OLAP2DataCube: An Ontowiki Plug-In for Statistical Data Publishing

Percy E. Rivera Salas[†§], Michael Martin[‡], Fernando Maia Da Mota[†],
Sören Auer[‡], Karin Breitman[†], Marco A. Casanova[†]
[†]*Depto. Informática*
Pontifícia Universidade Católica do Rio de Janeiro, Brazil
{*psalas,fmota,karin,casanova*}*@inf.puc-rio.br*
[‡]*AKSW, Computer Science*
University of Leipzig, Germany
{*lastname*}*@informatik.uni-leipzig.de*
[§]*Globo.com, Rio de Janeiro, Brazil*
percy@corp.globo.com

***Abstract*—Statistical data is one of the most important sources of information, relevant for large numbers of stakeholders in the governmental, scientific and business domains alike. In this article, we introduce an Ontowiki plugin that extracts and publishes statistical data in RDF. We illustrate the plugin with a comprehensive use case reporting on the extraction and publishing on the Web of statistical data about 10 years of Brazilian government.**

I. Introduction

Statistical data is one of the most important sources of information, relevant for large numbers of stakeholders. In the governmental domain, statistical data provides an anatomy of society outlining strong and weak points of governance thus providing crucial input for policy and decision makers. In science, statistical data representing observations or measurements is often a fundamental artifact to verify or refute scientific theories. In the business domain, statistical data about product sales, market developments or economic indicators provide crucial input for strategic decisions of the management. The elicitation of statistical data is very time and resource demanding, in particular in scenarios where different organizations are involved. This is particularly true for public statistical data, where local, regional, state-level, national/federal and supranational organizations are involved in the definition of statistical criteria and the elicitation of statistic ground truth. In order to aggregate and integrate statistical data it is of paramount importance that the statistical criteria are semantically described and linked to suitable ontologies or background knowledge bases.

In this article, we overview how statistical data can be managed on the Web using Linked Data. We present the *OLAP2DataCube* plug-in to efficiently transform large analytical databases, represented according to the Online Analytical Processing (OLAP) paradigm, to RDF. *OLAP2DataCube* uses the RDF Data Cube Vocabulary, which is based on the popular SDMX standard[1] and designed particularly to represent multidimensional statistical data using RDF. The vocabulary also uses the SDMX feature of content oriented guidelines (COG), which define a set of common statistical concepts and associated code lists that can be re-used across datasets.

As a comprehensive use case we report about the creation of *dados.gov.br* – the extraction and publishing data about 10 years of Brazilian government. The dados.gov.br information catalog has over 1,300 historic data series that reflect government activity during the mandate president Luiz Inacio "'Lula"' da Silva (2003 to 2010). The dataset comprises more than 4 million observations covering three levels of administration in Brazil. It is expressed in more than 30 million RDF triples

The remainder of the paper is organized as follows. section II describes the OLAP2DataCube plugin. section III contains the dados.gov.br use case. section IV discusses related work. Finally, section V contains conclusions and lessons learned.

II. Extracting and publishing statistical data

In this section, we present the OLAP2DataCube plug-in for extracting statistical data from OLAP sources. OLAP2DataCube is implemented as a plug-in extension into *OntoWiki* [1]. OntoWiki is a tool for browsing and collaboratively editing RDF knowledge bases. It differs from other Semantic Wikis insofar as OntoWiki uses RDF as its natural data model instead of Wiki texts. Information in OntoWiki is always represented according to the RDF statement paradigm and can be browsed and edited by means of views. These views are generated automatically by employing the ontology features such as class hierarchies or domain and range restrictions. OntoWiki adheres to the Wiki principles by striving to make the editing of information as simple as possible and by maintaining a comprehensive revision history. This history is also based on the RDF statement paradigm and allows to roll-back prior change-sets. OntoWiki incorporates a number of Linked

[1]http://sdmx.org

Data features, such as exposing all information stored in OntoWiki as Linked Data as well as retrieving background information from the Linked Data Web. Apart from providing a comprehensive user interface, OntoWiki also contains a number of components for the rapid development of Semantic Web applications, such as the RDF API Erfurt, methods for authentication, access control, caching and various visualization components. In addition to ontology engineering tasks, OntoWiki provides ontology evolution functionality, which can be used to further transform the newly converted statistical data. Furthermore, OntoWiki provides various interfaces (in particular Linked Data and SPARQL interfaces) to publish and query RDF data.

A. OLAP2DataCube

A cube is represented in a relational database as a set of tables, organized in the shape of a star or a snowflake. *Star schemas* are composed of one or more fact tables that reference dimension tables. *Snowflake schemas*, on the other hand, are a more complex variation, where dimension tables are normalized into multiple, related tables.

The input to the OLAP2DataCube[2] plugin is a relational database with an star model. Its output is a tripleset, mapped from the OLAP cube using the RDF Data cube vocabulary.

The process encompasses three stages: (1) relational database metadata extraction and table categorization, (2) cube definition, and (3) RDF mapping. We detail each stage in the sequel.

1) Metadata extraction and Table categorization: In this step we query the database data dictionary and extract existing metadata, e.g. tables, primary keys (PKs) and foreign keys (FKs), and distinguish between fact and dimension tables. The categorization is done (manually) based on the analysis of table relationships. For example, a table with several FKs to other tables is likely to be a fact table. On the other hand, a table with no FKs is more likely to be a dimension table.
2) Cube definition: In this step, we define a cube, guided by the following choices:
   a) *Fact Table Selection:* The user chooses one of the fact tables identified in the table categorization step.
   b) *Dimension Table Selection:* The user selects dimension tables that are related to the chosen fact table.
   c) *Metadata Annotation:* To facilitate future use and promote interoperability, the user provides additional information about the dataset in question. This can be, for example, name, description, and units of measures (if appropriate). The metadata can be stored in a separate dimension table, and accessed as a special dimension table.

[2]https://github.com/AKSW/olapimport.ontowiki

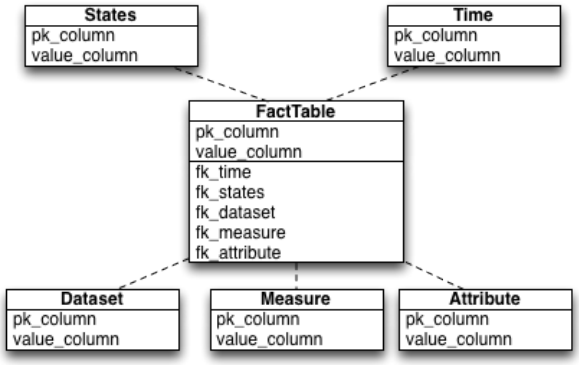

Figure 2. Database schema used as example in section III.

Prefix	Table Type
fact	Fact
dim	Dimension
attr	Attribute
mea	Measure
data	Dataset

Table I
EXEMPLARY SELECTED TRANSFORMATION RULES.

It is important to note that the defined cube does not necessarily have to be a cuboid (three-dimensional cube), but it may be multidimensional. The boundary is the number of dimension tables associated with the chosen fact table.
The OLAP2DataCube OntoWiki plugin provides an interactive interface that guides users during the selection process. Figure 1 depicts the plugin in action, as seen by the user. In this particular example, the fact table (`dado_ficha`) was selected (in a previous step, not depicted). The plugin is now prompting for the selection of the dimension table(s) that should be part of the data cube, by displaying the total range of possibilities (all dimension tables related to the `dado_ficha` fact table).
3) Mapping: In this stage the cube is mapped to RDF. The cube definition, conceived in the previous step, is internally transformed into an SQL query, which extracts the envisioned data from the relational database.

In the following, we exemplify some of the transformation rules used in the process. The schema we used as example in section III is depicted in Figure 2 and the conventions are listed in Table I is adopted. The SQL fragments below follow this schema.

1) The values selected by the SQL query are taken from the fact, dimension and attribute tables chosen by the user in the cube construction step. The prefix indicates the table type. For the database schema of our running

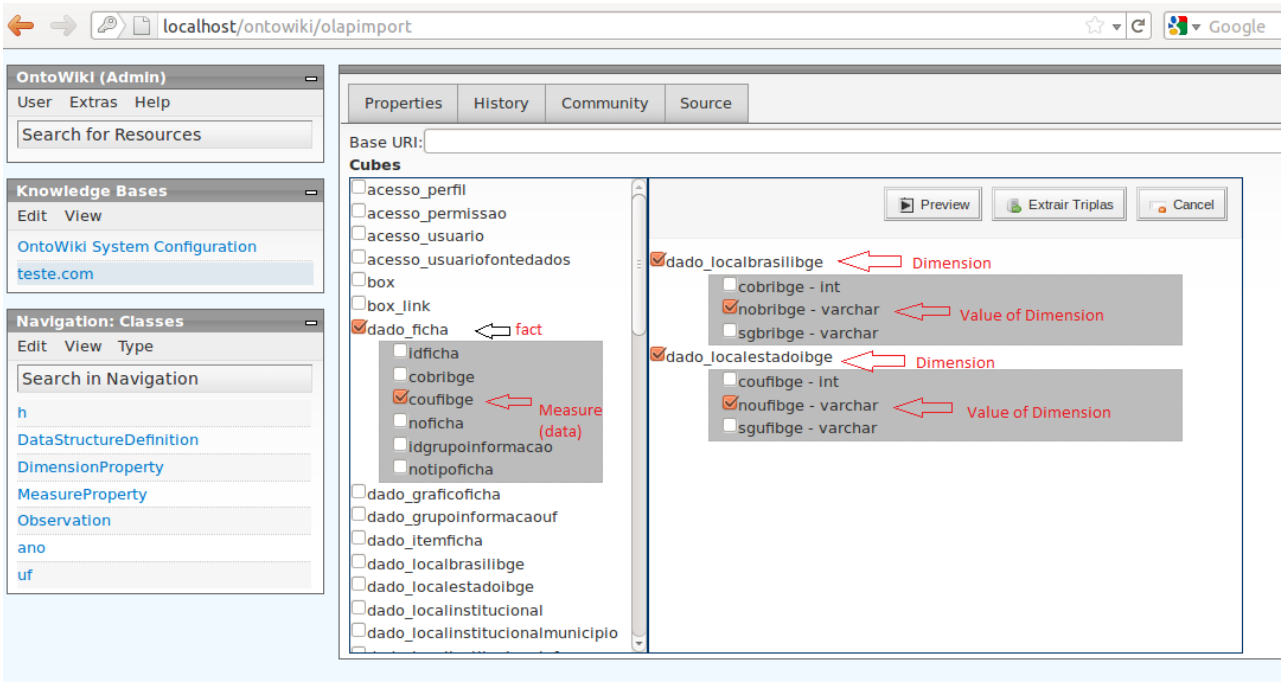

Figure 1. The OLAP2DataCube OntoWiki extension.

example (see Figure 2), we would have:

```
SELECT dim1.value_column, dim2.value_column, fact.value_column
  data.value_column, meas.value_column, attr.value_column,
```

2) Each selected dimension table generates an individual JOIN operation with the fact table. In this operation each of the tables is identified with its type prefix. Use FKs and PKs relationships as identified during step 1. Again, for the database schema of our running example, we would have:

```
FROM FactTable AS fact, Time AS dim1, States AS dim2
  WHERE fact.fk_time = dim1.pk_column
    AND fact.fk_states = dim2.pk_column
```

3) For each selected special dimension table (i.e. dataset, measure, attribute) generate a JOIN operation with the fact table. Again, for the database schema of our running example, we would add the following assertions to the `FROM` and `WHERE` clauses:

```
FROM ..., Dataset AS data, Measure AS meas, Attribute AS attr
WHERE  ...
    AND fact.fk_dataset   = data.pk_column
    AND fact.fk_measure   = meas.pk_column
    AND fact.fk_attribute = attr.pk_column
```

Query results are then mapped to corresponding concepts in the RDF Data Cube vocabulary [2]. We exemplify some mapping rules in the sequel:

1) Each dimension table is defined as an instance of `qb:DimensionProperty`, so that each tuple in them is an instance of the new dimension. Special dimension tables receive similar treatment.

```
#New dimension definition
dim:Time       rdf:type    qb:DimensionProperty ;
               rdfs:label  "Year" .
#New dimension individuals
times:T2010    rdf:type    dim:Time;
               rdfs:label  "2010"^^xsd:int
```

2) Tuples that result from the SQL query are instances of the type `qb:Observation`, and are mapped taking column labels into consideration, as their label reflects the type of data they represent (dimension, dataset, attribute, measure or fact).

```
# New dimension definition
observations:O1 rdf:type qb:Observation ;
               dim:Time times:T2010 ;
               dim:State states:Rio_de_Janeiro ;
               qb:dataset datasets:Emprego_Criado ;
               sdmx:unitMeasure attributes:Emprego ;
               measures:Emprego "1031473" .
```

III. Dados.gov.br – Exposing 10 years of statistics about Brazil on the Web of Data

Efforts towards the publication of Open Government Data (OGD) in Brazil can be traced back to 2009, when the Information Organizing Committee of the Presidency (COI-PR), started to amass large amounts of aggregated government data for digital publication. The goal of the committee was to create a central information catalog of

public activity, with the intent of improving governance, and monitoring government activity. This catalog was originally created to serve the President of the Republic and his team of advisors, as a reliable source of official data. The project was so successful that, reflecting open data principles, the catalog was made available to the general public in 2010.

In September 2011 Brazil became a member of the Open Government Partnership[3], a multinational initiative to promote worldwide adoption of OGD. As a participating member, Brazil committed to public transparency and action in securing open publication of official data. The commitment comprises political, as well as technical landmarks, including a presidential mandate for the launch of the Brazilian open government data portal.

The Dados.Gov.br information catalog comprises over 1,300 historic data series that reflect government activity during the mandate of President Luiz Inacio 'Lula' da Silva (2003 to 2010). The COI management team proposed a standard organization to classify the data, based on two dimensions: territorial (country, states, cities) and time (year or month). Data series were classified in several hierarchical thematic trees, that branched from general to more specific subjects, e.g., infrastructure, citizenship and social inclusion, as well as more specific subjects that define third and fourth level trees. Data (not in Linked Data format) is publicly available[4].

As a result of our publishing effort (employing the techniques described in the previous sections), we obtained an anatomy of 10 years of Brazilian government reaching in some cases even 30 years back in time. Table II summarizes the results of our publishing effort. The dataset comprises more than 4 million observations covering three levels of administration in Brazil. It is expressed in more than 30 million RDF triples, linked to DBpedia and GeoNames. The conversion took approximately 60 hours, which appears reasonable due to the amount of raw data (1GB) and the transformation process stretching over the stages extraction/transformation, serialization, insertion/loading. The time consuming steps, here, are the first and last stages. During the first stages we had to run extensive SQL queries to extract data from the database. Not only was that time consuming, but also slowed down due to the fragmentation of the data in 900+ separate datasets.

Criterion	Measurement
Base data	
Data size	1GB
Data entries	4,514,612
Conversion Process	
Triples	31,120,766
Conversion Time	$\approx$ 3,600 min
Data About	
Municipalities	5,320
States	28
Series	937
Years	27
Data Sources	77
Observations (qb:Observation)	4,110,045
Municipality	4,016,902
State	87,304
Brazil	5,839
Dimensions (qb:Dimension)	6
Datasets (qb:DataSet)	937
Measures (sdmx:unitMeasure)	119

Table II
RESULTS STATISTICS (HTTP://PURL.ORG/GOVDATACUBE).

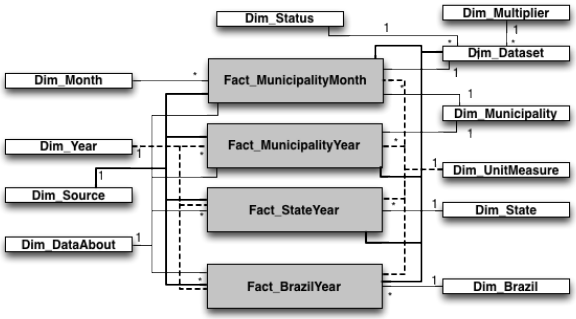

Figure 3. The Dados.gov.br OLAP data model.

IV. RELATED WORK

Related work can be roughly divided into other RDF triplification approaches and statistical data publishing.

Currently most of the work in the area of triplification focuses on generating RDF from relational database content. There is a wide range of approaches developed in this regard ranging from very simple scripts such as *Triplify* [3] over standalone solutions such as *D2R* [4] up to integrated tools such as *Virtuoso RDF Views* [5]. Under the auspices of the W3C, the *RDB2RDF working group* is currently standardizing the *R2RML* mapping language for the mapping and transformation of relational data to RDF. One of the few works in the area of transforming statistical data to RDF is [6], which explores the opposite direction to our approach, i.e., the transformation of statistical Linked Data for use in OLAP systems.

Statistical Data and Metadata eXchange (SDMX, [7]) is an initiative started in 2001 to foster standards for the exchange of statistical information. The SDMX sponsoring institutions are the Bank for International Settlements, the European Central Bank, Eurostat, the International Monetary Fund (IMF), the Organisation for Economic Co-operation and Development (OECD), the United Nations Statistics Division and the World Bank. The SDMX message formats have two basic expressions, SDMX-ML (using XML syntax) and SDMX-EDI (using EDIFACT syntax and based on

[3]http://www.opengovpartnership.org/
[4]https://i3gov.planejamento.gov.br/

the GESMES/TS statistical message). Experiences and best practices regarding the publication of statistics on the Web in SDMX have been published by the United Nations [8] and the Organisation for Economic Co-operation and Development [9].

The representation of statistics in RDF started with SCOVO [10], [11] and continued with the successor RDF Data Cube Vocabulary [2]. The Data Cube Vocabulary is closely aligned with SDMX [11]. Examples of statistics published as RDF adhering to the Data Cube vocabulary and visualized for human consumption include the EC's INFSO Digital Agenda Scoreboard[5] and the LOD2 Open Government Data stakeholder survey [12].

V. Conclusions, lessons learned, and future work

In this paper we introduced the OLAP2DataCube plug-in for extracting and publishing statistical data on the Data Web. We also presented a large-scale use case of statistic data publishing in Brazil. We validated the plug-in by converting a very large database composed of a little over nine hundred datasets, whose data spans nearly 30 years of Brazilian government (amounting to 30 million triples). It took the proposed plug-in nearly three days to complete the task, placing it well among existing tools. It also served to demonstrate OLAP2DataCube's robustness and scalability.

The work described in this article is a first step towards a larger research and development agenda aiming at facilitating the life-cycle of statistical data on the Web. As promising future directions, we may quote in particular the following. First, we will invest on the semi-automated generation of links and visualizations. Currently, it is still cumbersome to configure the linking and visualization tools. A possible approach to simplify the generation of configurations is the use of user provided examples or the analysis of navigation logs for learning suitable configurations automatically. Second, we may quote the semi-automatic integration and comparison of statistic data from distributed sources, which could ultimately lead to a rich and diverse Statistical Data Web.

Our plan is to continue to use the Ontowiki platform and the plug-in architecture. We believe beneficial to our purposes because it provides a common framework of shared functionality, that helps reduce users learning curve, and the cognitive overload of learning to use different tools; secondly, Ontowiki has an active community of users, that developed over a dozen plug-ins for the framework. We expect to capitalize over existing plug-ins, that can be pipelined to produce new functionality.

[5]http://ec.europa.eu/information_society/digital-agenda/scoreboard/

Acknowledgment

This work was partly supported by CNPq, under grants 301497/2006-0, 305824/2010-4, 475717-2011-2, and 557128/2009-9, by FAPERJ under grant E-26/170028/2008 and by Globo.com

References

[1] S. Auer, S. Dietzold, and T. Riechert, "OntoWiki - A Tool for Social, Semantic Collaboration." in *ISWC*, ser. LNCS, vol. 4273. Springer, 2006.

[2] "The rdf data cube vocabulary," Tech. Rep., 2010. [Online]. Available: http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html

[3] S. Auer, S. Dietzold, J. Lehmann, S. Hellmann, and D. Aumueller, "Triplify: Light-weight linked data publication from relational databases," in *WWW*. ACM, 2009. [Online]. Available: http://doi.acm.org/10.1145/1526709.1526793

[4] C. Bizer and R. Cyganiak, "D2r server - publishing relational databases on the semantic web," Poster at ISWC, 2006. [Online]. Available: http://www4.wiwiss.fu-berlin.de/bizer/pub/Bizer-Cyganiak-D2R-Server-ISWC2006.pdf

[5] O. Erling, "Automated Generation of RDF Views over Relational Data Sources with Virtuoso," 2009. [Online]. Available: http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/VOSSQL2RDF

[6] B. Kämpgen and A. Harth, "Transforming statistical linked data for use in olap systems," in *I-SEMANTICS 2011*, 2011.

[7] "Statistical data and metadata exchange (sdmx)," Standard No. ISO/TS 17369:2005, Tech. Rep., 2005.

[8] "Guidelines for statistical metadata on the internet," United Nations, Economic Commission for Europe (UNECE), Tech. Rep., 2000.

[9] "Management of statistical metadata at the oecd," 2006.

[10] M. Hausenblas, W. Halb, Y. Raimond, L. Feigenbaum, and D. Ayers, "Scovo: Using statistics on the web of data." in *ESWC*, ser. LNCS, vol. 5554. Springer, 2009. [Online]. Available: http://dblp.uni-trier.de/db/conf/esws/eswc2009.html#HausenblasHRFA09

[11] R. Cyganiak, S. Field, A. Gregory, W. Halb, and J. Tennison, "Semantic statistics: Bringing together sdmx and scovo." in *LDOW*, ser. CEUR Workshop Proceedings, vol. 628. CEUR-WS.org, 2010. [Online]. Available: http://dblp.uni-trier.de/db/conf/www/ldow2010.html#CyganiakFGHT10

[12] M. Martin, M. Kaltenbck, H. Nagy, and S. Auer, "The open government data stakeholder survey," in *OKCon*. OKFN, 2011. [Online]. Available: http://okcon.org/2011/programme/the-open-government-data-stakeholder-survey