

A Mediator for Statistical Linked Data

Lívia Ruback
Departamento de Informática
PUC-Rio
+55-21-3527-1500
lrodrigues@inf.puc-rio.br

Sofia Manso
Departamento de Informática
PUC-Rio
+55-21-3527-1500
ssilva@inf.puc-rio.br

Percy E. Rivera Salas
Departamento de Informática
PUC-Rio
+55-21-3527-1500
psalas@inf.puc-rio.br

Marcia Pesce
Departamento de Informática
PUC-Rio
+55-21-3527-1500
mpesce@inf.puc-rio.br

Sérgio Ortiga
Departamento de Informática
PUC-Rio
+55-21-3527-1500
sortiga@inf.puc-rio.br

Marco A. Casanova
Departamento de Informática
PUC-Rio
+55-21-3527-1500
casanova@inf.puc-rio.br

ABSTRACT

This paper introduces a mediation architecture to help describing and consuming statistical data, exposed as RDF triples, but stored in relational databases. The architecture features a catalogue of *linked data cube descriptions*, created according to the Linked Data principles. The catalogue has a standardized description for each data cube actually stored in each statistical (relational) database known to the mediation environment. The mediator offers an interface to browse the linked data cube descriptions and exports the data cubes as RDF triples, generated on demand from the underlying databases.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications – *Statistical Databases*.

General Terms

Design.

Keywords

Statistical Data, Linked Data, Mediation Architecture.

1. INTRODUCTION

Currently, statistical data are mostly stored in relational databases and available through interfaces designed for humans, which are inappropriate for software agents. The raw data are processed, validated and stored in tables, which are aggregated, ensuring the confidentiality of individuals and entities [4]. Applications dealing with statistical data usually include Online Analytical Processing (OLAP), a set of tools and algorithms for querying large statistical databases. In OLAP, data are perceived as

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SAC'13, March 18-22, 2013, Coimbra, Portugal.

Copyright 2013 ACM 978-1-4503-1656-9/13/03...\$15.00.

multidimensional structures known as *data cubes* [6], which represent a star schema view of the relational database.

The Linked Data principles [8] can be profitably applied to statistical data, in the sense that the principles offer a strategy to provide the missing semantics of the data. If followed, the Linked Data principles will connect statistical data with related data sources, creating an interconnected data space that enables a rich analysis of the data [4].

This paper introduces a mediation architecture to help describing and consuming statistical data, exposed as RDF triples, but stored in relational databases. The architecture features a catalogue of *linked data cube descriptions*, created according to the Linked Data principles. The mediator offers an interface to browse the linked data cube descriptions and exports the data cubes as RDF triples, generated on demand from the underlying data sources.

The remainder of the paper is organized as follows. Section 2 discusses the representation of data cubes in RDF. Section 3 describes the mediation architecture. Section 4 presents related work. Finally, Section 5 contains the conclusions. and presents directions for future work.

2. DATA CUBES IN RDF

2.1 Formal Definition of Data Cubes

Following Cyganiak et al [13], 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) 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 of the form

$$R \subseteq D_1 \times \dots \times D_m \times A_1 \times \dots \times A_n \times O$$

commonly referred to as a *data cube* or simply as a *cube*.

In fact, a tuple of values from the dimension domains identifies a single observation measure value and the associated attribute values. Hence, R is actually a function of the form

$$R: D_1 \times \dots \times D_m \rightarrow A_1 \times \dots \times A_n \times O$$

2.2 Data Cube Description in RDF

The RDF Data Cube vocabulary [13] was specifically designed to publish data cubes on the Web in such a way that it can be linked to related RDF datasets. The model that underlies the Data Cube vocabulary is compatible with the cube model that is supported by SDMX, an ISO standard for exchanging and sharing statistical data and metadata among organizations.

Very briefly, the RDF Data Cube vocabulary 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`.

2.3 Triplification of Data Cubes

There is not a single strategy to *triplify* a data cube, that is, to create a set of RDF triples that represent the cube. We adopt what we call the *complete triplification strategy*, defined as follows.

Let $R \subseteq D_1 \times \dots \times D_m \times A_1 \times \dots \times A_n \times O$ be a relation that represents a data cube, as in Section 2.1.

The complete triplification strategy represents R by reification, 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 $(u, d_1, x_1), \dots, (u, d_m, x_m), \dots, (u, a_1, y_1), \dots, (u, a_n, y_n), (u, o, z)$.

Figures 1 and 2 present an example of triplification. Each dimension table is defined as an instance of `qb:DimensionProperty`, so that each tuple in them is an instance of the new dimension, as presented in Figure 1.

Tuples that result from the SQL query are instances of `qb:Observation`. Figure 2 shows triples that capture an observation. Note that the third line is “`dim:Time times:T2010`” and thereby corresponds to the instance of the dimension `dim:Time` defined in Figure 2.

```

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

```

Figure 1. Dimensions definition and individuals.

```

1 # New dimension definition
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" .

```

Figure 2. Instance of an observation.

3. THE MEDIATION ARCHITECTURE

Figure 3 summarizes the major components of the mediation architecture.

A *Wrapper* for an underlying relational database provides star-shaped view schemas describing statistical data stored in the database. We stress that the data cubes may be organized in the underlying database in any way, using several tables. However, the wrapper exposes each data cube through a single star-shaped schema, whose mapping to the underlying tables is internal to the wrapper.

The *Catalogue* contains *public* and *private* data. Public data refers to the data cube descriptions (again, including their dimensions and dimension values) that are exposed to the applications. A data cube description is stored as a set of RDF triples, called a *linked data cube description*.

We stress that a linked data cube description contains triples describing the dimensions and attributes of a data cube, perhaps including dimension domain values. However, a linked data cube description does not contain triples that capture the observations, i.e, it is not a complete materialization of a data cube in RDF; the data cube still remains in the relational database.

The catalogue also includes public RDF *sameAs* triples that relate resources in linked data cube descriptions with resources located in external data sets. For example, if there is a dimension resource representing the City of Rio de Janeiro, there will be a *sameAs* triple relating this resource with the DBpedia entry for the City of Rio de Janeiro, Brazil.

Private data refers to the information required internally. For each linked data cube description in the catalogue, there is at least one mapping to a star-shaped view schema of an underlying database, which is used to retrieve the observations (of the data cube). Similar mappings are required to retrieve the dimension values.

By assumption, each linked data cube description corresponds to one or more star-shaped schemas, in the sense that the data cube may be redundantly stored in different databases or even in the same database. The mediator is free to choose any one of the star-shaped schemas to materialize the data cube.

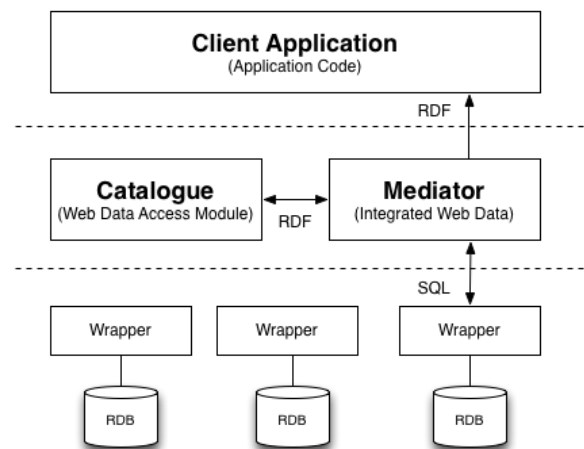


Figure 3. Overview of the Architecture.

The *Mediator* mediates access to the underlying statistical relational databases and exposes catalogue data to the applications. It allows an application to select a linked data cube description, stored in the catalogue, and to apply certain transformations to the cube. It converts the data (i.e., the cube) returned by a wrapper to RDF, passing the triples to the application that submitted the request.

A *Client Application* is any application that interacts with the mediator to access the catalogue and the underlying databases.

A client application had already been developed, RdXel [16], which browses the catalogue with the keyword given by the user, requests and displays the observations of the selected cube to the mediator.

4. RELATED WORK

Most triplification tools generate RDF from relational databases [1][2][5]. One of the few works that explores the transformation of statistical Linked Data for use in OLAP systems is [9].

The *Statistical Data and Metadata eXchange* (SDMX) [12] is an initiative started in 2001 to foster standards for the exchange of statistical information. Experiences and best practices regarding the publication of statistics on the Web in SDMX have been published by the UN [14] and the OECD [11].

The representation of statistics in RDF started with SCOVO [7], [3] and continued with its successor, the RDF Data Cube Vocabulary [13]. The Data Cube Vocabulary is closely aligned with SDMX [16]. Examples of statistics published as RDF adhering to the Data Cube vocabulary the LOD2 Open Government Data stakeholder survey [10].

OLAP2DataCube [15] is an Ontowiki plug-in for statistical data publishing for extracting and publishing statistical data on the Web. Our approach was based on the experience obtained during the development of OLAP2DataCube.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we presented an approach for extracting and publishing statistical data on the Web of Data through a mediator. The approach combines RDF data cubes descriptions, stored in a central catalogue, with the dynamic generation of RDF triples that describe the data cube observations. The approach avoids storing the complete triplification of all data cubes redundantly, which might be infeasible. It also allows data cube descriptions to be linked and combined with related information on the Web.

As future work, we intend to implement additional client applications, such as a mashup data cube utility that combines two or more data cubes from heterogeneous data sources. We also plan to implement a mediator federation that connects several data cube catalogues seamlessly.

6. ACKNOWLEDGMENTS

This work was partly supported by CNPq, under grants 301497/2006-0, 475717/2011-2 and by FAPERJ under grant E-26/103.070/2011.

REFERENCES

- [1] S. Auer, S. Dietzold, J. Lehmann, S. Hellmann, and D. Aumueller. Triplify: light-weight linked data publication from relational databases. In *Proc. 18th Int'l. Conf. on World Wide Web*, pp. 621–630. ACM, 2009.
- [2] C. Bizer and R. Cyganiak. D2r server - publishing relational databases on the semantic web. *5th Int'l. Semantic Web Conf.*, p. 26, 2006.
- [3] R. Cyganiak, S. Field, A. Gregory, W. Halb, and J. Tennison. Semantic statistics: Bringing together sdmx and scovo. In C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas, ed., *LDOW*, vol. 628 of *CEUR Workshop Proc.*, 2010.
- [4] R. Cyganiak, M. Hausenblas, and E. McCuire. Official Statistics and the Practice of Data Fidelity. In D. Wood, ed., *Linking Government Data*, pp. 135–151. Springer, 2011.
- [5] O. Erling and I. Mikhailov. Rdf support in the virtuoso dbms. *Networked Knowledge-Networked Media*, pp. 7–24, 2009.
- [6] L. Etcheverry and A. A. Vaisman. Enhancing olap analysis with web cubes. In *Proc. ESWC 2012*, pp. 469–483, Springer, 2012.
- [7] M. Hausenblas, W. Halb, Y. Raimond, L. Feigenbaum, and D. Ayers. Scovo: Using statistics on the web of data. In *Proc. ESWC 2009 Heraklion*, pp. 708–722. Springer, 2009.
- [8] T. Heath and C. Bizer. Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1(1):1–136, 2011.
- [9] B. Kämpgen and A. Harth. Transforming statistical linked data for use in olap systems. In *Proc. 7th Int'l. Conf. on Semantic Systems*, pp. 33–40, ACM, 2011.
- [10] M. Martin, M. Kaltenböck, H. Nagy, and S. Auer. The open government data stakeholder survey. *OKCon. OKFN*, 2011.
- [11] Management of statistical metadata at the OECD. <http://www.oecd.org/dataoecd/26/33/33869551.pdf>
- [12] Statistical data and metadata exchange (sdmx). <http://sdmx.org/>.
- [13] The rdf data cube vocabulary w3c working draft. <http://www.w3.org/TR/vocab-data-cube/>.
- [14] U. N. S. Commission, U. N. E. C. for Europe, and C. of European Statisticians. *Guidelines for statistical metadata on the Internet*. Statistical standards and studies. United Nations, 2000.
- [15] P. E. Salas, M. Martin, F. Maia Da Mota, K. Breitman, S. Auer, and M. A. Casanova. Publishing Statistical Data on the Web. *Proc. IEEE ICSC 2012 (Sept. 19th- 21st, 2012)*. Palermo, Italy.
- [16] M. Pesce. RdXel: um conjunto de ferramentas para a manipulação de dados estatísticos em RDF por meio de planilhas. 2012. Master's Thesis, Pontifical Catholic University of Rio de Janeiro, Brazil.