6
Conclusions

6.1 Contributions

In this thesis we proposed StdTrip process, a process and accompanying tool to guide users during the conceptual modeling stages of the triplification process, i.e., the translation from the relational to the RDF-triples model. In order to promote interoperability and reuse — facilitating the integration with other datasets —, StdTrip emphasizes a standard-based *a priori* design of triples.

To validate the proposed process, we developed a companion tool that helps the users in the process of modeling their original databases in terms of well-known — *de facto* — standard RDF vocabularies. StdTrip was a finalist at the Triplification Challenge, a yearly organized competition that awards prizes to the most promising approaches using Semantic Web and Linked Data technologies [Salas et al. 2010].

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 the adoption of standards and vocabulary reuse, we needed to provide a tool that "had it all in one place". The StdTrip approach served as an educational tool by "reminding" — or by introducing new — RDF vocabulary concepts to non expert users.

6.2 Limitations and Future Work

We believe our approach can be further improved as follows. First of all, as discussed in Section 4.3, typically the terminology used to describe the relational database, including table and column names, is inadequate to be externalized. To exemplify, we could think of a relationship element named *country_id* that relates *City* and *Country*, an acronym *tb_cust* that could represent a table *Customer* or, even worst, an attribute *Ir675F* representing

an ISBN code. In such cases, the StdTrip process tackles this lack of semantics with the following techniques.

- A domain expert (e.g. database administrator) first defines an external vocabulary, i.e., a set of terms that will be used to communicate the data materialized 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, when selected for the StdTrip process, should have their internal names replaced by the definitions in the external vocabulary, more meaningful and therefore best suited for data integration.
- A common user could replace the inadequate terminology, by consulting documents that fully describe the data represented in the database (e.g. glossary, data dictionary).

It is important to note that, currently none of these techniques is supported by an automatic, or even semi-automatic way, during the triplification, making this operation practically unfeasible in the absence of a domain expert, or a document that describes the database domain. In future work we plan to add semi-automatic techniques in order to help users decide and choose adequate terms to characterizes their own data, in the following ways:

- We can take advantage of instance based approaches, such as the one proposed by [Wang et al. 2004], to suggest more adequate attribute names based on the data stored in the dataset. For example, an attribute named *Ir675F*, in the format XXX-XXXXXXXXXX (where Xs are numbers) may be automatically identified as an ISBN number.
- Taking into consideration that the relationships in the ER model — derived from the relational model — often lack meaningful names, we can use the semantics of the elements related by these relationships and apply Natural Language Processing algorithms to suggest terms that better describe the relationship in question. For example. A relationship attribute named *country_id*, which relates the entities *City* and *Country*, can be replaced by *isPartOf*, in order to obtain an statement *City isPartOf Country*.
- 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.

Secondly, as we mentioned at the beginning of the Chapter 4, we assume that the input of the StdTrip is a relational database in third normal form

(3NF). This assumption has some drawbacks in practice, as many databases might not be well normalized. Without support for database normalization, users might be tempted to directly take the databases as input even if badly designed. We plan to tackle this drawback in the following ways:

- Following the approach of [Du & Wery 1999] and [Wang et al. 2000], we plan to automate the process of finding functional dependencies within data in order to eliminate data duplication in the source tables, and to algorithmically transform a relational schema to third normal form.
- We also plan to offer more input options, such as W-Ray [Piccinini et al. 2010], in which a set of database views, capturing the data that should be published, is manually defined. In this sense, another interesting and helpful input option could be using a valid SQL query against the input database.
- We noticed that most relational databases use autonumber column to set tables identifiers (Primary Key). This autonumber does not properly work as an identifier for well-known entities, such as people, institutions or organizations. Therefore, we plan to include the option of replacing the table primary keys, for a more suitable options that better identifies what the table represents whenever possible. For example, The table *Person* uses as primary key an autonumber column named *person_id*. We could change the identifier for a column named *SSN*, which provides a more meaningful label to the *Person* table.

Finally, as users are likely to be confronted with more than one choice during the StdTrip process, e.g., ***foaf:****Person* or ***foaf:****Agent*, we plan to reuse previous mapping files and to include a rationale capturing mechanism to register design decisions during the modeling process (stages discussed in Sections 4.5 and 4.6). A what-who-why memory would be a beneficial asset for future improvements and redesign of the dataset.