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

Percy E. Salas[1], **Karin Breitman**[1] **(Advisor)**

[1]Departamento de Informática – Pontifícia Universidade Católica do Rio de Janeiro
R. Mq. de S. Vicente, 225 – 22451-900 – Rio de Janeiro – RJ – Brazil

{psalas,karin}@inf.puc-rio.br

Abstract. *The Linked Data standard — which is based on the publication of data in the Web as RDF triples — is gaining increasing attention as a way to promote open data interoperability. The publication of data originally stored in relational databases as RDF is known as the RDB-to-RDF approach. The first step in this process requires the mapping of database concepts to RDF vocabularies, used as the base for generating the RDF triples. The selection of such vocabularies is extremely important, because the more standards are reused, the easier it will be to interlink the resulting datasets to other previously published. However, the tools available today, instead of providing adequate support for reuse of standard vocabularies, reinforce the creation of new ones. In our work, we present the StdTrip process, which provides general guidelines for performing the mapping phasis of the RDB-to-RDF approach, promoting the reuse of standard RDF vocabularies.*

1. Introduction

The open data paradigm is defined as the process by which information is produced, archived and distributed in open raw formats in a way that it can be shared, discovered, accessed and easily manipulated by those desiring such data. The Semantic Web provides a common framework that allows data to be shared and reused across applications, enterprises, and community boundaries. Particularly, for representing open data, W3C recommends the Linked Data standard, which is based on the representation of data in the form of sets of RDF triples.

The publication of data originally stored in relational databases requires the conversion of a myriad of RDB datasets to RDF triplesets. This process is known as RDB-to-RDF approach or triplification. 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, to be used as the base for generating the RDF triples. The construction of this vocabulary is extremely important, because the more one reuses well known standards, the easier it will be to interlink the result to other existing datasets. This approach greatly improves the ability of third parties to use the information provided by governments in ways not previously available or planned, such as the creation of data mashups, i.e., the merge of data from different data sources, in order to produce comparative views of the combined information.

There are applications that provide support to the convertion phasis of the RDB-to-RDF approach, i.e., the mechanical process of transforming relational data to RDF triples,

[*]Percy E. Salas was supported by CAPES/PROEX.

such as Triplify [Auer et al. 2009], Virtuoso RDF views [Erling and Mikhailov 2009] and D2RQ [Bizer and Seaborne 2004]. However, they offer very little support to users during the mapping phasis. In our work, we present the StdTrip process, which aims at guiding the users in the task of converting relational data to RDF, providing support in the step of creating a conceptual model of the RDF datasets, i.e., the mapping phasis. Based on an *a priori* design approach, StdTrip promotes the reuse of standard — W3C recommended — RDF vocabularies, when possible, suggesting the reuse of vocabularies already adopted by other RDF datasets, otherwise. As such, the main contributions of this work comprise:

- The StdTrip process, an approach to guide users in the mapping phasis of the RDB-to-RDF process, promoting the reuse of standard RDF vocabularies.
- The StdTrip tool, an implementation of the respective approach that serves to demonstrate the feasibility of the proposed process.

Due to space limitation, this paper does not present a detailed description of the StdTrip process or the Stdtrip tool implementation. The full dissertation, the list of related journal and conference publications, courses, pending submissions and implementations details are available in the Web[1].

The rest of this paper is organized as follows. In Section 2, we discuss the interoperability problems and explain the *a priori* matching approach. In Section 3, we present the StdTrip process. In Section 4, we present the list of related publications. Finally, in Section 5, we conclude this paper.

2. The Interoperability Problem

In the words of Bizer, Cyganiak and Heath [Bizer et al. 2007], *"in order to make it as easy as possible for client applications to process data"*, when publishing open data in the Linked Data standard, one *"should reuse terms from well-known vocabularies wherever possible"*. Therefore, one *"should only define new terms if he can not find required terms in existing vocabularies"*. In that way, ontologies that describe the published data serve as the global schema upon which is based the interoperability with other datasets.

In the RDB-to-RDF mapping phasis, a preliminary stage consists in the definition of a generic ontology that describes how the RDB schema concepts are represented in terms of generic RDF classes and properties. The sheer adoption of this ontology, however, is not sufficient to secure interoperability. In a distributed and open system, such as the Semantic Web, different parties tend to use different ontologies to describe specific domains of interest, raising interoperability problems.

Ontology alignment techniques can be applied to solve heterogeneity problems. Such techniques are closely related to schema matching approaches, which consist of taking two schemata as input and producing a mapping between pairs of elements that are semantically equivalent. Matching approaches may be classified as syntactic vs. semantic and, orthogonally, as *a priori* vs. *a posteriori* [Casanova et al. 2007]. Both syntactic and semantic approaches work *a posteriori*, in the sense that they start with existing datasets, and try to identify links between the two. A third alternative — the *a priori* approach — is proposed in [Casanova et al. 2007], where the author argues that,"when specifying databases that will interact with each other, the designer should first select an appropriate

[1] `http://www.inf.puc-rio.br/~psalas/ctd2012`

standard, if one exists, to guide design of the exported schemas. If none exists, the designer should publish a proposal for a common schema covering the application domain".

Matching two schemata that were designed according to the *a priori* approach, is an easier process as there is a consensus on the semantics of terminology used, thus avoiding possible ambiguities. Unfortunately, this is not what happens in practice. Most teams prefer to create new vocabularies — as do the vast majority of tools that support this task —, rather than spending time and effort to search for adequate matches. We believe that this fact is mainly due to the distributed nature of the Web itself, i.e., there is no central authority one can consult to look for a specific vocabulary. Semantic search engines, such as Watson, works as an approximation for such mechanism. Notwithstanding, there are numerous standards that designers can not ignore when specifying triple sets, and publishing their content.

3. StdTrip process

The StdTrip process aims at guiding users during the mapping phasis of the RDB-to-RDF approacch. Most tools that support this task do that by mapping tables to RDF classes, and attributes to RDF properties, with little concern regarding the reuse of existing standard vocabularies [Auer et al. 2009] [Erling and Mikhailov 2009]. Instead, these tools create new vocabularies using the internal database terminology, such as the table and attribute names. We believe that the use of standards in schema design is the only viable way for guaranteeing future interoperability [Breitman et al. 2006]. The StdTrip process is anchored in this principle, and strives to promote the reuse of standards by implementing a guided process comprised by six stages. StdTrip architecture is represented in Figure 1

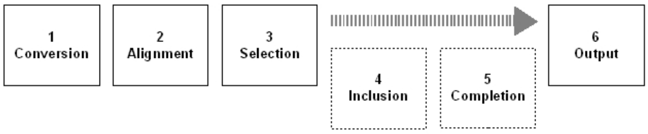

Figure 1. StdTrip Architecture.

Stages 1 to 6 were named, respectively, Conversion, Alignment, Selection, Inclusion, Completion and Output according with the main operation performed in each one. While stages 1, 2, 3 and 6 are obligatory, stages 4 and 5 are optional. Each stage is summarized as follows:

1. Conversion: This stage consists in transforming the structure of the relational database in an RDF ontology. It takes as input the relational database schema, which contains the metadata of the RDB. This stage is comprised by two major operations. In the first operation, we transform the relational database schema into an Entity-Relationship (ER) model. In the second operation, we transform the Entity-Relationship model, resulting from the previous operation, into an OWL ontology.
2. Alignment: The alignment stage is where lies the essence of our approach. As the name suggests, is in this stage that we apply existing ontology alignment algorithms. We aim at finding correspondences between a generical ontology — obtained in the previous stage — and standard well-known RDF vocabularies.

The alignment operation is supported by the *K-match* ontology alignment tool, which is based on a collaborative approach to find matches between ontology terms. Better than proposing "yet another ontology matching tool", *K-match* capitalizes from years of collaborative research and results obtained by the Semantic Web community, particularly during the OAEI contest[2].

The alignment process is comprised by three steps: the first step consists in the execution of different ontology matchers, the second step combines the results of the previous step applying aggregation strategies, and the final step applies one of several selected strategies to choose the match candidates for each ontology term. The steps of the *K-match* alignment process is illustrated by the Figure 2

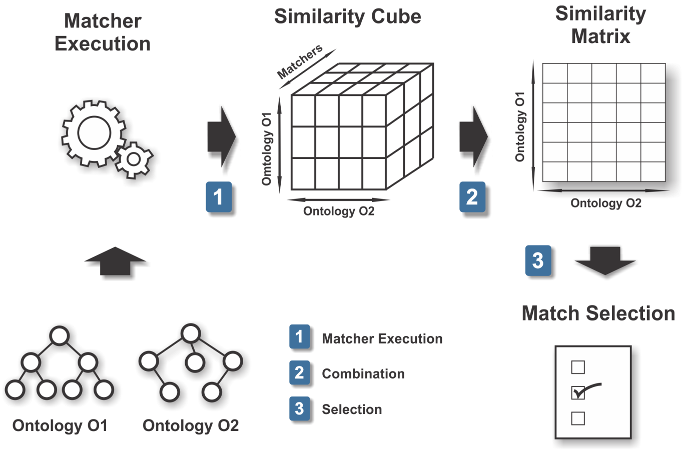

Figure 2. The *K-Match* overall matching process

3. Selection: During the selection stage human interaction plays an essential role. Ideally, the user should know well the application domain, because he or she will have to choose the vocabulary elements that best represent each concept in the database. The user will select each vocabulary element from a list of possibilities, listed in decreasing order of similarity value obtained as the result of the previous stage.
4. Inclusion: There are cases where the selection stage does not yield any result, i.e., either there is no element in the known vocabularies that matches the concept in the database, or none of suggestions in the list is considered adequate by the user. For such cases we provide a list of terms from other, vocabularies in the web that might be possible matches. The choice of these vocabularies is domain-dependent, and the search, based on keywords, is done using a semantic web searching tools. The rationale is the following, "if your concept is not covered by any of the known standards, look around and see how others dealt with it. By choosing a vocabulary in use, you will make your data more interoperable in the future, than by creating a brand new vocabulary."

   This stage is accomplished with the aid of existing mechanisms for searching semantic documents offered by Semantic Web searchers, namely Watson[3], which executes a keywords based search. In order to improve the quality of the results,

[2] `http://oaei.ontologymatching.org/`

[3] `http://kmi-web05.open.ac.uk/WatsonWUI/`

it is crucial to follow some *"tuning"* and configuration guidelines to get the best out of this type of service.

5. Completion: If, for some terms, none of the previous stages was able to provide appropriate RDF mapping, the user will have to define a new vocabulary. During this stage, we help users providing recommendations and best practices on how his or her vocabulary should be published on the Web, how choose an appropriate URI namespace, and its constituent elements (classes and properties).
6. Output: This is not properly a stage, rather the output of the StdTrip process, which produces two artifacts.
   (a) **A mapping specification file.** This artifact serves as the core parameterization for a RDB-to-RDF conversion tool. The specification file format can be easily customized for several approaches and tools that provide support to the mechanical process of transforming RDB into a set of RDF. Among them, there are the formats used by Triplify [Auer et al. 2009], Virtuoso RDF views [Erling and Mikhailov 2009] and D2RQ [Bizer and Seaborne 2004], and also R2RML [Das et al. 2010], the new standardized language to map relational data to RDF.
   (b) **"Triples Schema".** The second artifact is an ontology representing the original database schema, with the corresponding restrictions, and maximizing the reuse of standard vocabularies.

4. Publications

- *Book Chapter*
  - Salas, P. E., Viterbo, J., Breitman, K., and Casanova, M. A. (2011b). StdTrip: Promoting the Reuse of Standard Vocabularies in Open Government Data. In Wood, D., editor, *Linking Government Data*, pages 113–133. Springer New York
- *Journal Papers*
  - Breitman, K., Salas, P. E., Viterbo, J., Saraiva, D., Gama, V., Casanova, M. A., Pires, R., Franzosi, E., and Chaves, M. (2012). Open Government Data in Brazil. *Intelligent Systems, IEEE* – ***(Qualis A1)***
  - Marx, E., Salas, P. E., Breitman, K., Viterbo, J., and Casanova, M. A. (2012). RDB2RDF plugin: Relational to RDF plugin for eclipse. *Software: Practice and Experience* – ***(Qualis A2)***
- *In Conference Proceedings*
  - Salas, P. E., Breitman, K., Casanova, M. A., and Viterbo, J. (2010a). StdTrip: An a priori design approach and process for publishing Open Government Data. In *Proceedings of the XXV Brazilian Symposium on Databases*, XXV Brazilian Symposium on Databases 2010, pages 41– 48, Belo Horizonte, MG, Brazil. SBC – ***(Qualis B3)***
  - Salas, P. E., Breitman, K., Viterbo F., J., and Casanova, M. A. (2010b). Interoperability by design using the StdTrip tool: an a priori approach. In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS'10, pages 43:1 – 43:3, New York, NY, USA. ACM
  - Salas, P. E., Marx, E., Mera, A., and Viterbo, J. (2011a). RDB2RDF plugin: relational databases to RDF plugin for eclipse. In *Proceedings of the 1st Workshop on Developing Tools as Plug-ins*, TOPI11, pages 28–31, New York, NY, USA. ACM

5. Conclusion

In our work, we introduced StdTrip, a process and tool that emphasize the standard-based, *a priori*, design of triples, in order to promote the reuse of vocabularies, hence facilitating open data interoperability, i.e., the integration among datasets in the Linked Data cloud.

StdTrip was initially conceived to serve as an aid in a training course “Publishing Open Government Data in Brazil”, an initiative of W3C Brasil to promote the adoption of the Linked Data technology by Brazilian government agencies. Target audiences were assumed to have no familiarity with Semantic Web techniques in general, nor with RDF vocabularies, in particular. To promote the reuse of vocabularies and standards, we designed a tool that “has it all in one place”, i.e., supports all the operations needed to create a RDB-to-RDF map that can be used in the conversion of relational datasets.

As further work — discussed in [Breitman et al. 2012] —, StdTrip was successfully applied in the creation of *dados.gov.br*, a Brazilian open government data repository containing statistical data reflecting government activity during the president Luiz Inacio ”‘Lula”’ da Silva mandates (2003 to 2010). The dataset, which comprises about 1,300 historic data series containing more than 4 million observations, was expressed in more than 30 million RDF triples being initially linked to DBpedia and GeoNames.

References

Auer, S., Dietzold, S., Lehmann, J., Hellmann, S., and Aumueller, D. (2009). Triplify: light-weight linked data publication from relational databases. In *WWW ’09: Proceedings of the 18th international conference on World wide web*, pages 621–630, New York, NY, USA. ACM.

Bizer, C., Cyganiak, R., and Heath, T. (2007). How to publish linked data on the web. Retrieved December 14, 2010, from `http://www4.wiwiss.fuberlin.de/bizer/pub/LinkedDataTutorial/`.

Bizer, C. and Seaborne, A. (2004). *D2RQ-treating non-RDF databases as virtual RDF graphs*.

Breitman, K., Casanova, M. A., and Truszkowski, W. (2006). *Semantic Web: Concepts, Technologies and Applications (NASA Monographs in Systems and Software Engineering)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Breitman, K., Salas, P. E., Viterbo, J., Saraiva, D., Gama, V., Casanova, M. A., Pires, R., Franzosi, E., and Chaves, M. (2012). Open Government Data in Brazil. *Intelligent Systems, IEEE*.

Casanova, M. A., Breitman, K., Brauner, D., and Marins, A. (2007). Database conceptual schema matching. *IEEE Computer*, 40(10):102–104.

Das, S., Sundara, S., and Cyganiak, R. (2010). R2RML: RDB to RDF mapping language. W3C RDB2RDF working group. Retrieved March 20, 2012, from `http://www.w3.org/TR/r2rml/`.

Erling, O. and Mikhailov, I. (2009). Rdf support in the virtuoso dbms. *Networked Knowledge-Networked Media*, pages 7–24.