

Interoperability by Design Using the StdTrip Tool: An a priori approach

Percy E. Salas Karin K. Breitman José Viterbo F. Marco A. Casanova
Departamento de Informática – Pontifícia Universidade Católica do Rio de Janeiro
Rua Marquês de S. Vicente, 225 - Rio de Janeiro, Brazil - CEP 22451-900
{psalas, karin, viterbo, casanova}@inf.puc-rio.br

ABSTRACT

A database conceptual schema is a high-level description of how database concepts are organized, typically as a set of classes of objects and their attributes. Triplification is the process by which a database schema, and its instances, are transformed into a RDF dataset. A major step 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 a vocabulary, to be used as the base in which to generate the RDF triples from. 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. Most triplifying engines today provide support to the mechanical process of transforming relational to RDF data. However, to best of our knowledge, none provide user support during the conceptual modeling stage. In this paper, we present StdTrip, a tool that guides users in this process. If possible, the tool promotes the reuse of standard, W3C recommended RDF vocabularies, or otherwise suggests the reuse of vocabularies already adopted by other RDF datasets.

Categories and Subject Descriptors

H.3.1 [Information Storage And Retrieval]: Content Analysis and Indexing

General Terms

Design

Keywords

Interoperability, Conceptual Modeling, Linked Data, Ontology Matching

1. INTRODUCTION

The process of interlinking newly produced datasets to existing ones consists of connecting the subject URI from one dataset with an object URI from another dataset by using links, expressed as RDF triples [2]. This match opera-

tion takes two vocabularies as input and produces a mapping between elements of the two. Many techniques support this process. e.g. ontology alignment, schema matching and data fusion. Good surveys of such techniques are presented by [3], [8] and [12].

Matching approaches may be classified as syntactic vs. semantic and, orthogonally, as a priori vs. a posteriori [5]. The syntactic approach consists of matching two vocabularies based on syntactical hints, such as attribute data types and naming similarities. The semantic approach uses semantic clues to generate hypotheses about vocabulary matching. It generally tries to detect how real world objects are represented in different datasets, and leverages on the information obtained to match different URIs. 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. This task is particularly time consuming, effort intensive and difficult to automate, as hinted by the simple example that follows, illustrated by Figure 1.

Consider two triple datasets D1 and D2, whose application domains are not entirely clear. Assume that D1 has a set of classes named *Games*, with properties *Name* and *ESRB* (Entertainment Software Rating Board), and D2 has a set of classes named *Gaming*, with properties *Name*, *Price*, and *Rating*, as shown in Figure 1. Using only syntactical similarity, *Games* would probably match with *Gaming*, and the *Name* property in both sets would definitely match with one another, but *ESRB* would not match with *Rating*.

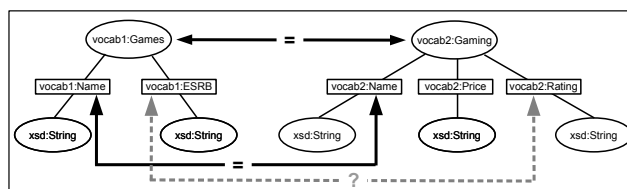


Figure 1: An example of interlinking datasets.

Now, if D1 and D2 describe stores that deal with computer game, e.g. BestBuy and Amazon, this matching is reasonable, though it still misses the match between *ESRB* and *Rating*, which can be assumed to refer to ratings assigned by the *ESRB*. However, if D1 describes the dataset of a travel agency specializing in safaris, matching *Game* (big game hunting) with *Gaming* (computer games) is obviously inaccurate. Unless the two datasets share a common vocabulary,

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.

I-SEMANTICS 2010 September 1–3, 2010, Graz, Austria.

© Copyright 2010 ACM 978-1-4503-0014-8/10/09 ...\$10.00.

there is no way to fully automate this process, human intervention will always be needed to identify possible matches and disambiguate dubious ones.

2. DESIGNING FOR INTEROPERABILITY: THE A PRIORI APPROACH

The a priori matching approach emphasizes that, “when specifying databases that need to interact with others, the designer should first select an appropriate standard, if one exists, to guide design of the resulting database. If none exists, the designer should publish a proposal for a common schema covering the application domain” [5].

The same philosophy is applicable to Linked Data, as stated by Bizer, Cyganiak and Heath, “in order to make it as easy as possible for client applications to process your data, you should reuse terms from well-known vocabularies wherever possible. You should only define new terms yourself if you can not find required terms in existing vocabularies” [2].

Unfortunately, that is not what happens in practice. Most (database) designers prefer creating new vocabularies (as do the vast majority of triplification tools) to spending the required time and effort to search for adequate matches [9]. We believe that is mostly due to the distributed nature of the Web itself, i.e., there is no central authority one can consult. Semantic search engines, such as Watson, function as an approximation.

Notwithstanding, there are numerous standards that designers cannot ignore when specifying triple sets and publishing their content, such as Dublin Core or Dbpedia ontologies. The term standard is used here in loose way, in that it encompasses vocabularies with different status (recommended, submitted, etc) in regard to standard authorities. Based on the notion that a good design, based on agreed upon standards, will promote and facilitate future interoperability, we propose the StdTrip tool, detailed in the next section.

3. STDTRIP TOOL

The StdTrip tool guides users during the conceptual modeling stages of the triplification process. A good metaphor for this process is a translation, from the relational model, to the RDF-triple model. Most triplifying tools today do that by mapping tables to RDF classes, and attributes to RDF properties, with no concern in identifying possible matches with existing standard vocabularies; instead, the tools create new vocabularies.

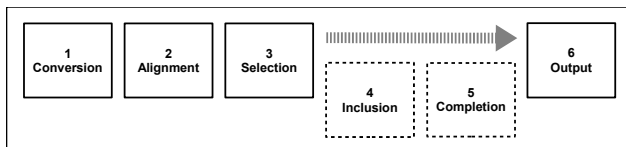


Figure 2: StdTrip Architecture.

We believe that the use of standards in schema design is the only viable way to guarantee future interoperability [4, 6, 10]. The StdTrip tool is anchored in this principle, and strives to promote the reuse of standards by implementing

a guided process that comprises six steps. The StdTrip architecture is represented in Figure 2. Steps 1 to 6 were named, respectively, Conversion, Alignment, Selection, Inclusion, Completion and Output, according to the main operation performed in each step. While Steps 1, 2, 3 and 6 are obligatory, Steps 4 and 5 are optional. Each step is summarized as follows:

1. **Conversion.** This step consists in transforming the structure of the relational database to an RDF ontology. In this stage, the designer may rely on approaches such as W-Ray [11], in which he manually defines a set of database views that capture which data should be published, and then specifies templates that indicate how RDF triples should be generated.
2. **Alignment.** This step uses the K-match ontology alignment tool¹ to match the ontology obtained in Step 1 with the set of standard vocabularies. The alignment process considers the ontology schema previously obtained as the source schema to be recursively aligned with each ontology representing the standard vocabularies. These ontologies are the target and each result in the alignment is allocated for each term. Eventually, the results are presented as suggestions for each term, i.e., for each schema element (table or attribute) a list of possible matches is presented. For example, a table named Person would be matched to foaf:maker, dc:creator.
3. **Selection.** This step presents to the user a list of possibilities from which he or she can select the vocabulary element that best represents each concept in the database.
4. **Inclusion.** If, for a given element, the process does not yield any result (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, StdTrip provides a list of triples from other vocabularies that might be a possible match. This is done using Watson, a Web interface for searching ontologies and semantic documents using keywords. 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 already in use, you will make it easier to interlink your vocabulary in the future, than by creating a brand new vocabulary.”
5. **Completion.** If none works, users are directed to the Best Practice Recipes for Publishing RDF Vocabularies [1].
6. **Output.** The tool outputs two artifacts: (1) a configuration file, to serve as the parameterization for a standard triplification tool. (2) an ontology that contains the mappings of the original database schema to standard RDF vocabularies.

¹K-match is a tool that combines the ontology matchers that obtained the highest rank for the last OAEI benchmark, i.e., Lily, Aroma and Anchor-Flood [7]. By combining the results from each matcher individually, K-match yields yet more accurate results.

4. SHORTCOMINGS AND CHALLENGES

The StdTrip tool was designed to support a training course on publishing Open Government Data in Brazil. The target audience was assumed to have no familiarity with Semantic Web techniques, in general, and with RDF vocabularies, in particular. To promote vocabulary and standard reuse, we needed to provide a tool that “had it all in one place”.

The StdTrip tool served an educational purpose by “reminding” or by introducing vocabulary concepts users were unaware of. This first version of the StdTrip tool, however, can be greatly improved. In what follows we discuss how we plan to tackle its shortcomings.

First, we intend to sophisticate the initial transformation step. In order to use the K-match tool we must create a RDF graph representation of the database schema. The process implemented today is similar to that used by the D2RServer tool, and is very crude. Several improvements are possible. First, the structure of the database itself might be useful, e.g., foreign keys should be mapped to object properties. Second, we are going to explore the use of database instances to infer the semantics of specific attributes, when their names or labels do not provide useful information. For example, an attribute named Af675T, with the following format XXX-XXXXXXXXXX (where Xs are numbers) may be automatically identified as ISBN numbers [14]. Finally, following the work of [13] that uses Wordnet extensions to expand and normalize the meaning of database comments, as used as a source for additional semantics.

Furthermore, as users are likely to be confronted with more than one choice, e.g., foaf:Person or foaf:Agent, it is reasonable to include a mechanism, even if informal, to capture design decisions during the modeling steps (3 and 4). A what-who-why memory is a beneficial asset for future improvements and redesign of the dataset.

Finally, by far, our largest challenge is the language barrier. All standards are in English when our intended use of StdTrip, eGov data, is essentially composed of data in Portuguese. So far, we have been manually translating the table and attribute names, that is, preprocessing the input to stage 1. We admit to have privately experimented with the Google translator API, and had (embarrassingly) good results. We are currently investigating the possibility of using Wordnet and extensions.

5. CONCLUSION

We introduced StdTrip, a tool that emphasizes the use of standard-based, a priori design of triples, in promoting interoperability, reuse and facilitates integration with other datasets in the LOD cloud. Combined with any triplification mechanism, StdTrip guides users in the process of modeling their original databases in terms of well-known, de facto RDF standard vocabularies.

An interesting byproduct of this process is the production of a mapping of the original relational database schema to standard RDF vocabularies. This artifact is valuable to the construction of mediators designed to integrate with the original database, e.g. traditional Deep Web interfaces.

6. ACKNOWLEDGMENTS

This research was made possible by grants E-26/170028/2008 from FAPERJ and 557.128/2009-9 from CNPq, at the Brazilian Web Science Institute.

7. REFERENCES

- [1] D. Berrueta, J. Phipps, A. Miles, T. Baker, and R. Swick. Best practice recipes for publishing rdf vocabularies, August 2008.
- [2] C. Bizer, T. Heath, D. Ayers, and Y. Raimond. Interlinking Open Data on the Web (Poster). In *In Demonstrations Track, 4th European Semantic Web Conference (ESWC2007)*, 2007.
- [3] J. Bleiholder and F. Naumann. Data fusion. *ACM Comput. Surv.*, 41(1):1–41, 2008.
- [4] K. Breitman, M. A. Casanova, and W. Truszkowski. *Semantic Web: Concepts, Technologies and Applications (NASA Monographs in Systems and Software Engineering)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
- [5] M. A. Casanova, K. Breitman, D. Brauner, and A. Marins. Database conceptual schema matching. *IEEE Computer*, 40(10):102–104, 2007.
- [6] M. A. Casanova, T. Lauschner, L. A. P. Leme, K. Breitman, A. L. Furtado, and V. Vidal. A strategy to revise the constraints of the mediated schema. In *Proc. of the 28th Int'l. Conf. on Conceptual Modeling*, volume 5829 of *Lecture Notes in Computer Science*, pages 265–279. Springer, Nov. 2009.
- [7] J. Euzenat, A. Ferrara, L. Hollink, and et al. Results of the ontology alignment evaluation initiative 2009. In *Proc. 4th of ISWC Workshop on Ontology Matching (OM)*, 2009.
- [8] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2007.
- [9] S. Kinsella, U. Bojars, A. Harth, J. G. Breslin, and S. Decker. An interactive map of semantic web ontology usage. In *IV '08: Proceedings of the 2008 12th International Conference Information Visualization*, pages 179–184, Washington, DC, USA, 2008. IEEE Computer Society.
- [10] L. A. P. Leme, M. A. Casanova, K. Breitman, and A. L. Furtado. Owl schema matching. *Journal of the Brazilian Computer Society*, 16(1):21–34, Apr. 2010.
- [11] H. Piccinini, M. Lemos, M. A. Casanova, and A. Furtado. W-Ray: A Strategy to Publish Deep Web Geographic Data. In *Proceedings of the 4th International Workshop on Semantic and Conceptual Issues in GIS (SeCoGIS 2010)*, to appear, 2010.
- [12] E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350, 2001.
- [13] S. Sorrentino, S. Bergamaschi, M. Gawinecki, and L. Po. Schema normalization for improving schema matching. In *Proc. of the 28th International Conference on Conceptual Modeling (ER '09)*, pages 280–293, Berlin, Heidelberg, 2009. Springer-Verlag.
- [14] J. Wang, J.-R. Wen, F. Lochovsky, and W.-Y. Ma. Instance-based schema matching for web databases by domain-specific query probing. In *Proc. of the 13th international conference on Very large data bases (VLDB '04)*, pages 408–419. VLDB Endowment, 2004.