A Dynamic Faceted Browser for Data Cube Statistical Data

Fadi Maali, Gofran Shukair, Nikolaos Loutas[1]
DERI, NUI Galway, Ireland
firstname.lastname@deri.org

1. Introduction

The European Commission (EC) encourages and facilitates Member States to open up government data and metadata [1-3]. In this vein, governments in turn have developed and are currently in the process of implementing open data strategies aspiring to increase transparency, encourage participation and improve government efficiency [4]. The amount of data published in open data portals is steadily growing. In some cases, data is published in human-readable, proprietary formats (e.g. pdf or doc), while in others data is published in machine-readable, open formats (e.g. xml, csv and rdf) [5].

Field research has shown that numerous open datasets contain statistical data. Statistical data is of paramount importance and have numerous applications in everyday life. Whether it's population, income, unemployment or interest rates, statistical data is a fundamental source of information for analysis, visualization, foundation of policy and decision making.

While there exists a number of mature standards to represent and publish statistical data, most notably SDMX[2], publishing statistical data as Linked Data promises a number of advantages. Using RDF, individual observations become web-addressable in a way that everyone can reference and annotate. RDF is a flexible, schema-less data model that eliminates the unified schema precondition as requirement of integration. Therefore, having the statistical data in RDF opens up the possibility of connecting it with other, not necessarily statistical data, and in a way that makes the data richer and more accessible. Beside these advantages, using RDF to represent statistical data makes complex queries feasible using SPARQL, the standard RDF querying language.

Data Cube vocabulary[3] is the state-of-the-art in representing statistical data in RDF. It is compatible with SDMX and increasingly being adopted. For example, data from the EU Digital Agenda Scoreboard[4] and the World Bank data[5] are available in Data Cube. A collection of Data Cube datasets is maintained at http://wiki.planet-data.eu/web/Datasets.

RDF isn't mainly designed for describing data in a human-readable way. The high dimensionality of statistical data adds to the challenges of providing the data to users in an understandable and easy-to-use manner. Moreover, writing SPARQL queries can be a

[1] Acknowledgement: The work presented in this paper has been funded in part by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

[2] http://sdmx.org/

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

[4] http://scoreboard.lod2.eu/index.php?page=export

[5] http://worldbank.270a.info/ published by a third party

challenging task even for seasoned practitioners. All these call for a generic browser that enables easy navigating, browsing and querying of Data Cube statistical data.

This position paper introduces a generic faceted browser for Data Cube statistical data. The browser provides a human-friendly interface to allow browsing, navigating and querying the data. Query formulation is enabled through a faceted interface dynamically built by examining the data structure. We present the architecture and the main functionalities provided by the tool next and conclude by discussing the faced challenges and lessonslearned.

2. Data Cube faceted browser

Figure 1 illustrates the main components of the browser and the user interaction with it. Navigation starts by entering a SPARQL endpoint URL in the browsing interface. The only prerequisite is that the endpoint support a subset of SPARQL 1.1 functions (COUNT and GROUP BY)[6]. The browser uses SPARQL protocol to examine the endpoint content and determine the available Data Cube datasets. For each dataset, it determines its dimensions and measures. This information is used to build a list of facets that can be used to browse and query the data.

Additionally, a full text search is supported. The search together with the facets help guiding the user to narrow results down to a particular subset that can then be exported as RDF data (in TURTLE) or as JSON (serialised in JSON-LD).

Presenting a single observation is also dynamically customised based on the data structure. In the summarised view, only properties that are defined as Data Cube dimensions, measures or attributes are presented. The full list of properties associated with an observation can still be viewed by clicking the details link. All values that appear to be HTTP URIs can be dereferenced (notice the ↗ icon next to country names and facets titles).

By integrating knowledge of core data model (i.e. Data Cube model) the browser can determine the useful facets of the data, mash together data from different datasets and meaningfully present the data.

A demo is available at: http://vmsgov03.deri.ie:8080/RDF-faceted-browser/start.html where any endpoint containing Data Cube data can be browsed. We also provide a demo endpoint containing the EU Digital Agenda Scoreboard Data.

[6] these function are widely supported even before SPARQL 1.1 via proprietary extensions

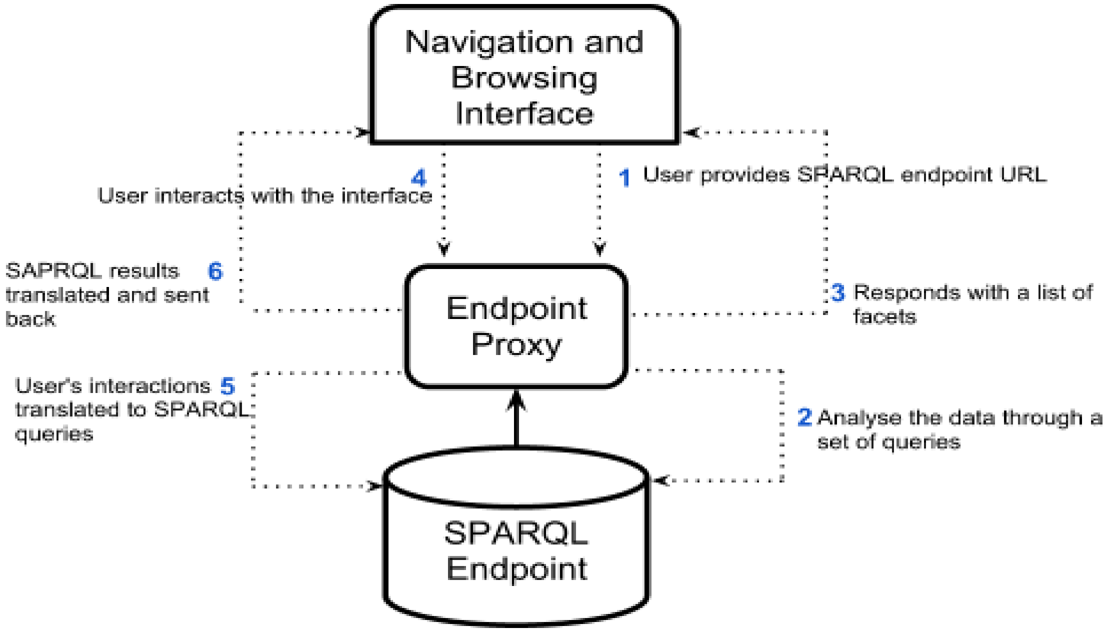
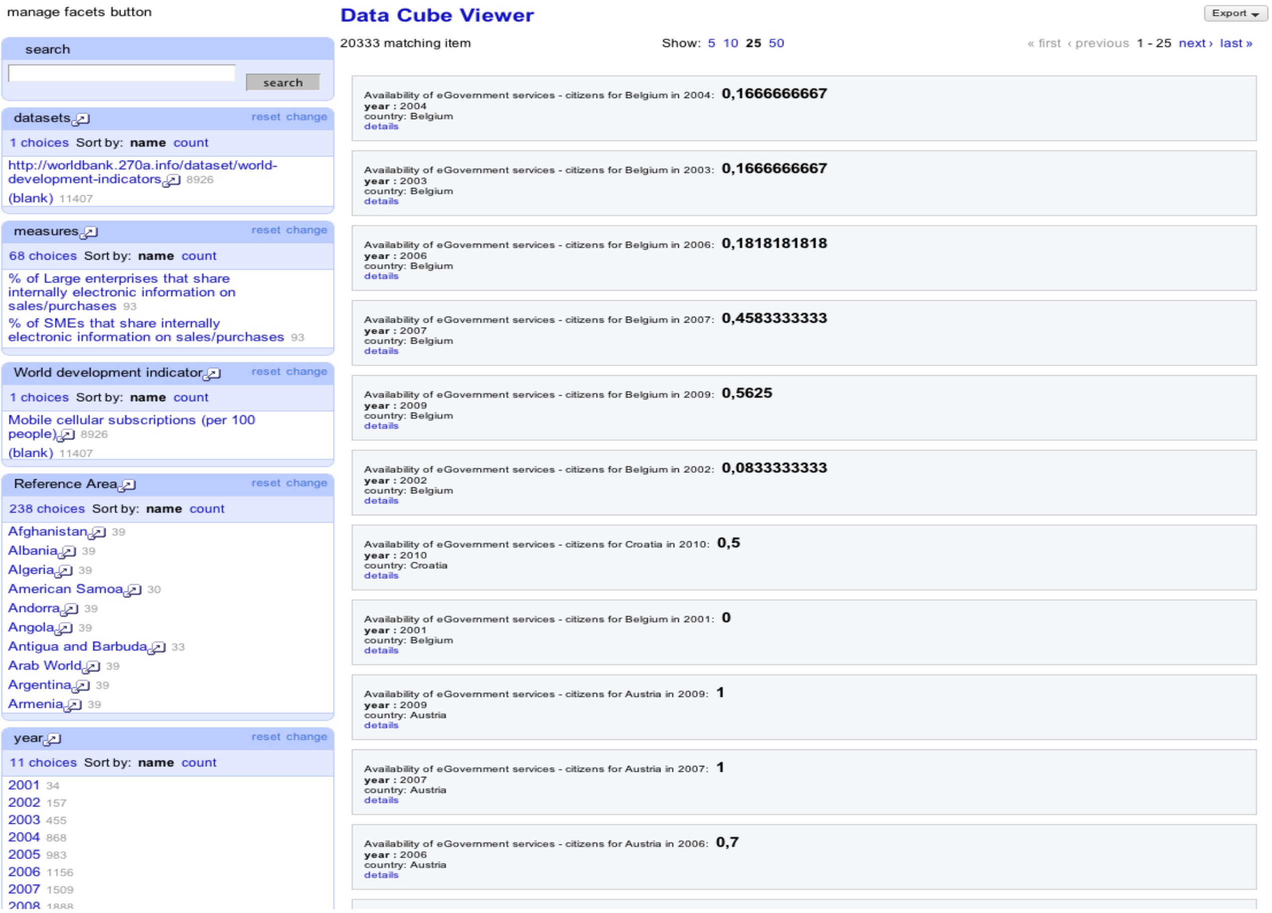

Navigation and Browsing Interface
1 User provides SPARQL endpoint URL
4 User interacts with the interface
SAPRQL results translated and sent back 6
Endpoint Proxy
3 Responds with a list of facets
User's interactions translated to SPARQL queries 5
2 Analyse the data through a set of queries
SPARQL Endpoint

Figure 1: Data Cube faceted browser structure


manage facets button

search

search

datasets — reset change

1 choices Sort by: **name** count

http://worldbank.270a.info/dataset/world-development-indicators 8926

(blank) 11407

measures — reset change

68 choices Sort by: **name** count

% of Large enterprises that share internally electronic information on sales/purchases 93

% of SMEs that share internally electronic information on sales/purchases 93

World development indicator — reset change

1 choices Sort by: **name** count

Mobile cellular subscriptions (per 100 people) 8926

(blank) 11407

Reference Area — reset change

238 choices Sort by: **name** count

Afghanistan	39
Albania	39
Algeria	39
American Samoa	30
Andorra	39
Angola	39
Antigua and Barbuda	33
Arab World	39
Argentina	39
Armenia	39

year — reset change

11 choices Sort by: **name** count

2001	34
2002	157
2003	455
2004	868
2005	983
2006	1156
2007	1509
2008	1888

Data Cube Viewer

Export

20333 matching item — Show: 5 10 **25** 50 — « first ‹ previous **1 - 25** next › last »

Availability of eGovernment services - citizens for Belgium in 2004: **0,1666666667**
year : 2004
country: Belgium
details

Availability of eGovernment services - citizens for Belgium in 2003: **0,1666666667**
year : 2003
country: Belgium
details

Availability of eGovernment services - citizens for Belgium in 2006: **0,1818181818**
year : 2006
country: Belgium
details

Availability of eGovernment services - citizens for Belgium in 2007: **0,4583333333**
year : 2007
country: Belgium
details

Availability of eGovernment services - citizens for Belgium in 2009: **0,5625**
year : 2009
country: Belgium
details

Availability of eGovernment services - citizens for Belgium in 2002: **0,0833333333**
year : 2002
country: Belgium
details

Availability of eGovernment services - citizens for Croatia in 2010: **0,5**
year : 2010
country: Croatia
details

Availability of eGovernment services - citizens for Belgium in 2001: **0**
year : 2001
country: Belgium
details

Availability of eGovernment services - citizens for Austria in 2009: **1**
year : 2009
country: Austria
details

Availability of eGovernment services - citizens for Austria in 2007: **1**
year : 2007
country: Austria
details

Availability of eGovernment services - citizens for Austria in 2006: **0,7**
year : 2006
country: Austria
details

Figure 2: main interface of Data Cube faceted browser

3. Discussion & Future Work

The high dimensionality of statistical data imposes an unavoidable complexity on both representing the data and using it. Data Cube is arguably a complicated model especially when compared to the lightweight ontologies increasingly being used recently in the Linked Data community. Along the same line, SPARQL, while being expressive and powerful, is a difficult query language especially for laypersons. In this paper, we presented a human friendly interface to statistical data represented in Data Cube. The faceted interface guides the user to formulate powerful queries without caring about the nitty-gritty details of SPARQL.

Two important premises underlying this work are:

- **A little semantics goes a long way**[7]: by making the tool aware of the core components of the Data Cube model, it is possible to slice and dice statistical datasets published independently by different entities without requiring them to use exactly the same terms.
- **Degrade gracefully**: with the Linked Data efforts moving beyond the bootstrapping phase, best practices and conventions for publishers are increasingly being crystalised. While tools should not assume the adoption of such practices, they should make their best to utilise them when available. A couple of examples from the tool we described:
  - Using common terms: as part of the Data Cube work, common terms defined originally in SDMX were defined as a set of extensions and were given URIs[8]. Examples of such terms include: gender, reference area and reference period. Publishers using these terms increase the utility of their data. The browser utilise this information, when used, to enhance data integration and presentation.
  - Dereferencing: URIs need to resolve to be part of the Linked Data. In practice, not all URIs do. The browser described here is able to enrich the user experience through dereferencing URIs; yet it continues to function when URIs do not resolve.

We also want to highlight a couple of design decisions that we think are useful and applicable in a wider scope:

- Centralised triplestore: while Linked Data enables distributed publication of data at a Web scale, querying the data through SPARQL still calls for loading the data in a central endpoint. While federated SPARQL querying is currently feasible, it still has a big penalty on performance and reliability. However, all needed to start browsing RDF datasets from multiple publishers is loading them into a central triple store, a straightforward cheap task. Relatedly, Erik Wilde argues that the killing application of the linked data might be in the Business Intelligence (read centralised warehouse) domain[9].
- We deliberately did not try to support visualization: being a generic browser, we decided not to include a visualization functionality in the browser. While supporting basic bar and pie charts would have been possible and cheap to implement, we believe that poor

[8] http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/vocab/

[7] We acknowledge Prof. James Hendler for this quote

[9] http://dret.typepad.com/dretblog/2011/05/from-ai-to-bi.html

visualisation is not only useless but actually harmful and misleading. Therefore, we didn't support out-of-the-box visualisation but enables exporting data in Turtle and in (the more visualisation friendly) JSON format.

We plan to extend the browser to enable the user to reconcile data coming from different datasets. We also plan to evaluate the use of SPARQL1.1 federated querying capabilities to enable browsing data from multiple endpoints.

References

1. European Commission (2010). Digital Agenda for Europe 2010 - 2020, Retrieved from: http://ec.europa.eu/information_society/digital-agenda/index_en.htm. [Accessed May 2012].
2. European Commission (2008). Results of the online consultation of stakeholders 'Review of the PSI directive'. Retrieved from: http://ec.europa.eu/information_ society/policy/psi/ docs/pdfs/online_consultation/report_psi_online_consultaion_ stakeholders.pdf. [Accessed May 2012].
3. European Commission, DG INFSO (2010). eGovernment Action Plan 2011 – 2015. Retrieved from: http://ec.europa.eu/information_society/activities/ egovernment/ action_plan_2011_2015/index_en.htm. [Accessed May 2012].
4. Huijboom, N., Van den Broek, T. (2011). Open data: an international comparison of strategies. In: European Journal of ePractice no 12. Retrieved from: http://www.epractice.eu/ files/European%20Journal%20epractice%20Volume%2012_1.pdf
5. Kalampokis, E., Tambouris, E., Tarabanis, K. (2011). A Classification Scheme for Open Government Data: Towards Linking Decentralized Data. In: International Journal of Web Engineering and Technology, Vol.6, No.3, pp.266-285.