Pontifícia Universidade Católica
do Rio de Janeiro

Sofia Ribeiro Manso de Abreu e Silva

Catalogue of Linked Data Cube Descriptions

Dissertação de Mestrado

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

Advisor: Prof. Marco Antonio Casanova

Rio de Janeiro
June 2013

Pontifícia Universidade Católica
do Rio de Janeiro

Sofia Ribeiro Manso de Abreu e Silva

Catalogue of Linked Data Cube Descriptions

Dissertation presented to the Programa de Pós-Graduação em Informática do Departamento de Informática do Centro Técnico e Científico da PUC-Rio, as partial fullfilment 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
UFF

Prof. José Eugenio Leal
Coordinator of the Centro
Técnico Científico — PUC–Rio

Rio de Janeiro, June 28^{th}, 2013

Sofia Ribeiro Manso de Abreu e Silva

Graduated in Applied Mathematics and Computation from the University of Aveiro (Portugal) in 2007. She joined the Master in Informatics at Pontifical Catholic University of Rio de Janeiro (PUC-Rio) in 2011. Her current research area is related to Statistical Linked Data and Semantic Web.

Bibliographic data

Ribeiro Manso de Abreu e Silva, Sofia

Catalogue of Linked Data Cube Descriptions / Sofia Ribeiro Manso de Abreu e Silva; advisor: Marco Antonio Casanova. — 2013.

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

1. Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2013.

Inclui bibliografia.

1. Informática – Teses. 2. Dados Estatísticos. 3. Linked Data. 4. Arquitetura de Mediação. 5. Triplificação. 6. RDF. 7. Data Cube Vocabulary. 8. R2RML. I. Casanova, Marco Antonio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Acknowledgements

First, I would like to express my gratitude to my advisor Marco Antonio Casanova for the useful comments, remarks and engagement through my learning process in this master. I am very thankful for his sharing of knowledge, support, motivation and trust during all my work. His encouragement and guidance was very important during this dissertation.

I would like to acknowledge the professors and colleagues, for the teachings, encouragement, help and friendship.

I also would like to thank the employees of the Informatics Department of PUC-Rio for their help, kindness and support.

To CAPES for the financial aid granted.

To my family, my sincere thanks. In particular, I would like to thank to my aunt Isabel and to my uncle António. Without them this master would not have been possible.

A big thanks to my friends for always being there for me.

I would like to make a special reference to my boyfriend Steve for his love and patience during these difficult times.

Last but not least, my deepest gratitude to my parents, Alice and João, for their unconditional love. I want to thank them for always believing in me, for giving me the strength to carry on and to make this happen. I will be grateful forever for their love.

Abstract

Ribeiro Manso de Abreu e Silva, Sofia ; Casanova, Marco Antonio (Advisor). **Catalogue of Linked Data Cube Descriptions**. Rio de Janeiro, 2013. 85p. MSc. Dissertation — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Statistical Data are considered one of the major sources of information and are essential in many fields as they can work as social and economic indicators. A statistical data set comprises a collection of observations made at some points of 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 processing the observations and, more importantly, helps combining 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 dissertation first describes a mediation architecture to help describing and consuming statistical data, exposed as RDF triples, but stored in relational databases. One of the features of this architecture is the Catalogue of Linked Data Cube Descriptions, which is described in detail in the dissertation. This catalogue has a standardized description in RDF of each data cube actually stored in statistical (relational) databases. Therefore, the main discussion in this dissertation is how to represent the data cubes in RDF, i.e., how to map the database concepts to RDF in a way that makes it easy to query, analyze and reuse statistical data in the RDF format.

Keywords

Statistical Data; Linked Data; Mediation Architecture; Triplification; RDF; Data Cube Vocabulary; R2RML.

Resumo

Ribeiro Manso de Abreu e Silva, Sofia ; Casanova, Marco Antonio. **Catálogo de Descrições de Cubos de Dados Interligados**. Rio de Janeiro, 2013. 85p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Dados estatísticos são considerados uma das principais fontes de informação e são essenciais em muitos campos, uma vez que podem funcionar como indicadores sociais e económicos. Um conjunto de dados estatísticos compreende um conjunto de observações feitas em determinados pontos de um espaço lógico e é muitas vezes organizado como o que se chama de cubo de dados. A definição correta dos cubos de dados, especialmente das suas dimensões, ajuda a processar as observações e, mais importante, ajuda a combinar as observações de diferentes cubos de dados. Neste contexto, os princípios de Linked Data podem ser proveitosamente aplicados à definição de cubos de dados, no sentido de que os princípios oferecem uma estratégia para proporcionar a semântica ausentes das suas dimensões, incluindo os seus valores.

Esta dissertação descreve inicialmente uma arquitetura de mediação para ajudar a descrever e consumir dados estatísticos, expostos como triplas RDF, mas armazenados em bancos de dados relacionais. Uma das características desta mediação é o Catálogo de Descrições de Cubos de Dados Interligados, que vai ser descrito em detalhes na dissertação. Este catálogo contém uma descrição padronizada em RDF para cada cubo de dados, que está realmente armazenado em cada banco de dados (relacional). Portanto, a principal discussão nesta dissertação é sobre a forma de representar em RDF cubos representando dados estatísticos e armazenados em bancos de dados relacionais, ou seja, como mapear os conceitos de banco de dados para RDF de uma forma em que seja fácil consultar, analisar e reutilizar dados estatísticos no formato RDF.

Palavras–chave

Dados Estatísticos; Linked Data; Arquitetura de Mediação; Triplificação; RDF; Data Cube Vocabulary; R2RML.

Table of Contents

List of Figures

List of Tables