The Journal of Systems and Software 100 (2015) 67–79

Contents lists available at ScienceDirect

The Journal of Systems and Software

journal homepage: www.elsevier.com/locate/jss

MostoDEx: A tool to exchange RDF data using exchange samples

Carlos R. Rivero [a,*], Inma Hernández [b], David Ruiz [c], Rafael Corchuelo [c]

[a] University of Idaho, 875 Perimeter Drive, MS 1010, Moscow, ID 83844-1010, United States
[b] Universidad Autonoma de Chile, C/ Carlos Antunez, 1920 Santiago, Chile
[c] University of Sevilla, ETSI Informática, Avda. Reina Mercedes s/n, Sevilla E-41012, Spain

ARTICLE INFO

Article history:
Received 2 September 2013
Received in revised form
17 September 2014
Accepted 19 October 2014
Available online 25 October 2014

Keywords:
Data exchange
RDF
Schema mappings

ABSTRACT

The Web is evolving into a Web of Data in which RDF data are becoming pervasive, and it is organised into datasets that share a common purpose but have been developed in isolation. This motivates the need to devise complex integration tasks, which are usually performed using schema mappings; generating them automatically is appealing to relieve users from the burden of handcrafting them. Many tools are based on the data models to be integrated: classes, properties, and constraints. Unfortunately, many data models in the Web of Data comprise very few or no constraints at all, so relying on constraints to generate schema mappings is not appealing. Other tools rely on handcrafting the schema mappings, which is not appealing at all. A few other tools rely on exchange samples but require user intervention, or are hybrid and require constraints to be available. In this article, we present MostoDEx, a tool to generate schema mappings between two RDF datasets. It uses a single exchange sample and a set of correspondences, but does not require any constraints to be available or any user intervention. We validated and evaluated MostoDEx using many experiments that prove its effectiveness and efficiency in practice.

1. Introduction

The current Web is progressively evolving into a Web of Data in which RDF (Resource Description Framework) data are becoming pervasive (Heath and Bizer, 2011). There are thousands of datasets available, many of which share a common purpose but have been developed by independent organisations in isolation (Bizer et al., 2009). There are many initiatives whose goal is to link these datasets, which is the first step to perform complex integration processes (Heath and Bizer, 2011).

Integration usually refers to several crucial tasks, such as data integration (Lenzerini, 2002), data warehousing (Marileo et al., 2012), model evolution (Flouris et al., 2008), model matching (Shvaiko and Euzenat, 2013), record linkage (Wang et al., 2013), or data exchange (Fagin et al., 2005). In this article, we focus on the latter, whose goal is to populate a target dataset using data that come from one or more source datasets. Data exchange has been paid much attention in the database context, i.e., relational, nested-relational, or XML (Arenas and Libkin, 2008; Fagin et al., 2005; Popa et al., 2002). Furthermore, the emergence of RDF is motivating some authors to work on data exchange in the context of the Web of Data (Barceló et al., 2013; Parreiras et al., 2008; Rivero et al., 2013b).

Data exchange is performed by means of schema mappings, which are declarative specifications of the relationships amongst a source and a target datasets (Alexe et al., 2011a). Generating schema mappings automatically is appealing because this relieves users from the burden of handcrafting them, so researchers have focused on helping users generate them (Qian et al., 2012). Many current tools are based on the data models to be integrated (Haas et al., 2005; Bonifati et al., 2005; Raffio et al., 2008; Mecca et al., 2009; Marnette et al., 2011; Rivero et al., 2013c). By data model, we refer to a sets of entities (that is, classes and properties) and a set of constraints that describe additional features of entities (for instance, class A is a specialisation of class B, property P has class C as its domain, and so on). In the Web of Data, there are many data models that comprise very few or no constraints at all, which typically results in data models that merely specify set of entities (Lausen et al., 2008; Heath and Bizer, 2011). Therefore, relying on data models with constraints to generate schema mappings is not appealing in the general context of the Web of Data.

There exist other tools that do not rely on data models. Unfortunately, they rely on handcrafting the schema mappings (Mocan and Cimpian, 2007; Maedche et al., 2002; Parreiras et al., 2008; Bizer and Schultz, 2010; Dou et al., 2005; Ressler et al., 2007), which is not appealing at all; and a few others rely on exchange samples (Alexe et al., 2008, 2006, 2011b; Qian et al., 2012), which make

* Corresponding author. Tel.: +1 2088856592.
E-mail addresses: crivero@uidaho.edu (C.R. Rivero), ichernandez@uautonoma.cl (I. Hernández), druiz@us.es (D. Ruiz), corchu@us.es (R. Corchuelo).

http://dx.doi.org/10.1016/j.jss.2014.10.033
0164-1212/

them more appealing, but require user intervention, or are hybrid and require constraints to be available. Note that an exchange sample is an example of source data and how it is exchanged into target data.

In this article, we present MostoDEx,[1] a tool to automatically generate schema mappings between two RDF datasets using a single exchange sample and a set of *n*:*m* correspondences. An exchange sample comprises a subset of source data and a subset of target data that is the expected result of exchanging the source data. Correspondences are hints that specify which entities in the source and target datasets correspond to each other, i.e., are somewhat related (Bellahsene et al., 2011). These schema mappings can be easily transformed into SPARQL queries.

Our tool does not rely on constraints of the source and target data models and does not require any user intervention, not even to repair the input exchange sample. We have validated our tool using ten data exchange problems amongst various real-world datasets. In our validation, the execution time never exceeded one second, and the data exchanged were as expected by experts in every case, which suggests that it is very efficient in practice and that the generated schema mappings are appropriate. Additionally, we have evaluated the performance of our tool when data exchange problems scale. We used four synthetic data exchange patterns proposed by MostoBM (Rivero et al., 2013a), a benchmark for testing data exchange proposals in the context of the Web of Data. We instantiated the synthetic data exchange patterns into 2000 non-trivial data exchange problems that we used to evaluate our tool. Our evaluation results suggest that our tool works well as the data exchange problems scale.

The rest of the article is organised as follows: Section 2 presents the tools related to MostoDEx and its main contributions to the state of the art; Section 3 presents some preliminaries that are necessary to understand the internal details of our tool; Section 4 describes how our tool works; Section 5 reports on the validity and scalability evaluation of MostoDEx; and, finally, Section 6 recaps on our main conclusions.

2. Related work

In this section, we present other existing tools that are related to MostoDEx. We present some tools that require the user to handcraft the schema mappings in Section 2.1, others are based on the constraints that comprise the source and target data models to be integrated in Section 2.2, and a last group of tools are based on samples of data to perform data exchange in Section 2.3. Finally, we analyse and discuss the drawbacks of these tools in Section 2.4, which motivated us to work on a new proposal.

2.1. Handcraft-based tools

There are a number of tools that focus on handcrafting schema mappings, which are expressed as queries but can be viewed as implicitly generating schema mappings: WSEE (Mocan and Cimpian, 2007), which stands for the Web Services Execution Environment, builds on a formal framework to describe correspondences in terms of first-order logic formulae that are used to generate schema mappings using the Web Service Modeling Language (WSML). This tool focuses on the problem of data exchange in the context of semantic-web services, i.e., web services that are enriched with semantic annotations to improve their discovery and composition (Forte et al., 2008). This tool is similar in spirit to MAFRA (Maedche et al., 2002) (MApping FRAmework), whose focus is on modelling correspondences in a general-purpose setting. The main difference with the previous tool is that WSEE goes a step beyond formalising correspondences and executes them using a WSML reasoner to exchange data.

MBOTL (Parreiras et al., 2008) (Model-Based Ontology Translation Language) builds on the framework of Model-Driven Engineering in which the ATL (ATLAS Transformation Language) metamodel is extended to support RDF data models, which allows to express constraints on them using OCL (Object Constraint Language). MBOTL comprises a mapping language by means of which users can express schema mappings that are later transformed into the SPARQL query language by means of a library of ATL transformations. This is similar in spirit to R2R (Bizer and Schultz, 2010) (RDF to RDF), OntoMerge (Dou et al., 2005), and Snoogle (Ressler et al., 2007), the difference is the language used to represent the schema mappings: R2R and Snoogle use SPARQL 1.0; whereas OntoMerge uses Web-PDDL schema mappings that are run by means of a first-order logic reasoner.

2.2. Constraint-based tools

They focus on generating schema mappings building on correspondences and constraints on the source and target data models. These tools are able to compute subsets of data in the source dataset that need to be exchanged as a whole, and subsets of data in the target dataset that need to be created as a whole (Rivero et al., 2013b). To compute them, they rely on user-defined constraints and the inherent constraints of certain data models, such as paths from the root to a leaf in a nested-relational data model, or hierarchy relations amongst classes in an RDF data model. Then, several combinations of these subsets of data are used to generate the final schema mappings (Popa et al., 2002).

Clio (Haas et al., 2005) is the state-of-the-art tool in this field. It takes a source and a target nested-relational data models, a number of constraints of each data model, and a number of 1 : 1 correspondences between them as input, and it generates schema mappings that can be easily transformed into different query languages, such as XQuery, XSLT, or SQL. HePToX (Bonifati et al., 2005) is similar to Clio but it focuses on XML data models, which are a superset of nested-relational data models. Clip (Raffio et al., 2008) allows to generate schema mappings based on *n*:1 correspondences, and it uses a mapping visual language that was specifically designed for nested-relational data models, which includes grouping functions, aggregation functions, or dependent correspondences. +Spicy (Mecca et al., 2009) allows to compute core schema mappings that generate non-redundant target data when performing data exchange. ++Spicy (Marnette et al., 2011) improves +Spicy by allowing more expressive target constraints. MostoDE (Rivero et al., 2013c) is able to work with RDF data models whose constraints are interpreted as graphs that are traversed to compute source and target kernels. A kernel comprises a subset of the source data model that needs to be exchanged as a whole, and a subset of the target data model that needs to be created as a whole. Kernels are translated into schema mappings that are represented in SPARQL 1.1.

2.3. Sample-based tools

These tools aim to generate schema mappings from a set of exchange samples. In the relational or nested-relational contexts, SPIDER (Alexe et al., 2006) helps users understand and maintain the schema mappings generated by Clio by extracting exchange samples from the source and target datasets, and it illustrates the following: (1) relationships in a specific schema mapping, (2) sample source data that this schema mapping would extract when performing data exchange, and (3) the target data generated by

[1] A technical report and a research prototype are available somewhere else (Rivero et al., 2013, 2013).