

International Journal of Semantic Computing
© World Scientific Publishing Company

PUBLISHING STATISTICAL DATA ON THE WEB

PERCY E. RIVERA SALAS, KARIN K. BREITMAN, MARCO A. CASANOVA

*Informatics Department
Pontifical Catholic University of Rio de Janeiro, Brazil
{psalas,fmota,karin,casanova}@inf.puc-rio.br*

MICHAEL MARTIN, SÖREN AUER

*AKSW, Computer Science
University of Leipzig, Germany
{lastname}@informatik.uni-leipzig.de
<http://aksw.org/>*

Received (Day Month Year)

Revised (Day Month Year)

Accepted (Day Month Year)

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 overview how statistical data can be managed on the Web. With *OLAP2DataCube* and *CSV2DataCube* we present two complementary approaches on how to extract and publish statistical data. We also discuss the linking, repair and the visualization of statistical data. As a comprehensive use case, we report on the extraction and publishing on the Web of statistical data describing 10 years of life in Brazil.

Keywords: Statistical Data; Linked Data; Open Government Data.

1. 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, specially 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

2 Percy E. Rivera Salas, Michael Martin, Karin K. Breitman, Sören Auer and Marco A. Casanova

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 two complementary approaches on how to extract, represent and publish statistical data. With the *OLAP2DataCube* tool, large analytical databases, represented according to the Online Analytical Processing (OLAP) paradigm, can be efficiently transformed into RDF. With the *CSV2DataCube* tool, statistical data available in CSV files spreadsheets can be easily converted into RDF. Both approaches use the RDF Data Cube Vocabulary, which is based on the popular SDMX standard^a and designed 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.

We also discuss the application of existing, general purpose link discovery tools for linking of statistical data. Interlinking various statistical dimensions (such as cities, regions or states with GeoNames) facilitates the unforeseen integration of independently gathered statistical data. We exhibit a comprehensive but generic solution for the visualization of statistical data by means of highly configurable charts.

As a comprehensive use case we report on the creation of *dados.gov.br* – the extraction and publishing of statistical data on the Web describing 10 years of life in Brazil. The *dados.gov.br* information catalog has over 1,300 historic data series that reflect government activity during the mandate of 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 being initially linked to DBpedia and GeoNames.

The remainder of the paper is organized as follows. Section 2 discusses the representation of statistical data in RDF. Section 3 describes the *OLAP2DataCube* and the *CSV2DataCube* tools. Section 4 addresses the problem of link discovery for statistical data. Section 5 covers the visualization of statistical data. Section 6 contains the *dados.gov.br* use case. Section 7 discusses related work. Finally, Section 8 contains conclusions and lessons learned.

2. Representation of Statistical Data in RDF

Following Cyganiak et al. [4], a *statistical data set* comprises a collection of *observations* made at some points across some logical space. The collection can be characterized by a set of *dimensions* d_1, \dots, d_m that define what the observations apply to, along with metadata attributes a_1, \dots, a_n describing what has been measured, how it was measured and how the observation measures o are expressed. The values of each dimension d_i (of each attribute a_j or of the observation measures o)

^a<http://sdmx.org>

are taken from a dimension domain D_i (an attribute domain A_j or an observation measures domain O , respectively).

A statistical data set therefore defines a relation $R \subseteq D_1 \times \dots \times D_m \times A_1 \times \dots \times A_n \times O$, commonly referred to as a *data cube* or simply as a *cube*. A tuple of values from the dimension domains identifies an observation measure value and the associated attribute values, that is, R is actually a function of the form $R : D_1 \times \dots \times D_m \rightarrow A_1 \times \dots \times A_n \times O$.

According to Noy and Rector [20], we may represent R by reification (‘Pattern 1: Introducing a new class for a relation’ in [20]), that is, by creating a new class r and treating the dimensions, attributes and observation measure as properties. Thus, a tuple $(x_1, \dots, x_m, y_1, \dots, y_n, z)$ in R is represented by $m + n + 1$ triples $(u, d_1, x_1), \dots, (u, d_m, x_m), \dots, (u, a_1, y_1), \dots, (u, a_n, y_n), (u, o, z)$. The OLAP2DataCube approach follows this reification strategy.

Cubes are often exported as spreadsheets, which are bi-dimensional matrices. This is possible by selecting a dimension D_i and treating $R : D_1 \times \dots \times D_m \rightarrow A_1 \times \dots \times A_n \times O$ as a function with two arguments $R : (D_1 \times \dots \times D_{i-1} \times D_{i+1} \times \dots \times D_m) \times D_i \rightarrow A_1 \times \dots \times A_n \times O$. Then, a tuple $(x_1, \dots, x_m, y_1, \dots, y_n, z)$ in R is represented by a tuple of values $(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_m)$ taken from the spreadsheet heading, a value x_i taken from a line of the spreadsheet and a tuple of values (y_1, \dots, y_n, z) obtained from the corresponding cell (usually just the observation measure value z). With this interpretation, one can then extract $m + n + 1$ triples $(u, d_1, x_1), \dots, (u, d_m, x_m), \dots, (u, a_1, y_1), \dots, (u, a_n, y_n), (u, o, z)$ to represent the tuple $(x_1, \dots, x_m, y_1, \dots, y_n, z)$ in R . Figure 3 shows an example of this strategy to represent cubes and how the CSV2DataCube tool helps the user through the process of extracting triples from a spreadsheet.

Both tools use the RDF Data Cube vocabulary^b [4], specifically designed to publish multidimensional statistics on the Web in such a way that it can be linked to related RDF datasets. Very briefly, to encode structural information about the observations, the RDF Data Cube vocabulary contains a set of concepts, such as `qb:DataStructureDefinition`, `qb:DataSet` and `qb:Slice`. It represents data cube dimensions, attributes, and measures as RDF properties. Each property is an instance of the abstract class `qb:ComponentProperty`, which in turn has sub-classes `qb:DimensionProperty`, `qb:AttributeProperty` and `qb:MeasureProperty`.

Finally, we observe that the dimension domain values, as well as the attribute domain values, should also be properly described through RDF triples, in much the same way as the conceptual vocabulary (see [4]). Section 4 further discusses this point and provides examples of dimension domain values described as triples.

^bqb: <http://purl.org/linked-data/cube#>

4 Percy E. Rivera Salas, Michael Martin, Karin K. Breitman, Sören Auer and Marco A. Casanova

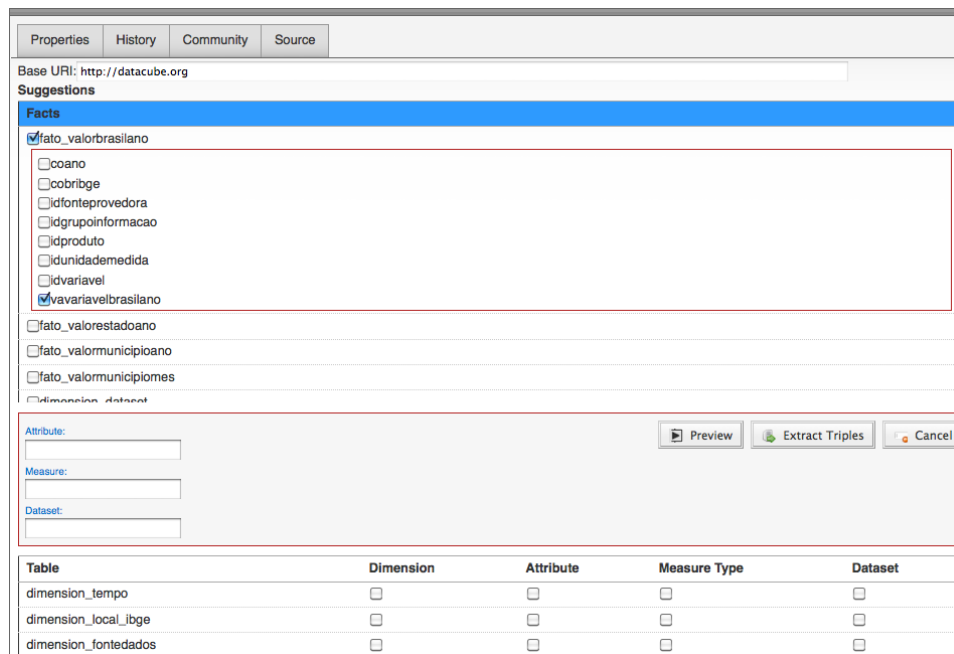


Fig. 1. The OLAP2DataCube OntoWiki extension.

3. Extracting and Publishing Statistical Data

In this section we present the OLAP2DataCube and the CSV2DataCube tools, two complementary approaches for extracting statistical data from OLAP and CSV sources, respectively. They are both implemented as plug-in extensions into *OntoWiki* [6]. *OntoWiki* is a tool that supports collaborative creation, maintenance and publication of RDF knowledge bases. 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.

3.1. 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^c plugin is a relational database, with an star

^c<https://github.com/AKSW/olapimport.ontowiki>

model. Its output is a tripliset, 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.

Metadata extraction In this step we query the database data dictionary and extract existing metadata, e.g. tables, primary keys (PKs) and foreign keys (FKs).

Table categorization In this step we distinguish between fact and dimension tables. The categorization is done (manually) based on the analysis of table relationships. For example, a table with several FK relationships to other tables is likely to be a fact table. On the other hand, a table with few relationships is more likely to be a dimension table.

Cube definition In this step, we define a cube, guided by the following choices:

- (1) *Fact Table Selection*: The user chooses one of the fact tables identified in the table categorization step.
- (2) *Dimension Table Selection*: The user selects dimension tables that are related to the chosen fact table.
- (3) *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.

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 now prompts the user to select 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).

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 6 is depicted in Figure 2

6 *Percy E. Rivera Salas, Michael Martin, Karin K. Breitman, Sören Auer and Marco A. Casanova*

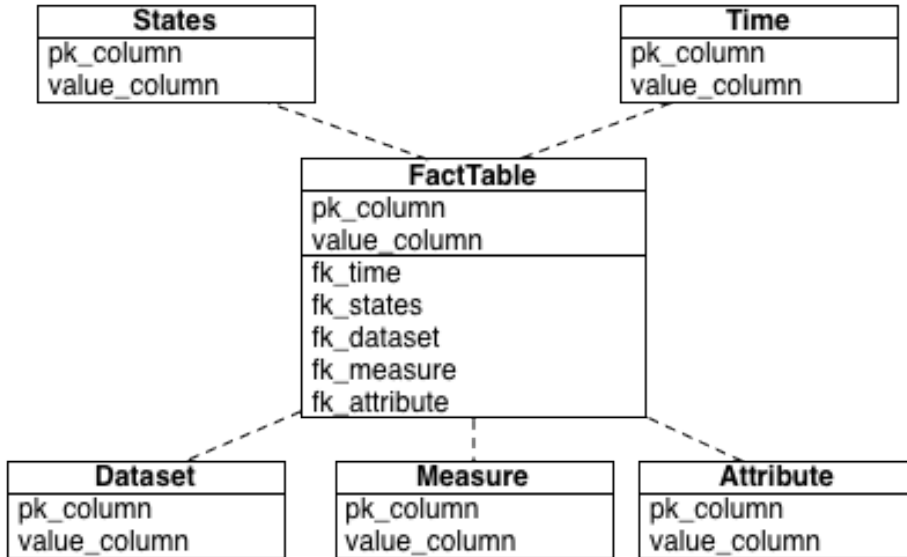


Fig. 2. The database schema used as example in Section 6.

Table 1. Convention used in transformation rules.

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

and the convention are listed in Table 1 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 definition step. The prefix indicates the table type. For the database schema of our running example (see Figure 2), we would have:

```

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

- (2) Each selected dimension table generates an individual JOIN operation with the fact table, using the FKs and PKs identified during step 1. In this operation each of the tables is identified with its type prefix. Again, for the database

schema of our running example, we would have:

```

1 FROM FactTable AS fact, Time AS dim1, States AS dim2
2 WHERE fact.fk_time = dim1.pk_column
3 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:

```

1 FROM ..., Dataset AS data, Measure AS meas, Attribute AS attr
2 WHERE ...
3 AND fact.fk_dataset = data.pk_column
4 AND fact.fk_measure = meas.pk_column
5 AND fact.fk_attribute = attr.pk_column

```

Query results are then mapped to corresponding concepts in the RDF Data Cube vocabulary [4]. 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 the table is an instance of the new dimension. Special dimension tables receive similar treatment.

```

1 #New dimension definition
2 dim:Time      rdf:type      qb:DimensionProperty ;
3               rdfs:label    "Year" .
4 #New dimension individuals
5 times:T2010   rdf:type      dim:Time;
6               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).

```

1 # New Observation instance
2 observations:01  rdf:type      qb:Observation ;
3                 dim:Time      times:T2010 ;
4                 dim:State      states:Rio_de_Janeiro ;
5                 qb:dataset      datasets:Emprego_Criado ;
6                 sdmx:unitMeasure attributes:Emprego ;
7                 measures:Emprego "1031473" .

```

Finally, the IRIs are generated from the base IRI <http://purl.org/GovDataCube>. Table 2 summarizes the IRIs generation rules used in the process.

3.2. CSV2DataCube

Statistical data is, in addition to OLAP systems, often collected and represented in simple spreadsheets. The CSV2DataCube^d tool described in this section facilitates the semi-automatic transformation of spreadsheets into data cubes.

^d<https://github.com/AKSW/csvimport.ontowiki>

Table 2. Examples of IRI Generation Rules.

Table Type	IRI Generation Rule
Dimension	<baseIRI>+"dimension/"<urlencode(dimensionName)>
Attribute	<baseIRI>+"attribute/"<urlencode(attributeName)>
Measure	<baseIRI>+"measure/"<urlencode(<measureName>)>
Dataset	<baseIRI>+"dataset/"<urlencode(<datasetName>)>
Observation	<baseIRI>+"item/"<md5(<dimensionsName>+<itemValue>)>

As is illustrated in Figure 3, when a spreadsheet containing multidimensional statistical data is imported into OntoWiki, it is presented as a table. This presentation of the data gives the users the ability to configure (1) dimensions and (2) attributes by manually creating them and selecting all elements belonging to a certain dimension and (3) the range of statistical items that are measured. The corresponding COG concepts are automatically suggested, using RDFa, when a user enters a word in the text box provided. It is also possible to save and reuse these configurations for other spreadsheets, which adhere to the same structure (e.g. for data published in consecutive years). Once the transformation is configured by the user, the Data Cube importer plugin for OntoWiki takes care of automatically transforming the spreadsheets into RDF adhering to the RDF Data Cube Vocabulary.

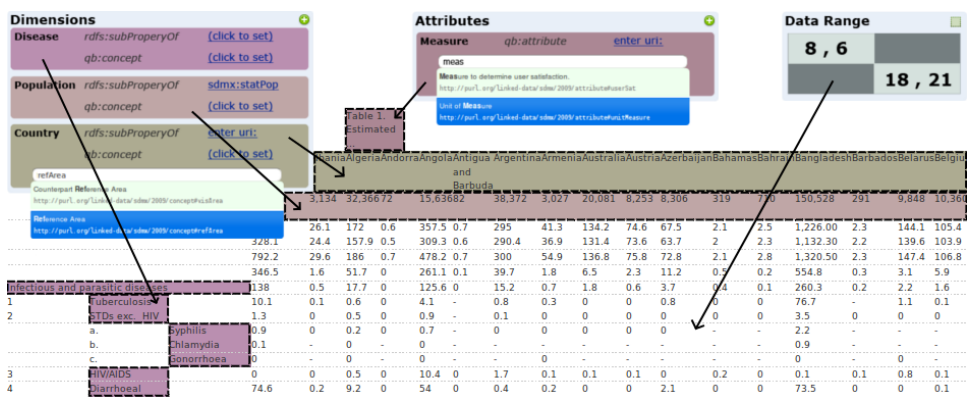


Fig. 3. Screenshot of the OntoWiki statistical data import wizard displaying a GHO table configured for conversion into RDF.

4. Linking Statistical Data

Establishing links between heterogeneous and distributed data sources is one of the fundamental features of the Web of Data. In this section we describe the application

```

1 <LIMES>
2   <PREFIX> [...] </PREFIX>
3   <SOURCE>
4     <ID>DBpedia</ID> <ENDPOINT>http://live.dbpedia.org/sparql</ENDPOINT>
5     <PAGESIZE>1000</PAGESIZE> <VAR>?x</VAR>
6     <RESTRICTION>?x rdf:type dbpedia-o:PopulatedPlace .
7       ?x dbpedia-o:country dbpedia-r:Brazil </RESTRICTION>
8     <PROPERTY>rdfs:label AS lowercase->nolang</PROPERTY>
9   </SOURCE>
10  <TARGET>
11    <ID>Dados</ID> <ENDPOINT>http://lod2.inf.puc-rio.br/sparql</ENDPOINT>
12    <PAGESIZE>1000</PAGESIZE> <VAR>?y</VAR>
13    <RESTRICTION>?y rdf:type dados-dim:localmunicipioibge</RESTRICTION>
14    <PROPERTY>rdfs:label AS lowercase->nolang</PROPERTY>
15  </TARGET>
16  <METRIC>levenshtein(x.rdfs:label, y.rdfs:label)</METRIC>
17  <ACCEPT><THRESHOLD> 1</THRESHOLD><RELATION>owl:sameAs</RELATION></ACCEPT>
18  <REVIEW><THRESHOLD>0.5</THRESHOLD><RELATION>owl:sameAs</RELATION></REVIEW>
19  <OUTPUT>N3</OUTPUT>
20 </LIMES>

```

Listing 1. LIMES spec. for discovering links between dados.gov.br and DBpedia.

of existing general purpose link discovery tools (such as LIMES [19] or SILK [22]) for linking of statistical data.

Interlinking various statistical dimensions (such as municipalities and states with DBpedia and GeoNames) facilitates the unforeseen integration of independently gathered statistical data. A fundamental difference between the link discovery in RDF knowledge bases and statistics represented in RDF is that in the former case links are established between the instance data items, while in the statistics case links are established between annotations of the instance data, i.e. the instances of the component property classes dimensions, attributes and measures.

We illustrate link discovery with the help of an example of a LIMES configuration for discovering links between *dados.gov.br* and *DBpedia*. Link discovery is a very resource and time intensive task, since potentially extremely large number of instances have to be compared. A naive approach to link discovery between dados.gov.br and DBpedia, for example, would require $4,110,045 * 3,500,000 \approx 14 * 10^{12}$ comparisons. LIMES [19] utilizes the mathematical characteristics of metric spaces to compute pessimistic estimates of the similarity between instances. These estimates are then used to filter out a large amount of those instance pairs that do not meet the mapping conditions. Thus, LIMES can reduce the number of comparisons needed during the mapping process by several orders of magnitude.

A suitable LIMES configuration is shown in the Listing 1. Lines 3-9 and 10-15 define the source and target knowledge bases by means of SPARQL endpoints. The `pagesize` tag contains the amount of data to be simultaneously retrieved, for each of the SPARQL endpoints. Each endpoint has an identifier and an endpoint URL (lines 4 and 11). The `var` and `restriction` tags in lines 5-7 and 12-14 define a selection of data items to be linked in each of the knowledge bases. The `property` tags in lines 8 and 14 determine which property of the data items should be used for matching. Additionally, a set of functions for object modifications are configurable

Table 3. Results of the link discovery between dados.gov.br, DBpedia and GeoNames.

Spatial concept	dados.gov.br Resources	Links to DBpedia	Links to GeoNames
Country	1	1	1
States	28	20	26
Municipalities	5320	545	3044

within the `property` tag. In this example, literal values will be set to lower case and language information will be ignored. The `metric` tag in line 16 defines which metric should be used to measure the similarity between data items. Finally, the acceptance and review statements (in lines 17-18) define what type of link should be generated for which threshold intervals.

Table 3 summarizes the results of the link generation process between dados.gov.br, DBpedia and GeoNames. While the link discovery results in a relatively high number of matched states, only few (i.e. 10%) of the municipalities could be matched to suitable DBpedia resources. We attribute this to the low coverage (and possibly non-standard spelling) of Brazilian municipalities in Wikipedia articles.

5. Visualizing Statistical Data

In order to hide the complexity of the RDF Data Cube vocabulary from users and to facilitate the browsing and exploration of cubes we developed the RDF Data Cube visualization component *CubeViz*^e as an OntoWiki extension.

As a starting point for using CubeViz, the desired data structure (`qb:DataSetDefinition`) and the dataset (`qb:DataSet`) have to be selected followed by the selection of aggregated components, as depicted in Figure 4. These components are defined as instances of type `qb:ComponentSpecification` which references different types of component properties `qb:DimensionProperty`, `qb:MeasurementProperty` and `qb:AttributeProperty`. To prepare the rendering of the selection form, CubeViz processes and analyses the cube employing a set of SPARQL queries to obtain all necessary structural information. In order to improve the performance of the analysis, CubeViz acts at the structural level of the cube and not at the observation level. In the case of a cube that does not have such an explicit structure, CubeViz is able to extract a generic one based on implicit definitions.

As a result of such a selection, a SPARQL query is created for retrieving all matching observations. Further configurations adjustable in CubeViz act on the visualization level. For instance, users (or domain experts) are able to select different types of graphs such as bar charts, pie charts, spline charts and scatter plots

^e<https://github.com/AKSW/cubeviz.ontowiki>

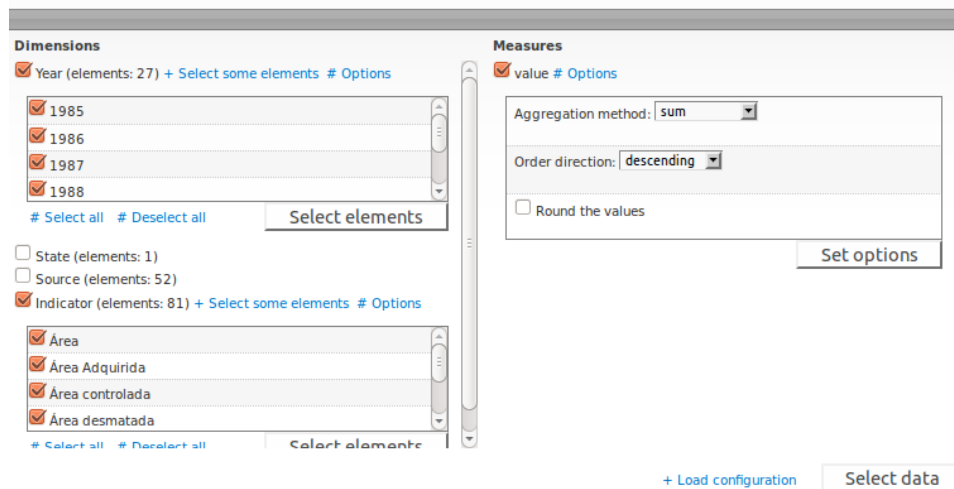


Fig. 4. CubeViz: Selection of structural resources.

depending on the selected amount of dimensions. To render visualized charts on the client (i.e. browser) side, we used the JavaScript library Highcharts^f, a result of which is depicted in Figure 5. Alternatively, PHPPlot^g can be used for server-side rendering of charts.

CubeViz contains a small number of methods to operate on the resulting observation measurements such as the aggregation methods SUM, AVG, MIN and MAX. The extensible architecture of CubeViz in combination with the OntoWiki extension system allows multiple enhancements to integrate further filter functions, mathematical operations as well as the integration of additional chart rendering libraries, chart types and their respective configuration.

6. 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

^f<http://www.highcharts.com/>

^g<http://sourceforge.net/projects/phpplot/>

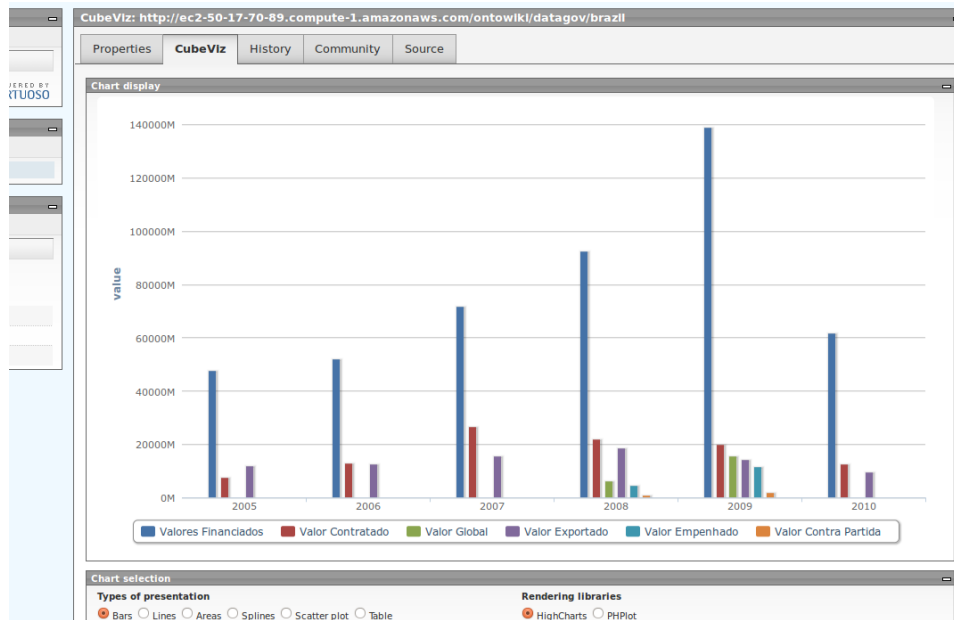


Fig. 5. Visualization of statistics of dados.gov.br model.

the general public in 2010^h.

In September 2011 Brazil became a member of the Open Government Partnershipⁱ, 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^j.

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^k.

As a result of our publishing effort (employing the techniques described in the previous sections), we obtained an anatomy of 10 years of life in Brazil reaching

^h<https://i3gov.planejamento.gov.br/>

ⁱ<http://www.opengovpartnership.org/>

^j<http://www.dados.gov.br/>

^k<https://i3gov.planejamento.gov.br/>

Table 4. Results Statistics (<http://purl.org/GovDataCube>).

Criterion	Measurement
Base data	
Data size	1GB
Data entries	4,514,612
Conversion Process	
Triples	31,120,766
Conversion Time	≈ 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

in some cases 30 years back in time. Figure 6 shows the dados.gov.br OLAP data model and Table 4 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 into more than 900 separate datasets.

7. Related Work

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

Triplification: 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

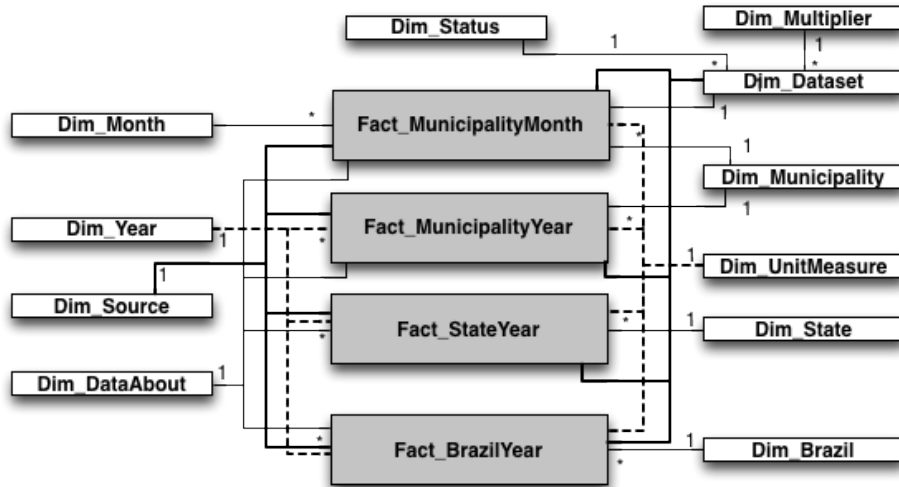


Fig. 6. The Dados.gov.br OLAP data model.

Triplify [5] over standalone solutions such as *D2R* [9] up to integrated tools such as *Virtuoso RDF Views* [14]. 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 [16], which explores the opposite direction to our approach, i.e., the transformation of statistical Linked Data for use in OLAP systems.

Statistical Data publishing: Statistical Data and Metadata eXchange (SDMX, [2]) 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 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 [1] and the Organisation for Economic Co-operation and Development [3].

The representation of statistics in RDF started with SCOVO [15, 10] and continued with the successor RDF Data Cube Vocabulary [4]. The Data Cube Vocabulary is closely aligned with SDMX [10]. Examples of statistics published as RDF adhering to the Data Cube vocabulary and visualized for human consumption include the

EC's INFSO Digital Agenda Scoreboard¹ and the LOD2 Open Government Data stakeholder survey [18].

Linked Governmental Data: Several governments started to publish governmental data on the Web. Tim-Berners Lee discussed a set of Design Issues [8] on how to publish governmental information in a re-usable way. One of the first Linked Data providers publishing governmental data was the UK Government (<http://data.gov.uk/>), hosting information about different governmental sectors of Great Britain including transport, legislation and finance [21]. A further provider of governmental data is the US Government (<http://data.gov/>). Due to the fact that this information was not made available as Linked Data, external groups started to transform and publish the information according the Linked Data principles [11]. Recent research work also aims to facilitate government data ecosystems through specialized portals [12] and distributed dataset catalogs [13]. Another important issue, which is particularly tackled by this paper for the statistics domain, is enabling interoperability of government data catalogs [17].

8. Conclusions, Lessons Learned and Future Work

We first presented two complementary approaches for extracting and publishing statistical data on the Data Web. Then, we discussed an efficient linking strategy and visualization tool. Finally, we presented a large-scale use case of statistic data published in Brazil.

The original data.gov.br OLAP database revealed quite a number of problems, the most crucial of which are discussed in what follows. First, we spent a significant amount of effort for *pre-processing* (i.e. reorganizing) the database into a star shaped OLAP. Second, *Encoding* – although standard – was the source of several problems. Expressions such as "São Paulo" were interpreted incorrectly ("S/u00e3o Paulo") which caused problems in the serialization; We applied a filter, to transform the whole content to UTF8, before feeding it to the tool chain (cf. [7]); Third, it turned out to be crucial to include as much background knowledge from the base data as possible to avoid *ambiguities* during the URI generation and linking process. For example, there are a large number of municipalities in Brazil with the same name (for example, twelve are called 'Vista Alegre'). By incorporating the type of municipality and the state they are located in, we were able to reduce this ambiguity substantially and create more qualitative links. Fourth, we were surprised by the relatively reasonable overall *processing time*. It took less than three days to process over nine hundred datasets, whose data spans nearly 30 years of government (amounting 30 million triples). In fact, pure RDF triple stores offered decent performance. However, the performance of a whole RDF processing tool chain is usually determined by its weakest element and thus performance is still an issue in most cases.

¹http://ec.europa.eu/information_society/digital-agenda/scoreboard/

We see the work described in this article as 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 may focus 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.

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, by Globo.com and by a grant from the European Union's 7th Framework Programme provided for the project LOD2 (GA no. 257943).

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] The rdf data cube vocabulary. Technical report, 2010.
- [5] S. Auer, S. Dietzold, J. Lehmann, S. Hellmann, and D. Aumueller. Triplify: Light-weight linked data publication from relational databases. In *WWW*. ACM, 2009.
- [6] S. Auer, S. Dietzold, and T. Riechert. OntoWiki - A Tool for Social, Semantic Collaboration. In *ISWC*, volume 4273 of *LNCS*. Springer, 2006.
- [7] S. Auer, M. Weidl, J. Lehmann, A. J. Zaveri, and K.-S. Choi. I18n of semantic web applications. In *ISWC2010*, LNCS. Springer, 2010.
- [8] T. Berners-Lee. Putting Government Data online. W3C Design Issue, 2009. <http://www.w3.org/DesignIssues/GovData.html>.
- [9] C. Bizer and R. Cyganiak. D2r server - publishing relational databases on the semantic web. Poster at ISWC, 2006.
- [10] R. Cyganiak, S. Field, A. Gregory, W. Halb, and J. Tennison. Semantic statistics: Bringing together sdmx and scovo. In *LDOW*, volume 628 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010.
- [11] L. Ding, D. DiFranzo, A. Graves, J. Michaelis, X. Li, D. L. McGuinness, and J. Hendler. Twc data-gov corpus: incrementally generating linked government data from data.gov. In *WWW*, pages 1383–1386. ACM, 2010.
- [12] L. Ding, T. Lebo, J. S. Erickson, D. DiFranzo, G. T. Williams, X. Li, J. Michaelis, A. Graves, J. G. Zheng, Z. Shangguan, J. Flores, D. L. McGuinness, and J. Hendler. Twc logd: A portal for linked open government data ecosystems. *J. of Web Semantics*, 2011.
- [13] J. S. Erickson, E. Rozell, Y. Shi, J. Zheng, L. Ding, and J. A. Hendler. Twc interna-

- tional open government dataset catalog. In *Proceedings of the 7th ICSS, I-Semantics '11*. ACM, 2011.
- [14] O. Erling. Automated Generation of RDF Views over Relational Data Sources with Virtuoso, 2009.
 - [15] M. Hausenblas, W. Halb, Y. Raimond, L. Feigenbaum, and D. Ayers. Scovo: Using statistics on the web of data. In *ESWC*, volume 5554 of *LNCS*. Springer, 2009.
 - [16] B. Kämpgen and A. Harth. Transforming statistical linked data for use in olap systems. In *I-SEMANTICS 2011*, 2011.
 - [17] F. Maali, R. Cyganiak, and V. Peristeras. Enabling interoperability of government data catalogues. In *Proc. of the 9th IFIP, EGOV'10*, 2010.
 - [18] M. Martin, M. Kaltenböck, H. Nagy, and S. Auer. The open government data stakeholder survey. In *OKCon*. OKFN, 2011.
 - [19] A. Ngonga Ngomo and S. Auer. Limes - a time-efficient approach for large-scale link discovery on the web of data. In *Proc. of IJCAI*, 2011.
 - [20] N. Noy and A. Rector. Defining N-ary Relations on the Semantic Web. Technical report, W3C, 2006.
 - [21] J. Sheridan and J. Tennison. Linking uk government data. In *WWW2010 Workshop on Linked Data on the Web (LDOW)*, 2010.
 - [22] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and maintaining links on the web of data. In *ISWC*, volume 5823 of *LNCS*, 2009.