

StdTrip: An *a priori* design approach and process for publishing Open Government Data*

Percy E. Salas¹, Karin K. Breitman¹, Marco A. Casanova¹, José Viterbo²

¹Department of Informatics – PUC-Rio
Rio de Janeiro, RJ – Brazil CEP 22451-900

²Departamento de Ciência e Tecnologia – UFF
Rio das Ostras, RJ – Brazil CEP 28.890-000

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

Abstract. *Open Government Data (OGD) consists in the publication of public information data in formats that allow it to be shared, discovered, accessed and easily manipulated by those desiring the data. This approach requires the triplification of datasets, i.e., the conversion to RDF of database schemas and their instances. 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 in which to generate the triples from. The construction of this vocabulary is extremely important, because the more standards are reused, the easier it will be to interlink the result to existing datasets. However, today's tools do not support reuse of standard vocabularies in the triplification process, but rather they create new vocabularies. In this paper, we present the StdTrip process that guides users in the triplification process, while promoting the reuse of standard, W3C recommended, RDF vocabularies in the first place and, if not possible, by suggesting the reuse of other vocabularies already in employed by other RDF datasets on the Web.*

1. Introduction

Open Government Data (OGD) means the publication of information produced, archived and distributed by public organizations (e.g. legal, financial, bibliographic) in open raw formats, and ways that make it accessible and readily available to all and allow reuse, such as the creation of data mashups, i.e., the merging of data from different data sources, producing comparative views of the combined information [Accar et al. 2009].

A database dump or zipped packages for bulk data download is a traditional – and crude – approach for publishing government data. In this case, third parties are capable of using tools to separate and extract the data from the HTML code, transforming it into a more automatic reusable format, and then mashing it up with other sources. However, this approach requires a large effort on the data consumer side. There are cases in which governments are providing access to information through specific APIs. In most cases, this means that the consumer has access to the data only in the way

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

the producer thinks it should be accessed, e.g., through certain methods. The consumer does not have access to the raw data or to a holistic view of it.

Nevertheless, the focus of OGD is on publishing data that can be shared, discovered, accessed, and easily manipulated by those desiring the data [Bennet & Harvey 2009]. The Semantic Web provides a common framework that allows data to be shared and reused across applications, enterprises, and community boundaries. It offers technologies to describe, model and query these data. With the adoption of the Semantic Web approach, public organizations would be able to publish datasets annotated with domain-specific vocabularies, and offer query interfaces for applications in which to access public information in a non-predefined way. This would greatly improve the ability of third parties to use the information provided by governments in ways not previously available or planned.

A fundamental step in this approach consists in the conversion of a myriad of public information datasets, represented by database schemas and their instances, to RDF datasets. A key issue in this process, known as triplification, 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 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 [Breslin et al. 2009].

There are triplifying engines that provide support to the mechanical process of transforming relational data to RDF triples [Auer et al 2009, D2R Server¹, OpenLink Virtuoso²]. However, they offer very little support to users during the conceptual modeling stage. In this paper, we present the StdTrip process that guides users in this process, while promoting the reuse of standard, W3C recommended, RDF vocabularies in the first place and, if not possible, by suggesting the reuse of other vocabularies already in employed by other RDF datasets on the Web.

The rest of this paper is divided as follow. In Section 2, we discuss the basic concepts involved in the process of interlinking newly produced datasets to existing ones. In Section 3, we explain the a priori matching approach. In Section 4, we present the StdTrip process to be used in the conceptual modeling stages of the triplification process. Finally, in Section 5, we discuss some limitations of our approach and the challenges to be met in the future.

2. Basic Concepts

Before describing our approach, we call attention to the process of interlinking newly produced datasets to existing ones. Briefly, this process consists of connecting the subject URI from one dataset with an object URI from another dataset by using links, expressed as RDF triples [Bizer et al. 2007].

This matching operation takes two vocabularies³ as input and produces a mapping between elements of the two. Many techniques support this process. e.g.

1 <http://www4.wiwiw.fu-berlin.de/bizer/d2r-server/>

2 <http://virtuoso.openlinksw.com/>

3 We use the term vocabulary in a very loose sense, inclusive of the notions of thesauri, ontologies and database schema, i.e., a generalization to designate any conceptual model that represents the organization of a data collection.

ontology alignment, schema matching and data fusion. Good surveys of such techniques are presented by [Rahm & Bernstein 2001], [Euzenat & Shvaiko 2007] and [Bleiholder & Nauman 2008].

Matching approaches may be classified as syntactic vs. semantic and, orthogonally, as a priori vs. a posteriori [Casanova et al. 2007]. 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.

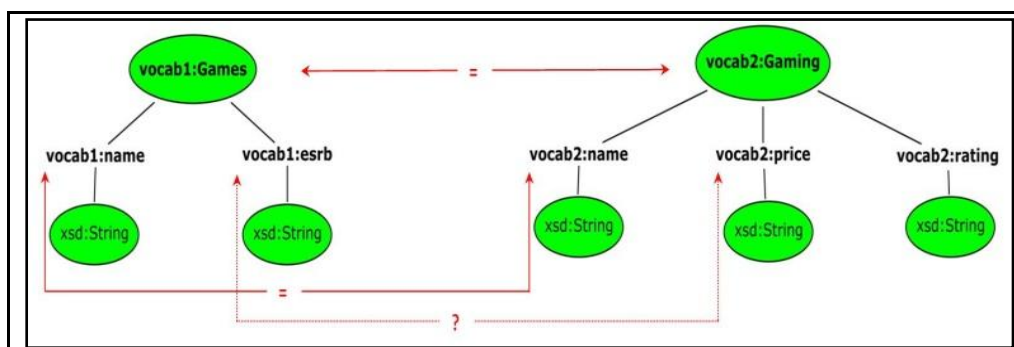


Figure 1. An example of interlinking datasets

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

3. 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” [Casanova et al. 2007]. The same philosophy is applicable to Linked Data – the Semantic Web

standard upon is based the publication of OGD –, 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” [Bizer et al. 2007].

Unfortunately, that is not what happens in practice. Most teams prefer to create new vocabularies (as do the vast majority of triplification tools), rather than spending time and effort to search for adequate matches [Kinsella et al. 2008]. 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 Watson⁴, function as an approximation. Notwithstanding there are numerous standards that designers can not ignore when specifying triple sets and publishing their content. Table 1 presents a list of some of these. Again, the term standard is used in loose way, in that it encompasses vocabularies with different status (recommended, submitted, etc.) in regards with standard authorities.

Based on the notion that good design, based on agreed upon standards, will promote and facilitate future interoperability, we propose the StdTrip Process, detailed in the next section.

Table 1. RDF Vocabularies

Ontology Name	Prefix	Namespace
Change Set	cs	http://purl.org/vocab/changeset/schema#
DBpedia Ontology	dbpedia	http://dbpedia.org/ontology/
Dcat: Data Catalog Vocabulary	dcat	http://www.w3.org/ns/dcat#
Dublin Core	dc	http://purl.org/dc/elements/1.1/
Dublin Core Terms	dcterms	http://purl.org/dc/terms/
FOAF: Friend Of A Friend	foaf	http://xmlns.com/foaf/0.1/
Geo: Geo Positioning	geo	http://www.w3.org/2003/01/geo/wgs84_pos#
GeoNames	gn	http://www.geonames.org/ontology#
MOAT: Meaning Of A Tag	moat	http://moat-project.org/ns#
Music Ontology	mo	http://purl.org/ontology/mo/
Programmes Ontology	po	http://purl.org/ontology/po/
SIOC: Semantically-Interlinked Online Communities	sioc	http://rdfs.org/sioc/ns#
SKOS: Simple Knowledge Organization System	skos	http://www.w3.org/2004/02/skos/core#
void: Vocabulary of Interlinked Datasets	void	http://rdfs.org/ns/void#

4. StdTrip Process

The StdTrip process aims at guiding users during the conceptual modeling stages of the triplification process. A good metaphor for it is that of translation, from the relational, 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 with identifying possible matches with existing standard vocabularies. Instead, these tools create new

4 <http://watson.kmi.open.ac.uk/WatsonWUI/>

vocabularies. However, we believe that the use of standards in schema design is the only viable way to guarantee future interoperability [Breitman et al. 2006, Casanova et al. 2009, Leme et al. 2010]. The StdTrip process is anchored in this principle, and strives to promote the reuse of standards by implementing a guided process (depicted in Figure 2), which comprises the six steps, 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 [Piccinini et al. 2010], 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 tool⁵ to match the ontology obtained in Step 1 with the set of standard vocabularies in Table 1 (and others, if the tool is so configured). This operation provides, for each schema element (table or attribute) a list of possible matches. 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 [Berrueta et al. 2008].
6. **Output.** The process outputs two artifacts: (1) a configuration file, to serve as the parameterization for a standard triplification tool. (2) an ontology that maximizes contains the mappings of the original database schema to standard RDF vocabularies.

5 K-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 [Euzenat et al. 2009]. By combining the results from each matcher individually, K-match yields yet more accurate results.

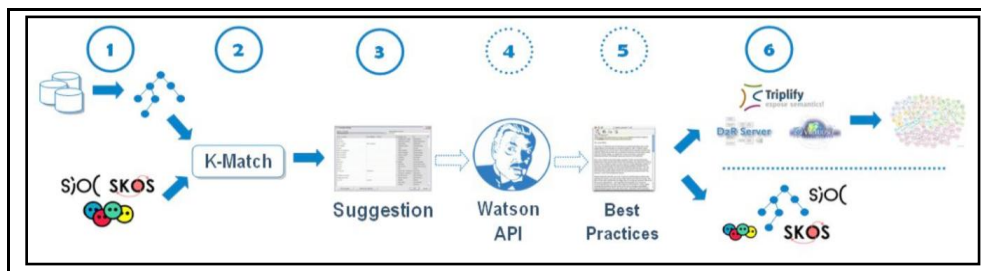


Figure 2. StdTrip Architecture

5. Conclusions

We introduced the StdTrip process, that emphasizes a standard-based, a priori design of triples to promote interoperability and reuse, and to facilitate integration with other datasets. In parallel with the StdTrip, but still work in progress, we are developing a companion tool that, combined with any triplification tool, guides users in the process of modeling their original databases in terms of well-known, de facto RDF standard vocabularies. StdTrip is a finalist at the Triplification Challenge, the yearly organized competition that awards prizes to the most promising approaches using Semantic Web and Linked Data technologies [Salas et al. 2010]. Winners will be known during the International Conference on Semantic Systems, in Austria this coming September.

StdTrip was initially conceived to serve as an aid in a training course on Publishing Open Government Data in Brazil. Target audiences were assumed to have no familiarity with Semantic Web techniques, in general, nor 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 approach served an educational purpose by “reminding” or by introducing vocabulary concepts users were unaware of.

The approach can be improved as follows. First, we must create an RDF graph representation of the database schema to be able to use the matching tool. The process implemented today is similar to the one used by the D2RServer tool, and it is very simple. Several improvements are possible. First, the structure of the database itself might be useful, e.g., foreign-keys should be mapped to object properties. We intend to experiment with reverse engineering the relational database, i.e., mapping relations schemes into entity-relationship diagrams [Casanova and Sá, 1984]. Entity relationship diagrams will provide good abstractions, e.g., subset, partonomy, that can be used to enrich and correlate RDF triple sets.

Secondly, we must observe that the database schema names, including table and column names, are typically inappropriate to be externalized. This implies that the designer must first define an external vocabulary, that is, a set of terms that will be used to communicate the data, materialized in the form of RDF triples, to Web users. That is to say that artificially generated primary keys, foreign keys that refer to such primary keys, attributes with domains that encode classifications or similar artifacts, if selected for the triplification process, should have their internal values replaced by their respective external definitions. For example, a classification code should be replaced by the description of the classification. Instance based approaches, such as the one proposed by Wang et al. [2004], might be useful. For example, an attribute named Ir675F, with the following format XXX-XXXXXXXXXX (where Xs are numbers) may easily be automatically identified as ISBN numbers. Finally, following the work of

[Sorrentino et al. 2009], we plan to use Wordnet extensions to expand and normalize the meaning of database comments, and use them 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 would be a good idea to include a rationale capturing mechanism, even if informal, to register 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.

References

- Accar, S., Alonso, J., Novak, K. (editors). "Improving Access to Government through Better Use of the Web". W3C Interest Group, 12 May 2009. Available at <http://www.w3.org/TR/egov-improving/>
- Auer, S., Dietzold, S., Lehmann, J., Hellmann, S. and Aumueller, D.. "Triplify: light-weight linked data publication from relational databases." Pp. 621-630 in Proceedings of the 18th international conference on World wide web. Madrid, Spain: ACM. 2009
- Bennet, D. and Harvey, A. "Publishing Open Government Data". W3C Work Group, 8 September 2009. Available at <http://www.w3.org/TR/gov-data/>
- Berrueta, D.; Phipps, J.; Miles, A. Baker, T.; Swick, R. "Best Practice Recipes for Publishing RDF Vocabularies". W3C Working Group. Available at <http://www.w3.org/TR/swbp-vocab-pub/>
- Bizer, C., Heath, T., Ayers, D. and Raimond, Y. Bizer, C., Heath, T., Ayers, D., Raimond, Y. "Interlinking Open Data on the Web"; Demonstrations Track at the 4th European Semantic Web Conference, Innsbruck, Austria. May 2007. Available at <http://www.eswc2007.org/pdf/demo-pdf/LinkingOpenData.pdf>
- Bleiholder, J.; Naumann, F. "Data fusion". ACM Computing Surveys 41(1), pages 1-41.
- Breitman, Karin K ; Casanova, M. A. ; Truszkowski, W. "Semantic Web: Concepts, Technologies and Applications". Londres: Springer, 2006. v. 1. 337 p.
- Breslin, J.; Passant, A.; Decker S. "The Social Semantic Web". Springer Verlag, 2009.
- Casanova, M.A., Amaral de Sá, J.E., "Mapping Uninterpreted Schemes into Entity-Relationship Diagrams: two Applications to Conceptual Schema Design". In: IBM Journal of Research and Development 28(1) pp. 82-94 (1984).
- Casanova, M.A.; Breitman, K.; Brauner, D. and Leme L.A. "Database Conceptual Schema Matching". Computer 40:102-104 (2007)
- Casanova, M.A., Lauschner, T., Leme, L., Breitman, K., Furtado, A., Vidal, V. "A Strategy to Revise the Constraints of the Mediated Schema". ER 2009: 265-279
- Euzenat, J. and Shvaiko, P. "Ontology matching". Springer-Verlag, 2007.
- Euzenat, J.; Ferrara, A.; Hollink, L. et al. "Results of the Ontology Alignment Evaluation Initiative 2009". In: Proc. 4th ISWC workshop on ontology matching (OM), Chantilly pp73-126, 2009
- Hausenblas, M.; Halb W. Interlinking of Resources with Semantics. Poster at the 5th European Semantic Web Conference (ESWC 08). Jun 2008.

- Kinsella, S.; Bojars, U.; Harth, A.; Breslin, J.G.; Decker, S. "An Interactive Map of Semantic Web Ontology Usage". *Information Visualisation*, 2008. IV '08. 12th International Conference, vol., no., pp.179-184, 9-11 July 2008
- Leme, L.A.; Casanova, M.; Breitman, K.; Furtado, A. "OWL schema matching" *Journal of Brazilian Computer Society*, Springer Verlag 16(1): 21-34 (2010)
- Piccinini, H., Lemos, M., Casanova, M.A. and Furtado, A. "W-Ray: A Strategy to Publish Deep Web Geographic Data". *SeCoGIS 2010*.
- Rahm and Bernstein, 2001Rahm, E. and Bernstein, P. "A survey of approaches to automatic schema matching". *The VLDB Journal*, 10(4):334–350.
- Raimond, Y. ; Sutton, C.; Sandler, M. "Automatic Interlinking of Music Datasets on the Semantic Web". In *WWW 2008 Workshop: Linked Data on the Web (LDOW2008)*, Beijing, China, 2008.
- Salas, P., Breitman, K. and Casanova, M.A. "Interoperability by Design Using the Std-Trip Tool: an a priori approach"(Triplification Challenge Submission), To Appear In *Proceedings of the International Conference on Semantic Systems 2010 (I-SEMANTICS'10)*
- Sorrentino, S.; Bergamaschi, MaciejGawinecki, Laura P. "Schema Normalization for Improving Schema Matching". *ER 2009*: 280-293
- Wang, J., Wen, J., Lochovsky, F., and Ma, W. "Instance-based schema matching for web databases by domain-specific query probing". In *Proc. of the 13th Int'l. Conf. on Very Large Data Bases*, pages 408–419.