Profiling Transformations in Heterogeneous and Large Scale Metadata Harvesting Processes

João Sequeira, João Edmundo, Hugo Manguinhas, Gilberto Pedrosa, José Borbinha
INESC-ID - Portugal
{joao.sequeira, joao.edmundo, gilberto.pedrosa, hugo.manguinhas, jlb}@ist.utl.pt

Keywords: metadata harvesting, data transformation, XML schema, profiling, data structure, value measures, metadata conformance.

1. Abstract

Nowadays, most organizations communicate automatically with each other through their information systems and the Internet. When those data structures are not unique, data transformation is a required process. Two fundamental issues of that process are the interpretation of the source and destination schemas, and the definition of the mapping relating them.

The definition of these mappings can also be called schema matching. Matching can be performed through two distinct methods, manually by humans or automatically by algorithms. Since matching can be a non-deterministic process, it suites best the manual process, but when the schemas are complex (and especially when there is an incomplete knowledge of the source schema) it might require a large intellectual effort. In these scenarios automatic processes can be used to produce drafts of the mapping, to be corrected and accepted later by a human. For example, complex algorithms based on methods such as machine learning can be used to produce mappings more approximated to the ones produced by humans; however, those algorithms need to be trained to perform correctly and learn with their input.

The motivation of this work is to contribute to improving interoperability in heterogeneous and large-scale data integration processes in the scope of digital libraries. Libraries, archives, museums and other related organizations face the need to share their resource descriptive metadata. This is the envisaged scenario in initiatives Europeana[1], TEL[2] and EuDML[3].

REPOX (Freire, 2006) is a framework for data harvesting and providing, with support for data transformation processes based on schema and data set profiling and matching.

2. REPOX

When used by a data provider (a library, museum, archive, etc.), REPOX can be an effective solution to expose data by OAI-PMH. For service providers, REPOX can be used to harvest and manage multiple data sources. For that task, additional services are available: profiling of the ingested data sets, "business intelligence" (customized data profiling), data transformation for exporting (by OAI or many other interfaces) and search services (REPOX can be deployed with a built-in integration with Solr[4], offering a service for search and retrieval of the harvested data sets).

Aggregators are a special class of OAI service provider, responsible for data harvesting processes from original providers, in order to expose these sets as a data provider to another "upper level" of harvesting. This concept is, for example, core in Europeana, where aggregators

[1] Europeana - http://dev.europeana.eu/

[2] TEL - The European Library- http://www.theeuropeanlibrary.org

[3] EuDML– European Digital Mathematics Library- http://www.eudml.eu/

[4] http://lucene.apache.org/solr/

are expected to represent countries, regions, or sectorial data providers (specialized archives or museums, etc.).

Besides all the use cases available to a Service Provider, to re-expose the records an Aggregator can also use the same cases available for Data Providers, as described in the Section 2 (in fact, an Aggregator also is a Data Provider, in this sense).

3. Problem Description

The main objective of this work is to improve the matching process in order to obtain more accurate mappings between schemas while expending less effort. User aided matching tools[5] are interfaces that help the user match the elements graphically. The problem with these tools is that the user needs to spend a huge effort (depending on the size of the schema and collections, we are assuming large collections and schemas) to learn the schemas and datasets in order to take the right decisions while matching the elements.

Automatic matching tools[6] (Rahm, 2001) (Giunchiglia, 2004) (Madhavan, 2001) implement algorithms that guess the best possible approximation of the mapping, so this mapping needs to be evaluated by a human, but still, that human has to spend a huge effort to know the schemas and datasets involved. The ideal solution should be an automatic method, as accurate as the manual method; the nearest method to accomplish this is a machine learning method (Berlin, 2002) (Doan, Madhavan, Domingos, & Halevy, 2004). But, these methods do not completely avoid the human effort in the process because they are not a hundred percent accurate. Therefore, the solution we propose is basically applying data profiling to the manual matching process, helping the user in his quest for knowledge about the schemas, datasets and compliance between them. Accordingly, we propose to develop some relevant measures to be calculated in order to provide the user with useful information for the mapping process. The actual assumption is that these measures might be better defined if we successfully segment the problem as shown in
Table 1.

The goal for this work is accomplished by the determination, study and test of measures to fulfill the cases described in the table above, integrating them in REPOX.

Table 1: Scenarios for schema matching

	Sample of only one data set available	**Samples of source and destination data sets available**
No schemas available	Case A - profile the data set to support further simple human reasoning and action based only on that	Case B - profile the two data sets and verification of attribute correspondence between the sets to support further human reasoning and action based on the comparing of these results
One schema available	Case C – Same as Case A, plus profiling of the schema usage completeness	Case D - Same as Case B, plus profiling of the extension of the schema usage (data sets not always use all the schema attributes, …)
Two schemas available	*Not Considered*	Case E - Same as Case D, now considering the extension of the usage of the two schemas-

References

Berlin, J. a. (2002). Database schema matching using machine learning with feature selection. *Proceedings of the 14th International Conference on Advanced Information Systems Engineering* (pp. 452--466). Springer-Verlag.

Doan, A., Madhavan, J., Domingos, P., & Halevy, A. (2004). Ontology matching: A machine learning approach. In *Handbook on Ontologies*, Springer-Verlag, ISBN 3-540-40834-7, 385--516.

[5] http://www.altova.com/mapforce.html, http://www.microsoft.com/biztalk/en/us/default.aspx, http://www.almaden.ibm.com/cs/projects/criollo/

[6] http://www.telusplanet.net/public/bmarshal/dataqual.htm

Freire, N., Borbinha, J. (2006). Metadata spaces: The concept and a case with REPOX. In Julio Gonzaloet. all, editors, Research and Advanced Technology for Digital Libraries, volume 4172 of Lecture Notes in Computer Science, pages 516–519. Springer Berlin / Heidelberg, 2006

Giunchiglia, F., Shvaiko, P., Yatskevich, M. (2004). S-match: an algorithm and an implementation of semantic matching. *Proceedings* of the First European Semantic Web Symposium (*ESWS'04*), Heraklion, Greece, May 10-12 The semantic web: research and applications. LNCS volume 3053, 61--75.

Madhavan, J.,Bernstein, P., Rahm, E. (2001). Generic Schema Matching with Cupid. In *Proceedings of the 27th International Conference on Very Large Data Bases* (VLDB '01), Peter M. G. Apers, Paolo Atzeni, Stefano Ceri, Stefano Paraboschi, Kotagiri Ramamohanarao, and Richard Thomas Snodgrass (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 49-58.

Rahm, E. a. (2001). A survey of approaches to automatic schema matching. (Springer, Ed.) *the VLDB Journal , 10*, pp. 334--350.