CubeViz – Exploration and Visualization of Statistical Linked Data

Michael Martin[*]
Universität Leipzig
Augustusplatz 10
04109 Leipzig, Germany
martin@informatik.uni-leipzig.de

Konrad Abicht
Universität Leipzig
Augustusplatz 10
04109 Leipzig, Germany
k.abicht@googlemail.com

Claus Stadler
Universität Leipzig
Augustusplatz 10
04109 Leipzig, Germany
cstadler@informatik.uni-leipzig.de

Sören Auer
Universität Bonn & Fraunhofer IAIS
Römerstraße 164
53117 Bonn, Germany
auer@cs.uni-bonn.de

Axel-C. Ngonga Ngomo
Universität Leipzig
Augustusplatz 10
04109 Leipzig, Germany
ngonga@informatik.uni-leipzig.de

Tommaso Soru
Universität Leipzig
Augustusplatz 10
04109 Leipzig, Germany
tsoru@informatik.uni-leipzig.de

ABSTRACT

CubeViz is a flexible exploration and visualization platform for statistical data represented adhering to the RDF Data Cube vocabulary. If statistical data is provided adhering to the Data Cube vocabulary, CubeViz exhibits a faceted browsing widget allowing to interactively filter observations to be visualized in charts. Based on the selected structural part, CubeViz offers suitable chart types and options for configuring the visualization by users. In this demo we present the CubeViz visualization architecture and components, sketch its underlying API and the libraries used to generate the desired output. By employing advanced introspection, analysis and visualization bootstrapping techniques CubeViz hides the schema complexity of the encoded data in order to support a user-friendly exploration experience.

Keywords

Statistics, exploration, linked data, visualization

1. INTRODUCTION

A vast part of the existing Linked Data Web consists of statistics (cf. LODStats[1] [5]) being represented according to the RDF Data Cube Vocabulary [4]. To hide the inherently complex, multidimensional statistical data structures and to offer user-friendly exploration the RDF Data Cube Explorer CubeViz[2] was developed. In this demo we will showcase how large data cubes comprising statistical data from different domains can be analysed, explored and visualized. CubeViz is based on the OntoWiki Framework [7] and consists of the following OntoWiki extensions:

- The *Integrity Analysis Component* (cf. section 3) evaluates the existence and the quality of selected RDF graphs according given integrity constraints.
- The *Facetted Data Selection Component* (cf. section 4) is retrieving the structure of the selected Data Cube using SPARQL [6] in order to generate filter forms. Those forms allow to slice the data cube according user interests.
- The *Chart Visualization Component* (cf. section 5) receives all observation as input, that correspond to the given filter conditions, in order to generate a chart visualization.

All components support the comprehensive CubeViz GUI shown in Figure 1. Before we introduce the three components in more detail, we give a brief introduction of the RDF Data Cube Vocabulary in the next section. We conclude the paper with links to publicly available deployments and a list of some upcoming features planned for the next release. Further information about CubeViz can be obtained in the repository wiki[3] or via a recorded webinar[4] comprising a comprehensive screencast.

2. THE RDF DATACUBE VOCABULARY

The *RDF Data Cube vocabulary* is a W3C recommendation for representing statistical data in RDF. The vocabulary is compatible with the Statistical Data and Medadata eXchange XML format (SDMX) [2], which is defined by an

[*] Corresponding Author

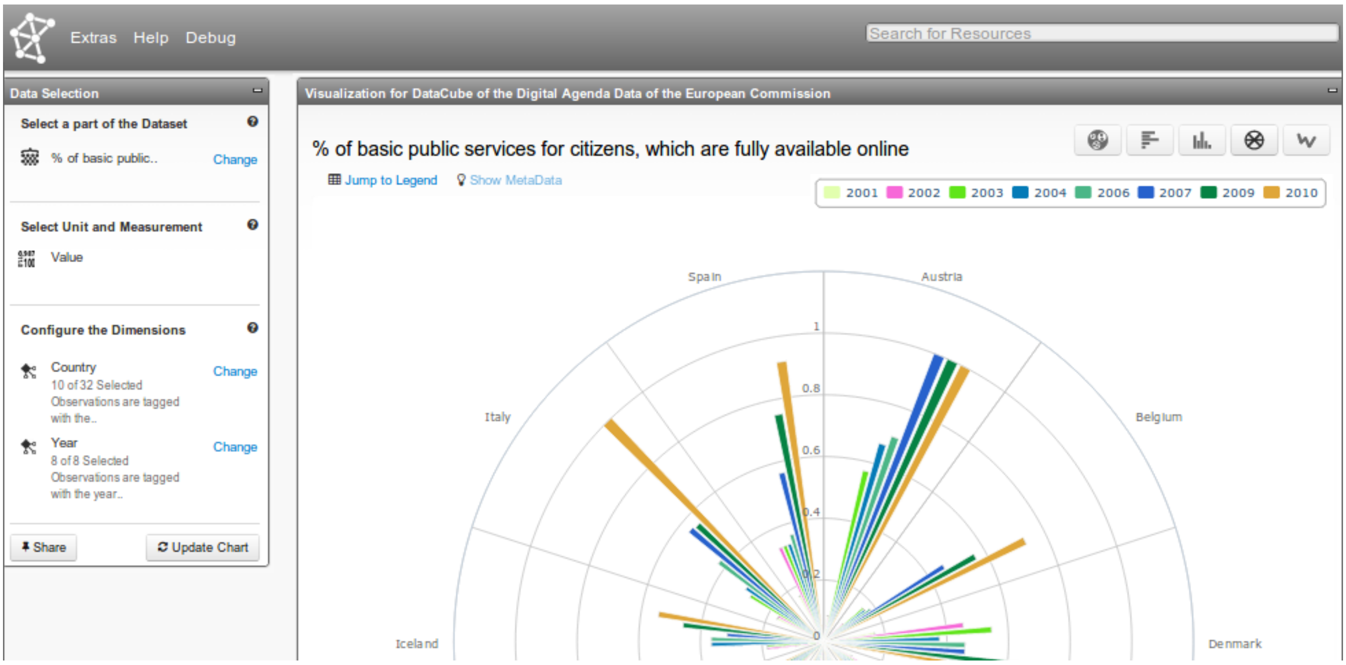
[1] `http://stats.lod2.eu/rdf_classes?search=Observation`

[2] `http://aksw.org/Projects/CubeViz`

[3] `https://github.com/AKSW/cubeviz.ontowiki/wiki`

[4] `http://www.youtube.com/watch?v=ZQc5lk1ug3M#t=1510`

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742848.



Data Selection

Select a part of the Dataset

% of basic public..

Select Unit and Measurement

Value

Configure the Dimensions

Country
10 of 32 Selected
Observations are tagged with the..

Year
8 of 8 Selected
Observations are tagged with the year..

Visualization for DataCube of the Digital Agenda Data of the European Commission

% of basic public services for citizens, which are fully available online

Figure 1: The CubeViz GUI visualizing a slice of a 2-dimensional RDF DataCube in a combined polar-column chart.

initiative chartered in 2001 to support the exchange of statistical data. Sponsoring institutions[5] of SDMX are among others *the Bank for International Settlements*, the *European Central Bank*, *Eurostat*, the *International Monetary Fund*, the *Organisation for Economic Co-operation and Development* (OECD), the *United Nations Statistics Division* and the *World Bank*. Experiences while publishing statistical data on the Web using SDMX were summarized by the *United Nations* in [1] and by the OECD in [3]

The core concept of the Data Cube vocabulary is the class `qb:Observation`[6], that is used to type all statistical observations being part of a Data Cube. Every observation has to follow a specific structure that is defined using the class `qb:DataStructureDefinition` (DSD) and referenced by a dataset resource (DS) of type `qb:DataSet`. Since every observation should refer to one specific DS (which again refers to the corresponding DSD) the structure of the observation is fully specified. DSD components are defined as set of dimensions (`qb:DimensionProperty`), attributes (`qb:AttributeProperty`) and measures (`qb:MeasureProperty`) to encode the semantics of observations. Those component properties are used to link the corresponding elements of dimensions, measure values and units with the respective observation resource. Furthermore, it is possible to materialize groups and slices of observations as well as hierarchical orders of dimension elements using respective concepts.

3. INTEGRITY ANALYSIS

As described in the W3C RDF Data Cube Recommendation document data cubes are structurally well-formed if they comply to specific integrity constraints[7]. Those constraints can be used to validate and if necessary to improve the quality of a data cube. For CubeViz, we translated those constraints into SPARQL queries using an `ASK`-clause returning boolean values. The queries were integrated into the Integrity Analysis Component of CubeViz, whose GUI is depicted in Figure 2. If a query returns false, the corresponding constraint is marked in the GUI in red and can be selected in order to reuse and modify them with a configured query editor. This functionality supports the discovery of potential modelling or conversion flaws.

Additionally, this component is used to introspect the selected RDF model for all included data cubes. If the introspection query (given in Listing 1) returns a positive result, the Faceted Data Selection and Chart Visualization components are activated.

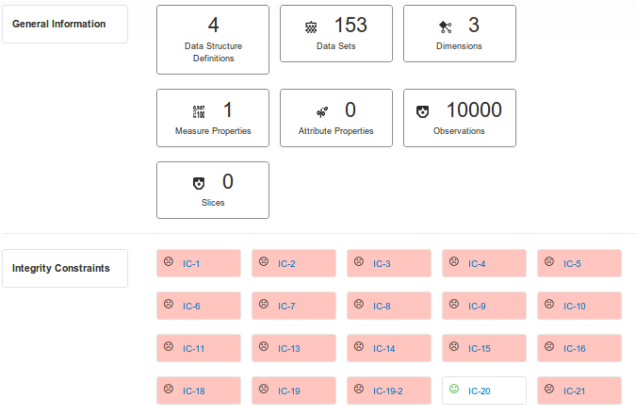
```
PREFIX qb:<http://purl.org/linked-data/cube#>
ASK FROM <http://example.org/> {
  ?observation   a              qb:Observation .
  ?observation   qb:dataSet     ?dataset .
  ?observation   ?dimension     ?dimelement .
  ?observation   ?measure       ?value .
  ?dataset       a              qb:DataSet .
  ?dataset       qb:structure   ?dsd .
  ?dsd           a       qb:DataStructureDefinition .
  ?dsd           qb:component   ?dimspec .
  ?dsd           qb:component   ?measurespec .
  ?dimspec       a        qb:ComponentSpecification .
  ?dimspec       qb:dimension   ?dimension .
  ?measurespec   a        qb:ComponentSpecification .
  ?measurespec   qb:measure     ?measure }
```

Listing 1: Data cube introspection query.

4. FACETED EXPLORATION

Given that the introspection was successful, specific structural parts of the identified data cube are queried in order to create a faceted search interface. All components of a DSD have to be integrated into any observation of the respective

[5] `http://sdmx.org/?page_id=6`
[6] Prefix qb:`http://purl.org/linked-data/cube#`
[7] `http://www.w3.org/TR/vocab-data-cube/#wf-rules`



General Information

4 Data Structure Definitions

153 Data Sets

3 Dimensions

1 Measure Properties

0 Attribute Properties

10000 Observations

0 Slices

Integrity Constraints

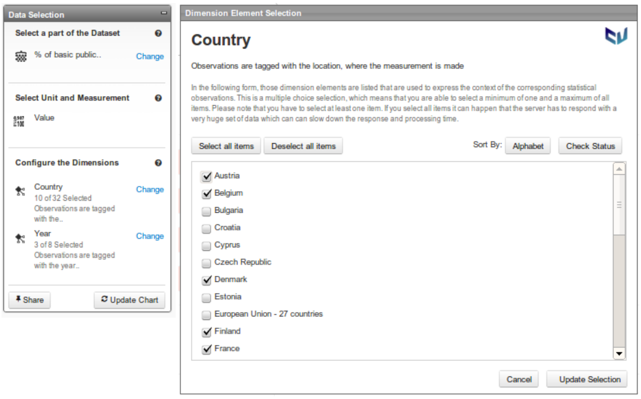
☹ IC-1 | ☹ IC-2 | ☹ IC-3 | ☹ IC-4 | ☹ IC-5

☹ IC-6 | ☹ IC-7 | ☹ IC-8 | ☹ IC-9 | ☹ IC-10

☹ IC-11 | ☹ IC-13 | ☹ IC-14 | ☹ IC-15 | ☹ IC-16

☹ IC-18 | ☹ IC-19 | ☹ IC-19-2 | ☺ IC-20 | ☹ IC-21

Figure 2: GUI presenting results of the statistical and integrity analysis.



Data Selection

Select a part of the Dataset

% of basic public.. Change

Select Unit and Measurement

Value

Configure the Dimensions

Country Change
10 of 32 Selected
Observations are tagged with the..

Year Change
3 of 8 Selected
Observations are tagged with the year..

Dimension Element Selection

Country

Observations are tagged with the location, where the measurement is made

In the following form, those dimension elements are listed that are used to express the context of the corresponding statistical observations. This is a multiple choice selection, which means that you are able to select a minimum of one and a maximum of all items. Please note that you have to select at least one item. If you select all items it can happen that the server has to respond with a very huge set of data which can can slow down the response and processing time.

☑ Austria
☑ Belgium
☐ Bulgaria
☐ Croatia
☐ Cyprus
☐ Czech Republic
☑ Denmark
☐ Estonia
☐ European Union - 27 countries
☑ Finland
☑ France

Figure 3: Facets and dialogues.

DS. In order to discover those observations the user has to select values that are referenced by those components. First the user needs to select a DS of a data cube in order to analyse the DSD that is the basis for all further facets. Second the user has to select the measure and attribute property used to identify the representation of values. The last mandatory facet is used to offer the selection of dimensions and its respective elements of interest. CubeViz is processing and visualizing values exactly as they are represented in the data cube and does not support aggregate functions such as `SUM`, `AVG`, `MIN` and `MAX`. As a consequence, users have to select at least one element of each dimension. Furthermore, if materialized slices are aggregated within the selected DS an optional facet will be generated to offer a selection from the retrieved slices.

Generation of Dialogues.

The detected facets and their generated GUI representations are integrated into a filter form. To select/deselect elements of facets for extracting subsets of the DS, respective interface elements are dynamically created. According the type of the facet (mandatory/optional) a configurable amount of elements (min/max) is selectable. Additionally, the label and textual description of components are retrieved using SPARQL queries and added to the interface. As illustrated in Figure 3 the selected amount of facet elements is displayed after confirmation. Already discovered RDF resources are cached on the client-side and will be re-used in the Chart Visualization component.

One of the major advantages of faceted exploration is the avoidance of possibly empty result sets. To avoid empty sets of observations after facet selection, the set of selectable elements of all further facets in combination with its respective count of observations is being calculated using respective SPARQL queries. Every selected combination of a component and its respective element is represented by a triple pattern that is conditionally used to retrieve the set of observations and all facet elements.

Initial Pre-selection.

To lower the barrier of exploring a data cube from scratch, an initial pre-selection algorithm is started after a positive introspection. As described in section 5 it is possible to integrate and configure charts visualizing one or multiple dimensions. The determined maximum amount of dimensions respectively chart axis is used as input for the pre-selection algorithm. After extracting all obligatory facets exactly one element per facet is pre-selected. According to the number of discovered dimensions and the maximum amount of processable chart axis, dimensions are randomly selected for which more than one element can be selected. To avoid confusing visualizations the amount of pre-selected elements is limited to 10 respectively 30% of the set of elements. During manual selection these limits are not relevant.

5. CHART VISUALISATION

In order to extract observations according to user interests, filter criteria from the facet selection component are translated into a corresponding SPARQL query. The resulting set of observation resources is serialized in JSON and sent back to the client. On the client-side the result set will be analysed according the amount of disjunctive dimensions and the respective amount of elements in order to select suitable charts. After identifying suitable chart implementations the first one is launched and renders the visualization using the received set of observations. All further suitable charts can be selected using a respective GUI element without querying the observations again.

Chart APIs.

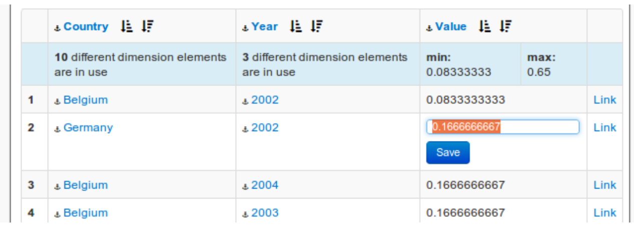
CubeViz comprises an abstraction layer to mediate between the retrieval of observations and the APIs used to generate and visualize charts. Currently, charts such as pie, bar, column, line and polar chart are implemented using the APIs *Data Driven Documents*[8] (D3js) and *HighCharts*[9]. Every chart is implemented using a defined interface and comprises a mechanism to convert the set of observations in combination with the meta data about dimension properties into the chart-specific input format.

Chart options.

Most of the implemented chart visualizations can be adjusted using preconfigured chart options. Hence, it is possible to enable the display of measure values in addition to its graphical representation, to switch axis / dimensions, to switch the value scale between linear and logarithmic or to

[8] `http://d3js.org/`

[9] `http://www.highcharts.com/`



	Country	Year	Value		
	10 different dimension elements are in use	3 different dimension elements are in use	min: 0.08333333	max: 0.65	
1	Belgium	2002	0.0833333333		Link
2	Germany	2002	0.1666666667 Save		Link
3	Belgium	2004	0.1666666667		Link
4	Belgium	2003	0.1666666667		Link

Figure 4: Interactive CubeViz legend.

enable a normal or percentage stacking. Additionally it is possible to combine charts such as a polar chart with a column chart (see Figure 1).

Element recognition.

On the Linked Data Web, URIs are used to identify resources. In a domain-agnostic tool such as CubeViz, it is not feasible to integrate static mappings between data items and their graphical representations. Most of the chart APIs have a limited amount of pre-defined colors used for colouring dimension elements or select colors completely arbitrarily. In order to display dimension elements in a deterministic manner and to support users to quickly recover selected elements in different charts we integrated a colouring algorithm that uniquely assigns URIs of each dimension element corresponding RGB color codes[10].

Interactive legend.

Below the generated charts an additional tabular representations of the selected data items is given (cf. Figure 4). On the one hand they can be used as legend containing additional meta data. On the other hand this view offers support for resolving data inaccuracies with functionality for editing values, that automatically updates the chart representation.

Sharing views.

After exploring, configuring and possible adaption of values users are able to share the created output. Sharing functionality is implemented via a button, which triggers the gathering of all information necessary to reproduce the created output, storing them server-side and returning a shareable link containing an identifying hash code for the particular view configuration. Furthermore, it is possible to export selected data as CSV and RDF in Turtle notation.

6. CONCLUSION AND FUTURE WORK

We presented the RDF Data Cube browser CubeViz, its architecture, analysis components and visualization interfaces. In addition to the exploration of locally stored RDF data cubes it is possible to access remotely published ones using a combination of the SPARQL backend and the SPARQL services component. Such a setup was deployed on the European Commission's IT infrastructure as part of the European Data Portal[11].

There are further deployments of CubeViz made online such as LinkedSpending[12], which contain government spendings from all over the world represented and published as Linked Data (more than 2.4 million observations in 247 datasets). Using LinkedSpending, interested users can gather information about greek spending on police in certain regions in 2012 for instance (jump in using the button *Example Visualization 2* on the start page).

CubeViz is publicly available for download[13] and its latest releases can be evaluated using an online demonstrator[14]. CubeViz is under active development and will be further extended with new features such as drill-down functionality, additional interactive and customizable charts, further chart APIs such as the *Google Charts API*[15], aggregate functions and mashup features to compare observations from different domains.

7. ACKNOWLEDGEMENT

This work was supported by a grant from the EU's 7th Framework Programme provided for the project LOD2 (GA no. 257943).

8. REFERENCES

[1] Guidelines for Statistical Metadata on the Internet. Technical report, United Nations, Economic Commission for Europe (UNECE), 2000.

[2] Statistical data and metadata exchange (SDMX). Technical report, Standard No. ISO/TS 17369:2005, 2005.

[3] Management of Statistical Metadata at the OECD, 2006.

[4] R. Cyganiak, D. Reynolds, and J. Tennison. The RDF Data Cube vocabulary. Technical report, W3C, 2013. http://www.w3.org/TR/vocab-data-cube/.

[5] J. Demter, S. Auer, M. Martin, and J. Lehmann. LODStats – An Extensible Framework for High-performance Dataset Analytics. In *Proceedings of the EKAW 2012*, Lecture Notes in Computer Science (LNCS) 7603. Springer, 2012. 29

[6] S. Harris and A. Seaborne. SPARQL 1.1 Query Language - W3C Recommendation. Technical report, World Wide Web Consortium (W3C), 2013. `http://www.w3.org/TR/sparql11-query/`.

[7] N. Heino, S. Dietzold, M. Martin, and S. Auer. Developing Semantic Web Applications with the OntoWiki Framework. In *Networked Knowledge - Networked Media*, Vol. 221 of Studies in Comp. Intelligence. Springer, 2009.

[10] `http://cold.aksw.org/`

[11] `https://open-data.europa.eu/cubeviz/`

[12] `http://linkedspending.aksw.org/`

[13] `https://github.com/AKSW/CubeViz/`

[14] `http://cubeviz.aksw.org/`

[15] `https://developers.google.com/chart/`