

Percy Enrique Rivera Salas

StdTrip: An a priori design
process for publishing Linked
Data

Dissertação de Mestrado

DEPARTAMENTO DE INFORMÁTICA
Postgraduate Program in Informatics

Rio de Janeiro
April 2011



Percy Enrique Rivera Salas

**StdTrip: An a priori design process for
publishing Linked Data**

Dissertação de Mestrado

Dissertation presented to the Postgraduate Program in Informatics of the Departamento de Informática, PUC-Rio as partial fulfillment of the requirements for the degree of Mestre em Informática.

Advisor: Prof. Karin Koogan Breitman

Rio de Janeiro
April 2011



Percy Enrique Rivera Salas

**StdTrip: An a priori design process for
publishing Linked Data**

Dissertation presented to the Postgraduate Program in Informatics of the Departamento de Informática PUC–Rio as partial fulfillment of the requirements for the degree of Mestre em Informática.

Prof. Karin Koogan Breitman

Advisor

Departamento de Informática — PUC–Rio

Prof. Marco Antonio Casanova

Departamento de Informática — PUC–Rio

Prof. José Viterbo Filho

Departamento de Ciência e Tecnologia — UFF

Prof. Antonio Luz Furtado

Departamento de Informática — PUC–Rio

Prof. José Eugenio Leal

Coordinator of the Centro Técnico Científico da PUC–Rio

Rio de Janeiro — April 01, 2011

All rights reserved.

Percy Enrique Rivera Salas

Graduated in Systems Engineering from Universidad Católica de Santa María (UCSM), Arequipa – Peru in 2009. He conducted training courses on Publishing Open Government Data for the W3C Brazil. His current research areas are Semantic Web and Linked Data. He is also a member of the W3C RDB2RDF Working Group.

Bibliographic data

Rivera Salas, Percy Enrique

StdTrip: An a priori design process for publishing Linked Data / Percy Enrique Rivera Salas; advisor: Karin Koogan Breitman – 2011.

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

Dissertação (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2011.

Inclui bibliografia

1. Informática – Teses. 2. Linked Data. 3. Triplification. 4. Alinhamento de Ontologias. 5. Reutilização de Ontologias. 6. Interoperabilidade. I. Koogan Breitman, Karin. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Acknowledgments

I wish to thank, first and foremost, my advisor Professor Dr. Karin Breitman, whose encouragement, guidance and everyday kindness made possible the realisation of this work.

I also would like to make a special reference to Dr. José Viterbo for his friendship, valuable insights and recommendations.

I owe my deepest gratitude to my beloved family, for encouraging me to face the challenges and for giving me the strength to carry on.

I remain indebted to many colleagues and professors at PUC-Rio for providing me the means to learn and understand.

To CAPES, for the financial support, without which this work would not have been possible.

Abstract

Salas, Percy; Koogan Breitman, Karin. **StdTrip: An a priori design process for publishing Linked Data**. Rio de Janeiro, 2011. 74p. MSc Dissertation — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Open Data is a new approach to promote interoperability of data in the Web. It consists in the publication of information produced, archived and distributed by organizations in formats that allow it to be shared, discovered, accessed and easily manipulated by third party consumers. This approach requires the triplification of datasets, i.e., the conversion of database schemata and their instances to a set of RDF triples. A key issue in this process is deciding how to represent database schema concepts in terms of RDF classes and properties. This is done by mapping database concepts to an RDF vocabulary, used as the base for generating the triples. The construction of this vocabulary is extremely important, because the more standards are reused, the easier it will be to interlink the result to other existing datasets. However, tools available today do not support reuse of standard vocabularies in the triplification process, but rather create new vocabularies. In this thesis, we present the StdTrip process that guides users in the triplification process, while promoting the reuse of standard, RDF vocabularies.

Keywords

Linked Data. Triplification. Ontology Matching. Ontology Reuse. Interoperability.

Resumo

Salas, Percy; Koogan Breitman, Karin. **StdTrip: Um processo de projeto a priori para publicação de “Linked Data”**. Rio de Janeiro, 2011. 74p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A abordagem de Dados Abertos tem como objetivo promover a interoperabilidade de dados na Web. Consiste na publicação de informações em formatos que permitam seu compartilhamento, descoberta, manipulação e acesso por parte de usuários e outros aplicativos de software. Essa abordagem requer a triplificação de conjuntos de dados, ou seja, a conversão do esquema de bases de dados relacionais, bem como suas instâncias, em triplas RDF. Uma questão fundamental neste processo é decidir a forma de representar conceitos de esquema de banco de dados em termos de classes e propriedades RDF. Isto é realizado através do mapeamento das entidades e relacionamentos para um ou mais vocabulários RDF, usados como base para a geração das triplas. A construção destes vocabulários é extremamente importante, porque quanto mais padrões são utilizados, melhor o grau de interoperabilidade com outros conjuntos de dados. No entanto, as ferramentas disponíveis atualmente não oferecem suporte adequado ao reuso de vocabulários RDF padrão no processo de triplificação. Neste trabalho, apresentamos o processo StdTrip, que guia usuários no processo de triplificação, promovendo o reuso de vocabulários de forma a assegurar interoperabilidade dentro do espaço da Linked Open Data (LOD).

Palavras-chave

Linked Data. Triplification. Alinhamento de Ontologias. Reutilização de Ontologias. Interoperabilidade.

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 12 |
| 1.1 | Problem Setting | 13 |
| 1.2 | Goal | 13 |
| 1.3 | Contributions | 13 |
| 1.4 | Organization | 14 |
| 2 | Background | 15 |
| 2.1 | Ontology | 16 |
| 2.2 | RDF | 17 |
| 2.3 | RDF Standard Vocabularies | 20 |
| 2.4 | Vocabulary Reuse | 22 |
| 2.5 | Publishing New Vocabularies | 23 |
| 2.6 | Linked Data | 24 |
| 2.7 | Mapping Relational Databases to RDF (RDB-to-RDF) | 25 |
| 2.8 | Ontology Matching | 28 |
| 2.8.1 | Syntactic Approach | 28 |
| 2.8.2 | Semantic Approach | 29 |
| 2.8.3 | <i>A priori</i> approach | 30 |
| 3 | Related Work | 31 |
| 3.1 | Triplify | 31 |
| 3.2 | D2RQ | 32 |
| 3.3 | Virtuoso RDF View | 32 |
| 3.4 | DB2OWL | 32 |
| 3.5 | RDBtoOnto | 33 |
| 3.6 | Ultrawrap | 33 |
| 3.7 | Automated Mapping Generation for Converting Databases into Linked Data | 33 |
| 4 | The StdTrip Process | 35 |
| 4.1 | The “a priori” Approach | 35 |
| 4.2 | The StdTrip Process | 36 |
| 4.3 | Conversion Stage | 37 |
| 4.3.1 | Relational model to Entity Relationship | 37 |
| 4.3.2 | Entity Relationship to OWL mapping | 41 |
| 4.4 | Alignment | 43 |
| 4.4.1 | K-Match | 43 |
| 4.5 | Selection | 45 |
| 4.6 | Inclusion | 46 |
| 4.7 | Completion | 47 |

| | | |
|----------|---|-----------|
| 4.8 | Output | 49 |
| 5 | StdTrip Tool | 50 |
| 5.1 | StdTrip Architecture | 50 |
| 5.2 | K-Match Architecture | 53 |
| 6 | Conclusions | 58 |
| 6.1 | Contributions | 58 |
| 6.2 | Limitations and Future Work | 58 |
| | Bibliography | 61 |
| A | Triplify mapping file for the Author-Publication example | 68 |
| B | Triple Schema for the Author-Publication example | 70 |
| C | RDFRendererVisitor format | 72 |
| D | OWLAXiomsRendererVisitor format | 74 |

List of Figures

| | | |
|----|---|----|
| 1 | An RDF Graph Describing Eric Miller [Manola & Miller 2004]. . | 19 |
| 2 | RDB-to-RDF Mapping Process | 27 |
| 3 | Syntatic Approach [Casanova et al. 2007]. | 29 |
| 4 | Semantic Approach [Casanova et al. 2007]. | 30 |
| 5 | StdTrip Architecture. | 36 |
| 6 | Author-Publication relational schema. | 37 |
| 7 | ER : Identification of ER elements for each table | 39 |
| 8 | ER : Definition of relationship cardinality for the Author-Publication example | 39 |
| 9 | ER : Definition of attributes | 40 |
| 10 | ER : Definition of entities and relationships identifiers | 41 |
| 11 | OWL ontology that resulted from applying the transformation process ER to OWL to the Author-Publication example | 43 |
| 12 | The <i>K-Match</i> overall matching process | 45 |
| 13 | OWL ontology after the <i>Selection</i> stage | 47 |
| 14 | StdTrip Architecture | 51 |
| 15 | K-Match Architecture | 54 |
| 16 | Results of the OAEI2009: Precision and recall [Euzenat et al. 2009]. | 55 |

List of Tables

| | | |
|---|---|----|
| 1 | Comparison of RDB-to-RDF approaches (*)Partial | 34 |
| 2 | Correspondence between ER schema and OWL ontology components | 42 |
| 3 | RDF Standard Vocabularies | 44 |
| 4 | Similarity Cube : Similarity values from a partial alignment between <i>O1</i> and <i>O2</i> for the Author-Publication example | 46 |
| 5 | Similarity Matrix : Combined similarity values combined of Table 4 for the Author-Publication example | 46 |

Ad Majorem Dei Gloriam

San Ignacio de Loyola, (1491 – 1556).