

Lívia Couto Ruback Rodrigues

LDC Mediator: A Mediator for Linked Data Cubes

Dissertação de Mestrado

Dissertation presented to the Programa de Pós-Graduação em Informática of the Departamento de Informática da PUC-Rio as partial fulfillment of the requirements for the degree of Mestre.

Advisor: Prof. Marco Antonio Casanova

Rio de Janeiro
September 2013

Lívia Couto Ruback Rodrigues

LDC Mediator: A Mediator for Linked Data Cubes

Dissertation presented to the Programa de Pós-Graduação em Informática of the Departamento de Informática do Centro Técnico Científico da PUC-Rio, as partial fulfillment of the requirements for the degree of Mestre.

Prof. Marco Antonio Casanova

Advisor

Departamento de Informática – PUC-Rio

Prof. Antonio Luz Furtado

Departamento de Informática – PUC-Rio

Prof. Luiz André Portes Paes Leme

Departamento de Informática – UFF

Profa. Giseli Rabello Lopes

Departamento de Informática - PUC-Rio

Prof. José Eugenio Leal

Coordenador Setorial do Centro

Técnico Científico – PUC-Rio

Rio de Janeiro, September 12th, 2013

All rights reserved.

Lívia Couto Ruback Rodrigues

Graduated in Computer Science from Universidade Federal de Juiz de Fora (UFJF), Minas Gerais - Brazil in 2009. She joined the Master in Informatics at Pontifical Catholic University of Rio de Janeiro (PUC-Rio) in 2011. Her main research interests cover Web Semantic and Linked Data.

Bibliographic data

Couto Ruback Rodrigues, Lívia

LDC Mediator: a Mediator for Linked Data Cubes / Lívia Couto Ruback Rodrigues; advisor: Marco Antonio Casanova. – 2013.

69 f. : il. (color) ; 30 cm

Dissertação (Mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2013.

Inclui bibliografia

1. Informática – Teses. 2. Dados Estatísticos. 3. Dados ligados. 4. Arquitetura de mediação. 5. Triplificação. 6. RDF. 7. Cubo de Dados OLAP. 8. REST. I. Casanova, Marco Antonio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Acknowledgments

I would like to say a special thank you to Marco Antonio Casanova, the best advisor I could ever ask for. His wisdom, knowledge, refined sense of humor and kindness were essential to motivate me to keep going and finish this work.

To Profa. Simone for her classes and for the best brownies recipe.

To PUC-Rio for giving me the opportunity to make such important lifelong friends.

To my family and friends. In particular to my parents Rubens and Izabel for their support raising me. To my sisters Flávia and Bianca for all the wonderful memories.

To my boyfriend and partner Adriano, for still being there no matter what. Even taking up dance classes to dance with me.

To CAPES for funding my research.

Abstract

Ruback, Livia; Casanova, Marco Antonio (Advisor). **LDC Mediator: A Mediator for Linked Data Cubes**. Rio de Janeiro, 2013. 69p. MSc. Dissertation – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A statistical data set comprises a collection of observations made at some points across a logical space and is often organized as what is called a *data cube*. The proper definition of the data cubes, especially of their dimensions, helps to process the observations and, more importantly, helps to combine observations from different data cubes. In this context, the Linked Data Principles can be profitably applied to the definition of data cubes, in the sense that the principles offer a strategy to provide the missing semantics of the dimensions, including their values. This work introduces a mediation architecture to help consume linked data cubes, exposed as RDF triples, but stored in relational databases. The data cubes are described in a catalogue using standardized vocabularies and are accessed by HTTP methods using REST principles. Therefore, this work aims at taking advantage of both Linked Data and REST principles in order to describe and consume linked data cubes in a simple but efficient way.

Keywords

Statistical Data, Linked Data, Mediation Architecture, Triplification, RDF, OLAP Data Cube, REST.

Resumo

Ruback, Livia; Casanova, Marco Antonio. **Mediador LDC: Um mediador de Cubos de Dados Interligados**. Rio de Janeiro, 2013. 69p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Um banco de dados estatístico consiste de um conjunto de observações feitas em pontos de um espaço lógico, e, muitas vezes, são organizados como cubos de dados. A definição adequada de cubos de dados, em especial de suas dimensões, ajuda a processar as suas observações e, mais importante, ajuda a combinar observações de cubos de dados diferentes. Neste contexto, os princípios de dados interligados podem ser proveitosamente aplicados à definição de cubos de dados, oferecendo uma estratégia para fornecer a semântica das dimensões, incluindo seus valores. Este trabalho introduz uma arquitetura de mediação para auxiliar no consumo de cubos de dados, expostos como triplas RDF e armazenados em bancos de dados relacionais. Os cubos de dados são descritos em um catálogo usando vocabulários padronizados e são acessados por métodos HTTP usando os princípios de REST. Portanto, este trabalho busca tirar proveito tanto dos princípios de dados interligados quanto dos princípios de REST para descrever e consumir os cubos de dados interligados de forma simples e eficiente.

Palavras-chave

Dados estatísticos, dados ligados, arquitetura de mediação, triplificação, RDF, Cubo de Dados OLAP, REST.

Table of Contents

1	Introduction	12
1.1	Motivation	12
1.2	Goal and Contributions	13
1.3	Dissertation Structure	14
2	Background	15
2.1	The World Wide Web architecture	15
2.2	The HTTP Protocol	16
2.3	The Linked Data concept	19
2.3.1.	The Linked Data Principles	19
2.3.2.	The Linking Open Data Project	20
2.4	RDF	21
2.5	SPARQL Query Language	23
2.6	Data Cubes	24
2.6.1.	OLAP Data Cube	24
2.6.2.	RDF Data Cube Vocabulary	25
2.7	RDF to RDB approaches	28
2.7.1.	Direct Mapping	29
2.7.2.	R2RML Vocabulary	29
2.8	The Web Service Architecture	31
2.8.1.	XML/SOAP/WSDL	32
2.8.2.	REST	34
2.8.2.1.	REST architecture	34
2.8.2.2.	RESTful Web services	36
2.9	Summary	36
3	Related Work	38
3.1	OLAP data approaches	38
3.2	RESTful Linked Data approaches	40

3.3	Summary	42
4	The OLAP2DataCube Catalog On Demand Framework	43
4.1	Overview of the OLAP2DataCube Catalog On Demand	43
4.2	Data cube consumption	46
4.2.1.	Stage 1: Selection of a Data Cube	46
4.2.2.	Stage 2: Requesting Data Cube Metadata	46
4.2.3.	Stage 3: Triplification of a Data Cube	46
4.3	Summary	47
5	The LDC Mediator	48
5.1	Combining REST and Linked Data	48
5.2	The LDC Mediator components	49
5.3	Implementation of the LDC Mediator	50
5.3.1.	The Mediator Engine package	51
5.3.2.	The Metadata Service package	53
5.3.3.	The RDB-to-RDF package	54
5.4	Consuming data cubes with the LDC Mediator	55
5.4.1.	Datasets collected	55
5.4.2.	The RESTful Web service definition	57
5.4.3.	Requesting a cube metadata	57
5.4.4.	Requesting cube observations	60
5.5	Summary	62
6	Conclusion	64
6.1	Contributions	64
6.2	Limitations and Future work	64
7	Bibliography	66

List of Figures

Figure 1: The three architectural bases of the Web	16
Figure 2: A HTTP request method example	17
Figure 3: The LOD Cloud Diagram at September 2011	21
Figure 4: An RDF graph example	22
Figure 5: A simple SPARQL query	23
Figure 6: The RDF Data Cube Vocabulary key terms	26
Figure 7: A Data cube example	26
Figure 8: Dimension properties	27
Figure 9: A measure property	27
Figure 10: A data cube description (the dataset definition part)	28
Figure 11: A data cube description (the data structure part)	28
Figure 12: Direct Mapping example	29
Figure 13: Overview of the R2RML Vocabulary	30
Figure 14: An R2RML mapping example	31
Figure 15: Web Services Architecture Stack	32
Figure 16: A SOAP request and response example	33
Figure 17: A WSDL structure	34
Figure 18: REST Data Elements	35
Figure 19: The QB4OLAP Vocabulary	39
Figure 20: Overview of the OLAP2DataCube Catalog On Demand	44
Figure 21: LDC Mediator components	49
Figure 22: Mediator Engine package	51
Figure 23: The Jersey implementation of the Data Cube resource	52
Figure 24: Metadata Service package	53
Figure 25: The generic method to execute a SPARQL	53
Figure 26: RDB-to-RDF Converter package	54
Figure 27: A residents star schema	56
Figure 28: List of cubes	57
Figure 29: Return of the cube metadata request	58

Figure 30: SPARQL Query to return a metadata cube subset	59
Figure 31: Searching dynamically triples related to a resource	59
Figure 32: R2RML mapping file to the residents cube	61
Figure 33: An observation generated	62

List of Tables

Table 1: The main HTTP methods _____	17
Table 2: Some of the HTTP status codes _____	18
Table 3: Comparing the related OLAP approaches _____	40
Table 4: The JAX-RS most important annotations _____	52
Table 5: DB2Triples parameters _____	54
Table 6: LDC RESTful API methods _____	57