implementation

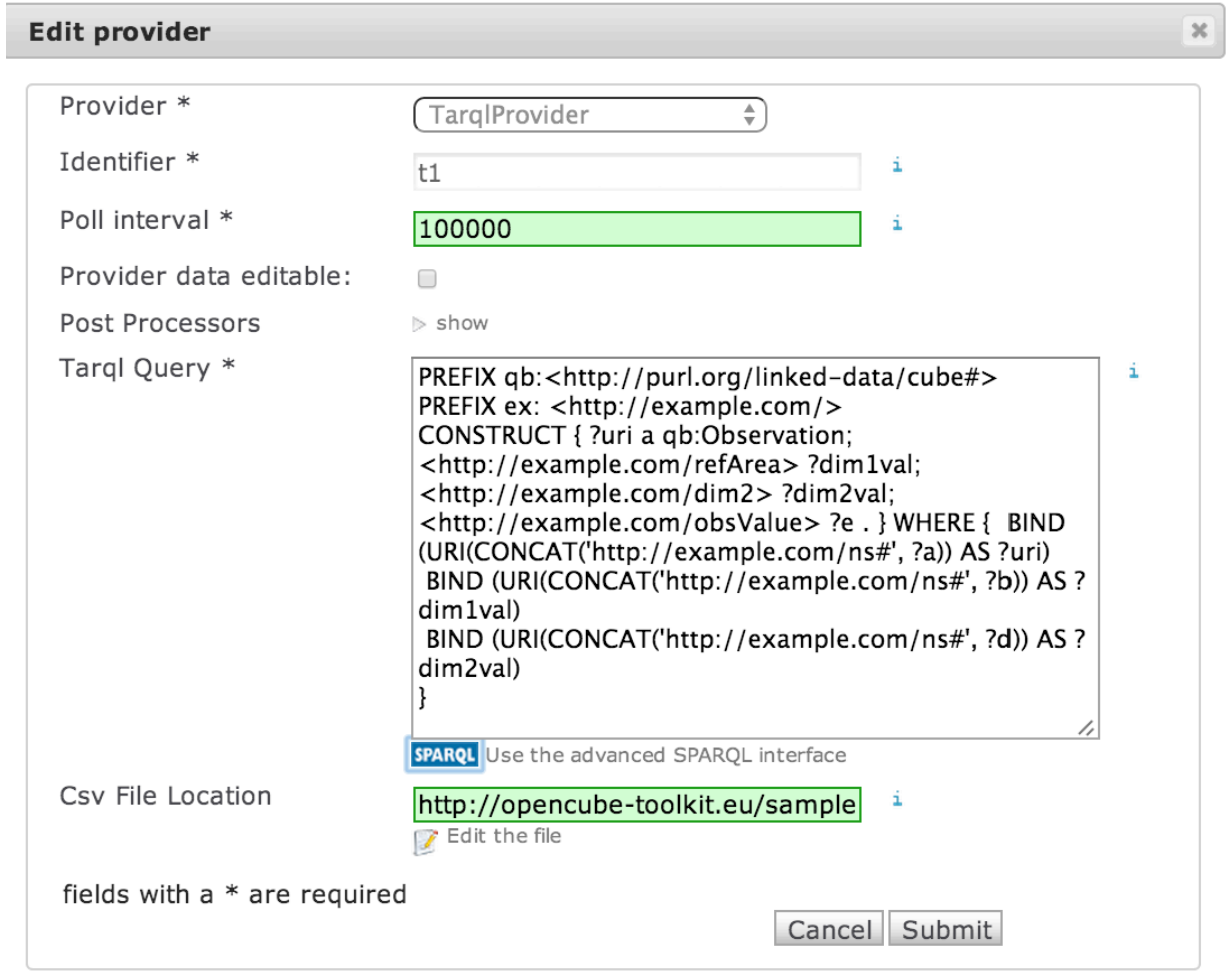
- Different steps of the lifecycle are realized by separate components.
- Two different implementation approaches are considered based on the underlying platform.
  - fluidOps' Information Workbench
  - Swirrl's PublishMyData
- Extensions for the commercial platforms and an Open-Source toolkit.

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

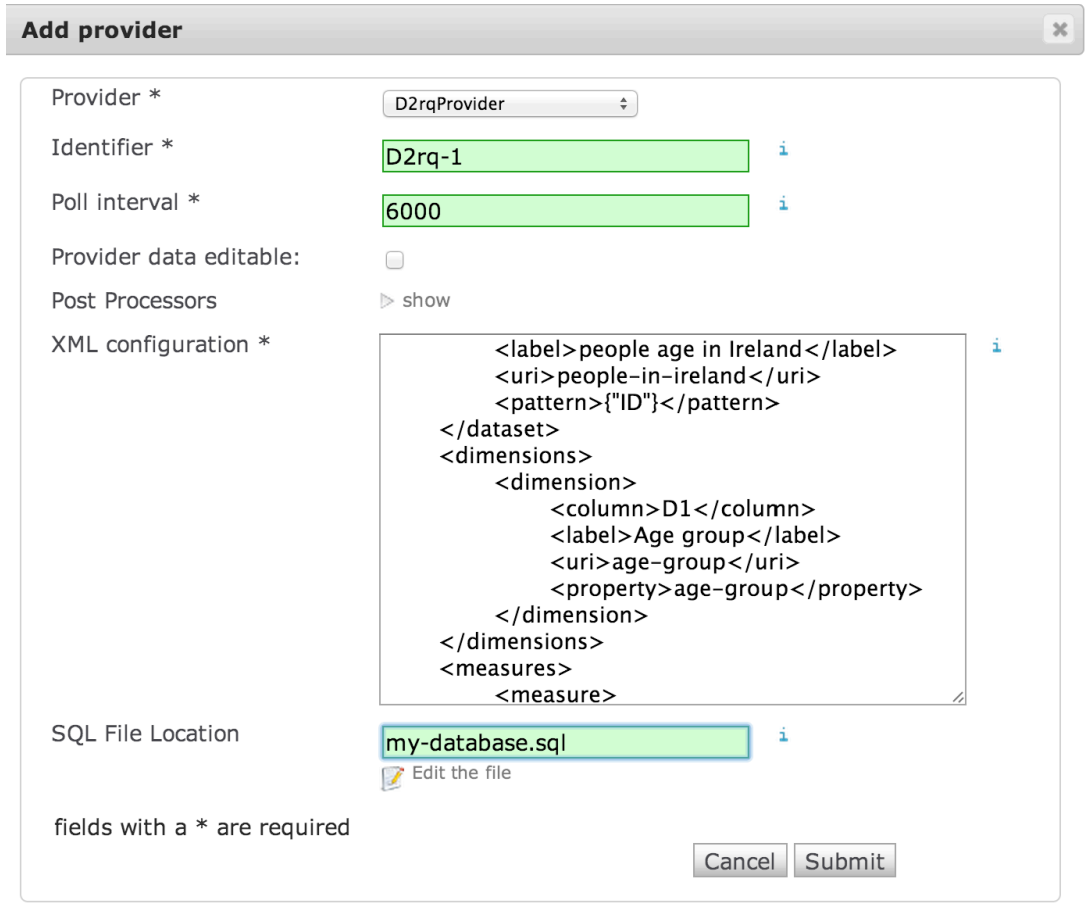
TARQL OpenCube Extension

- 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.



D2RQ/R2RML-QB Extension

- 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/
- 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
- Automatable/works with API
- Designed to support a graphical user interface (work in progress)
- Performs well with large datasets

Linked Data Manufacturing

Industrial-strength RDF production

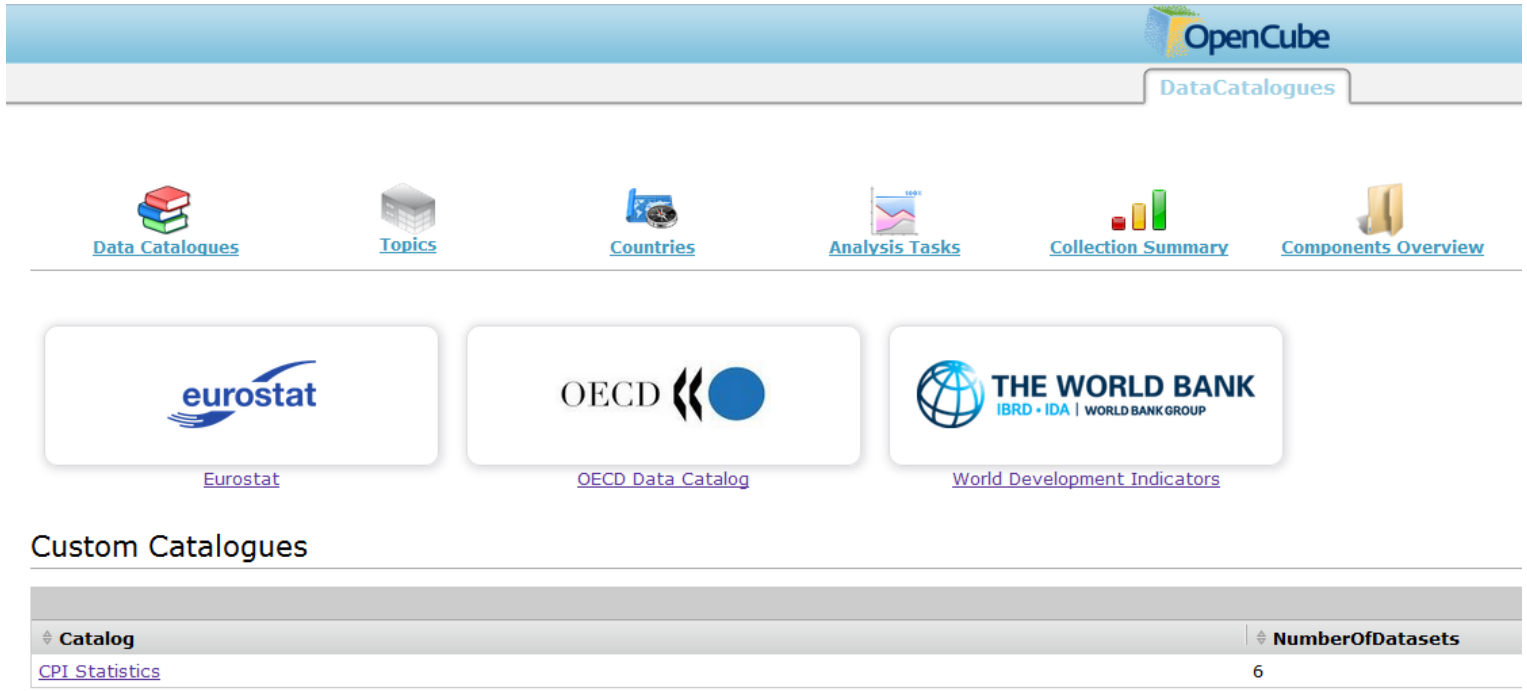
OpenCube

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



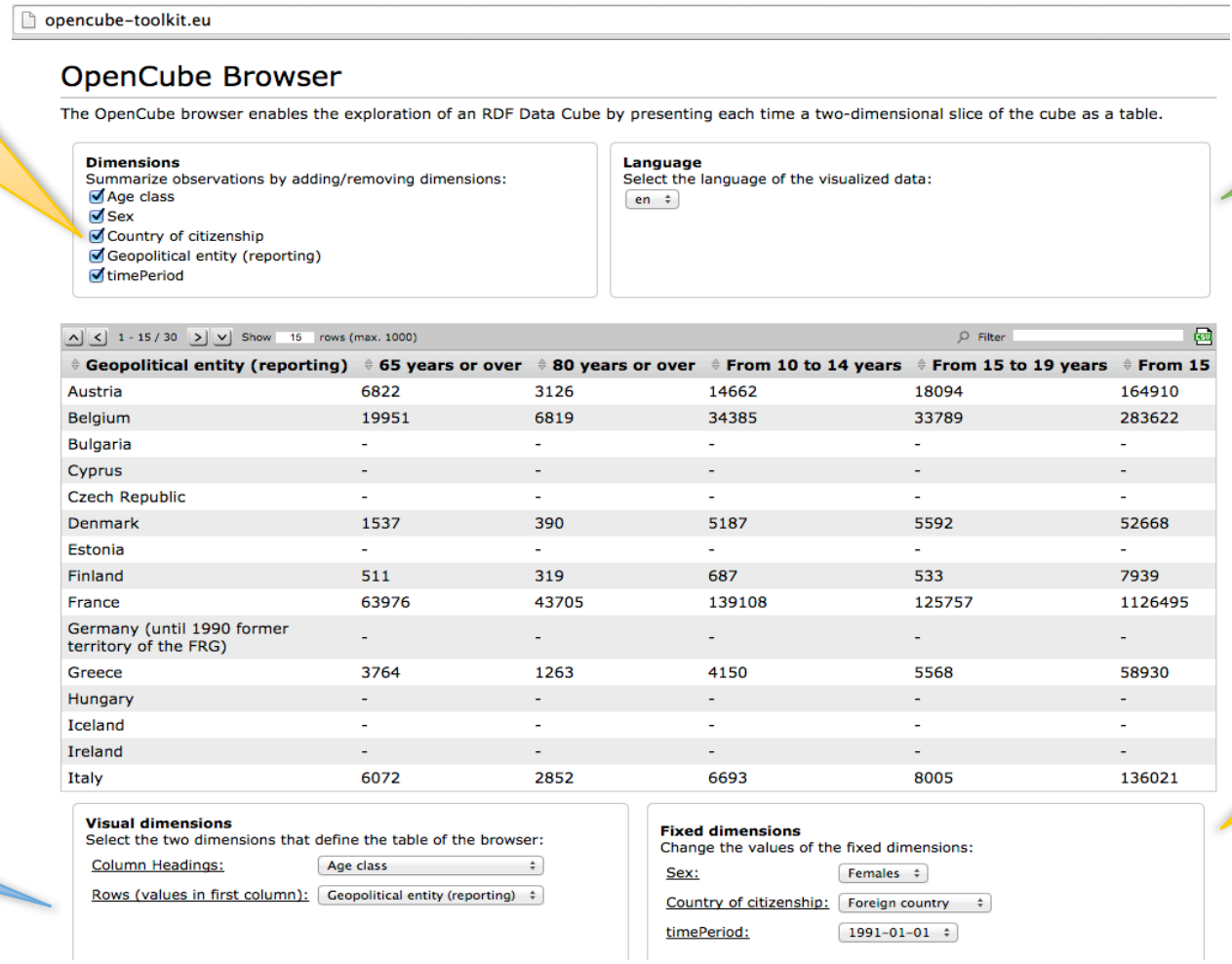
OpenCube browser

- 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 language

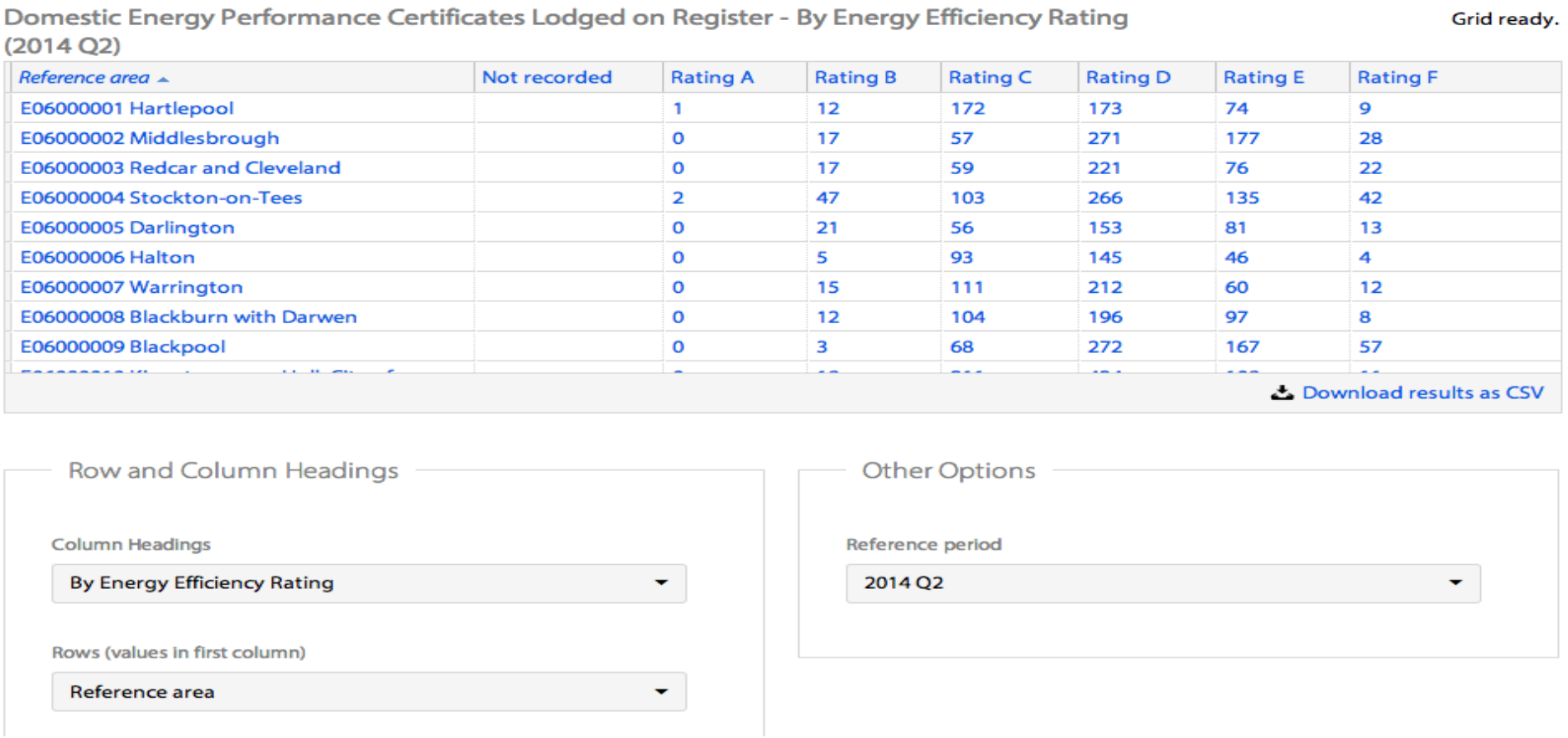


Change the axes of the table

Change the fixed values

Data cube grid view (PublishMyData extension)

- See http://opendatacommunities.org for live examples



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

Grid ready.

Reference area	Not recorded	Rating A	Rating B	Rating C	Rating D	Rating E	Rating F
E06000001 Hartlepool		1	12	172	173	74	9
E06000002 Middlesbrough		0	17	57	271	177	28
E06000003 Redcar and Cleveland		0	17	59	221	76	22
E06000004 Stockton-on-Tees		2	47	103	266	135	42
E06000005 Darlington		0	21	56	153	81	13
E06000006 Halton		0	5	93	145	46	4
E06000007 Warrington		0	15	111	212	60	12
E06000008 Blackburn with Darwen		0	12	104	196	97	8
E06000009 Blackpool		0	3	68	272	167	57

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

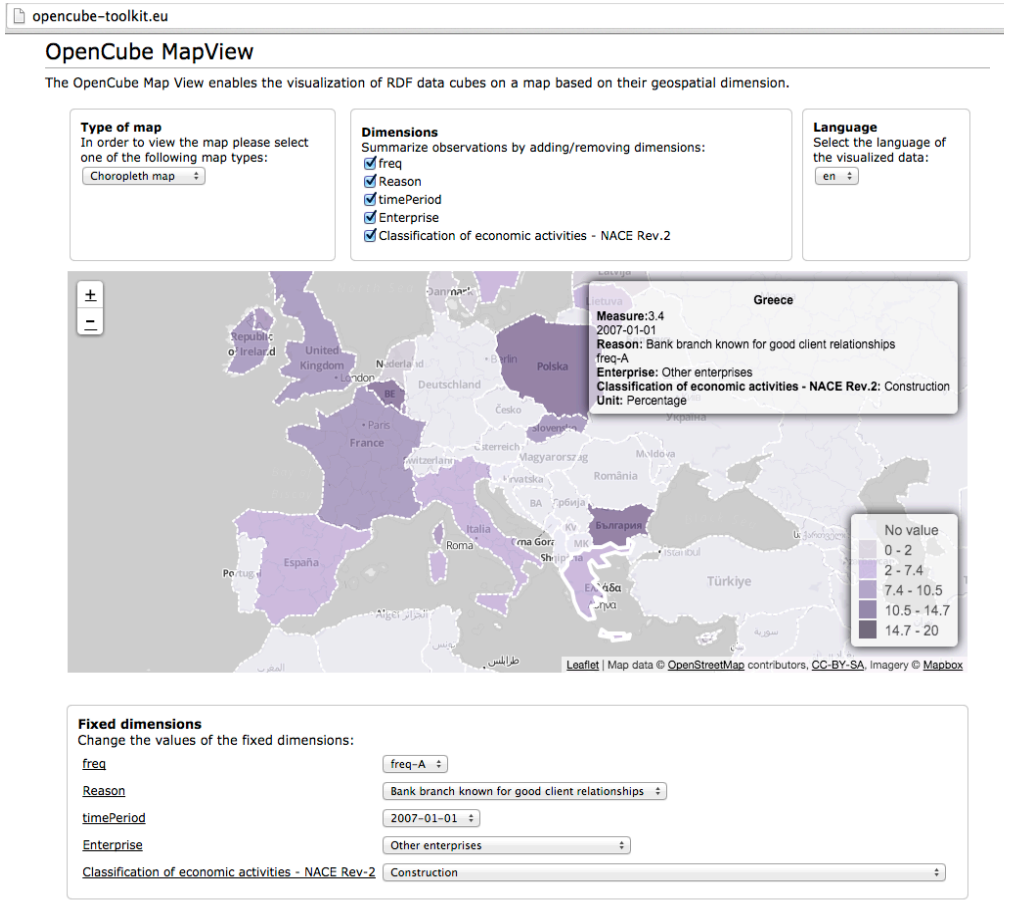
2014 Q2

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

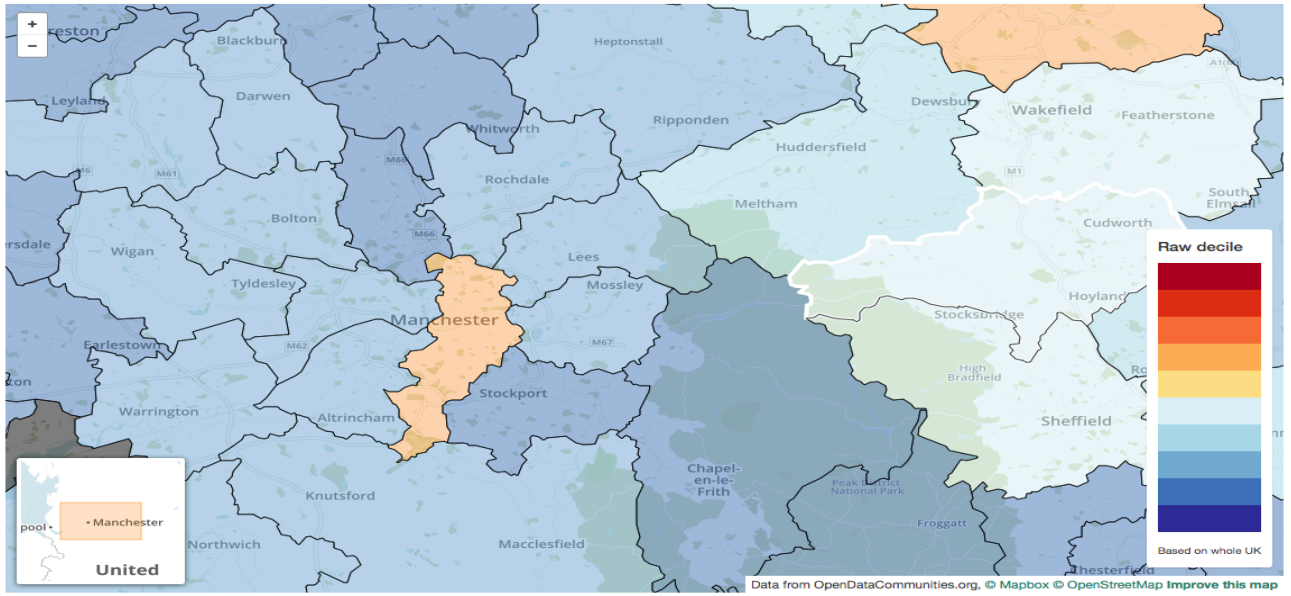
- It enables the visualization of RDF data cubes on a map based on their geospatial dimension.
- It supports:
  - Markers
  - Bubble
  - Choropleth maps (need for polygons)
- It is integrated in both
  - IWB and
  - PublishMyData



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

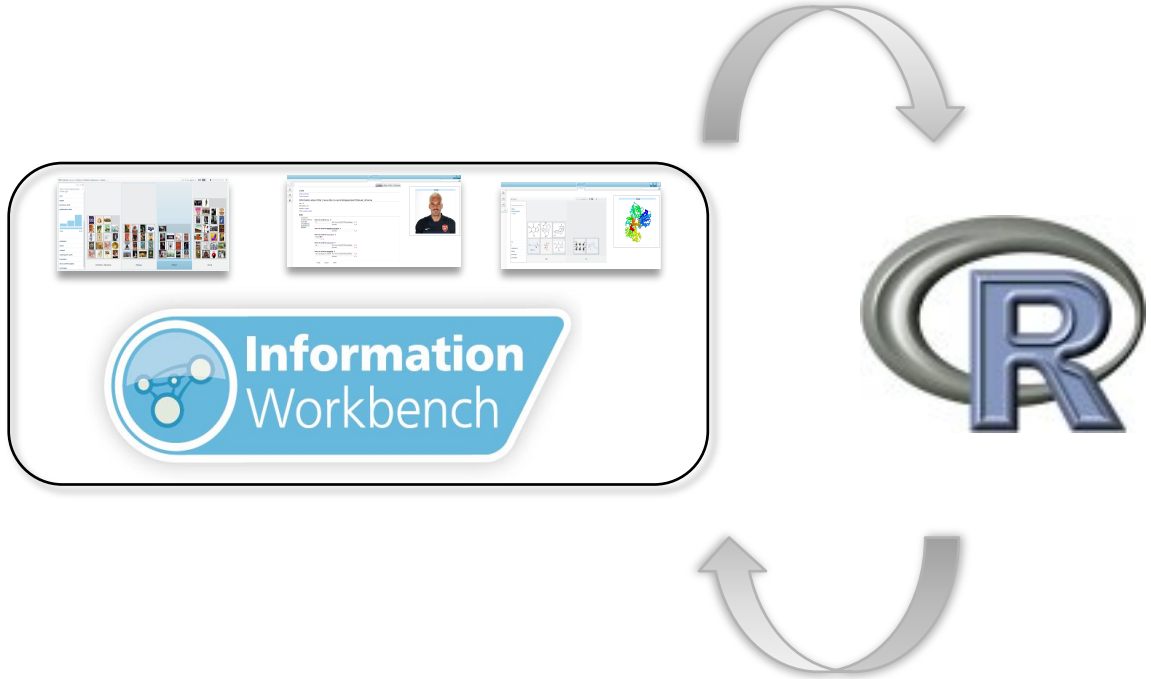
This dataset contains multidimensional data (a *data cube*) which can be displayed as a grid to compare two dimensions at a time.

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).

Permanent dwellings completed, 2009/10 to 2013/14, England, District By Tenure (All)

	Reference area	2009-2010	2010-2011	2011-2012	2012-2013	2013-2014
1	E06000001 Hartlepool	230	150	190	170	170
2	E06000002 Middlesbrough	320			220	
3	E06000003 Redcar and Cleveland	210	250	260	230	270
4	E06000004 Stockton-on-Tees	630	580	500	640	520

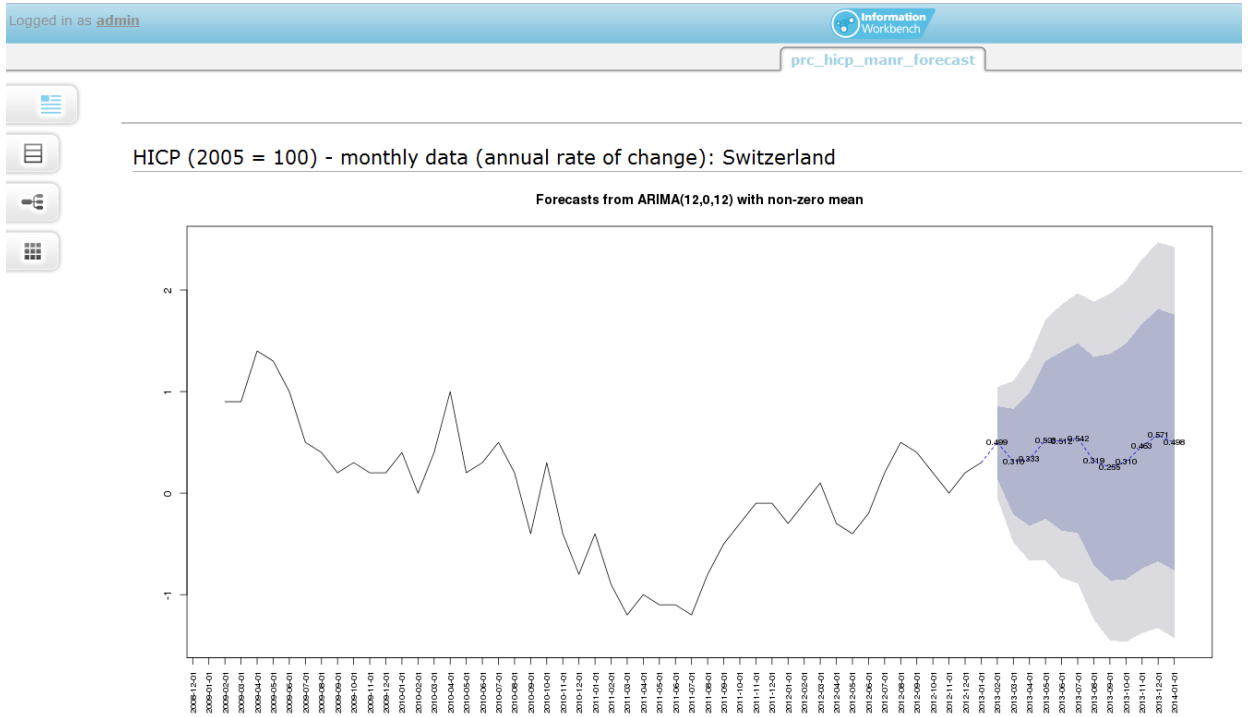
Support for advanced analytic tasks



- 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)

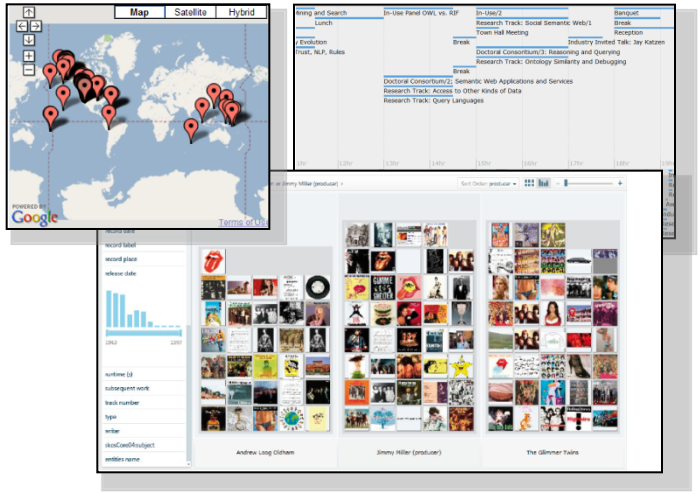
Integration with R

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

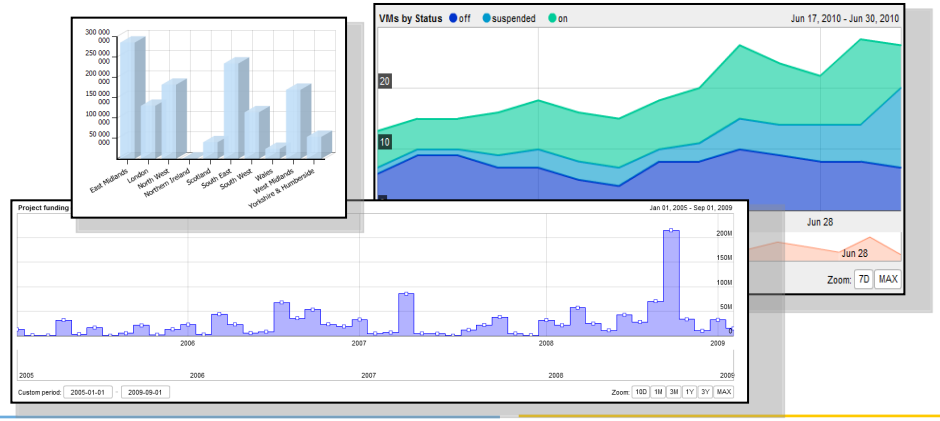


Data Cube Visualization

Visualization and Exploration



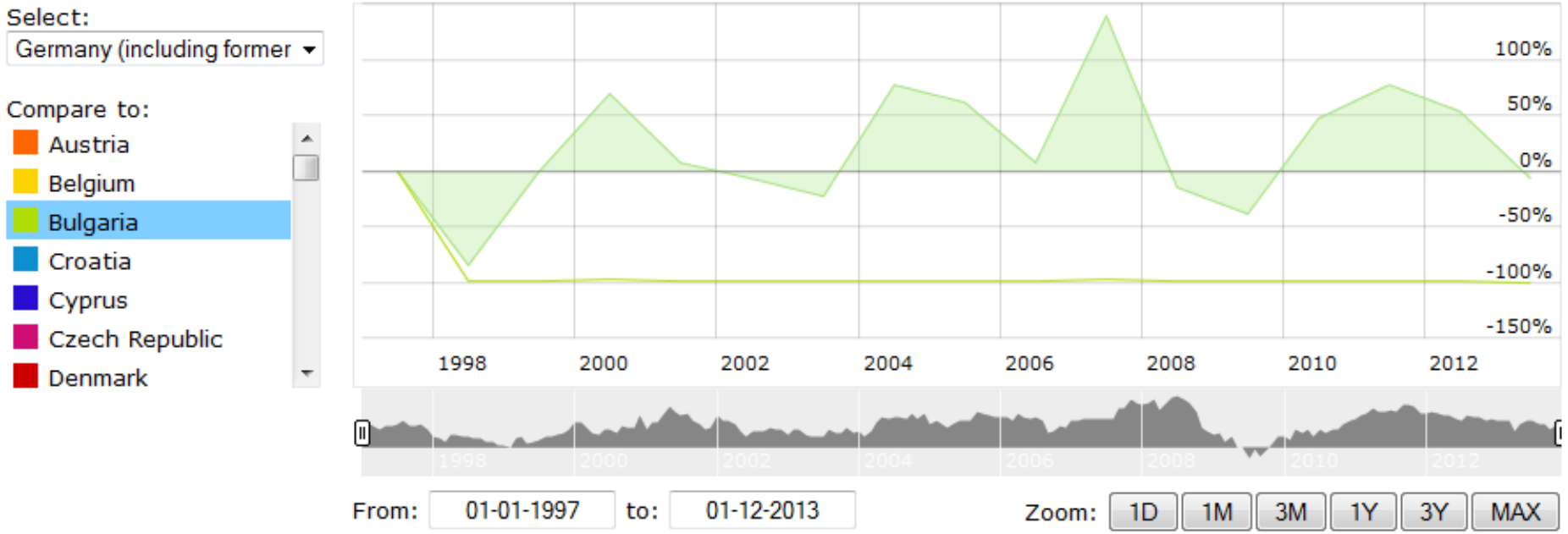
Analytics and Reporting



- 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

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-toolkit.eu