RDF rules for XML Data conversion to OWL Ontology

Christophe CRUZ, Christophe Nicolle
Laboratory Le2i
Université de Bourgogne
B.P. 47870, 21078
Dijon CEDEX, France
{christophe.cruz, cnicolle}@u-bourgogne.fr

Keywords: Ontology population, ontology enrichment, OWL ontology, XML data, RDF rules.

Abstract: The paper presents a flexible method to enrich and populate an existing OWL ontology from XML data based on RDF rules. Theses rules are defined in order to populate automatically the new version of the OWL ontology. Basic rules are defined to identify elements in XML schemas and an OWL schema. Advanced mapping rules are based on basic rules in order to define the mapping between XML schemas elements and OWL schema elements. In addition, this flexible method allows users to reuse rules for other conversions and populations.

1 INTRODUCTION

The knowledge defined in ontologies is used as an index to retrieve specific data (García, 2005), to infer new knowledge (Ha, 2005), to semantically annotate multimedia data (Castano, 2007), to find out Web Services automatically (Martin, 2007), or to match knowledge with other knowledge for a more general purpose.

XML schemas contain the knowledge of a domain that was specified by the author. This specification is only syntactic without any semantic definition. This is due to the fact, that XML data are used to exchange data between processes that were developed for this data. In order to permit the exploitation of the knowledge contained in XML schemas and instances, we propose an ontology enrichment and an automatic population process from XML data based on a manual mapping of XML schemas. The result of this process allows the use of a SPARQL engine to request data on the resulting OWL ontology, and allows the use of an inference engine in order to deduce new knowledge. In addition, multiple XML schemas and XML documents are integrated in the OWL ontology giving a single view on data.

Ontology enrichment is the activity of extending an ontology by adding new elements (e.g. concepts, relations, properties, axioms) (Castano, 2007). The enrichment process consists in annotating knowledge contained in XML schemas in order to convert it into an OWL schema (Faatz, 2004). The annotation process is manual. This is done by the user who is the only one who knows which part of the XML schema will be required in future processes. However, Schema matching is a manipulation process on schemas that takes two heterogeneous schemas as input and produces as output a set of mapping that identifies relations between the elements of the two schemas (Thang, 2008). An automatic matching process is of value in order to help in a semi-automatic way the annotation of XML schemas and for rough enrichment.

The ontology population process is the activity of adding new instances to an ontology (Castano, 2007). Presently, the data in XML document are converted in OWL instances automatically with rules generated from the previous “XML schema annotation” process.

The following section discusses about previous work done on this subject. Section three presents the principles of our method which is based on formal languages. Section four shows our method to enrich the ontology by matching schemas and to populate an ontology by an automatic process. A complete background concerning XML Data conversion to OWL ontology is described in (Cruz, 2008).

3 PRINCIPLE

The principle of our solution consists in making RDF stand-off annotations and RDF links between the schematic level of an OWL schema and the

schematic level of several XML schemas. As a consequence, the enriched ontology makes it possible to link the concepts of several XML schemas by amalgamating the attributes of common concepts. This process has also to check the consistency of the schema ontology in order to not define several identical concepts and to not allow the definition of cyclic "*rdfs*:*subClassOf*" graphs.

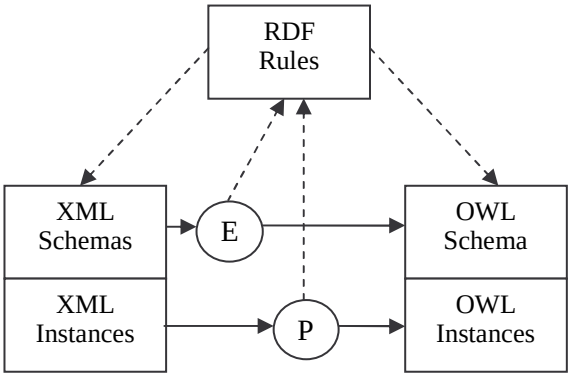

Figure 1: Principle of our "XSD to OWL" method.

The annotations and the links are used in a second time to defined rules in order to automatically populate the OWL within instances from XML instances from XML schemas annotated. The population has to follow some rules such as the imitation of attribute cardinalities and unique instances. Consequently, "advanced" rules have to model and specify restriction on attributes. This principle will be described in the next section.

The figure 1 presents the principle of our method to (E)nrichment and (P)opulation of an existing ontology. (E) is the process of enriching an ontology and (P) is the process of populating the ontology. These two processes use an RDF graph as rules to enrich and to populate the ontology. The rules in RDF are defined during the mapping process by referring elements of XML schemas elements and the OWL schema. In order to specify the relevant elements of an XML schema for the enrichment process, it is necessary to identify and mark these elements. These marks are called "schematic marks" and are external RDF annotations of XML structures. These "schematic marks" are stand-off annotations or external annotations.

The following section presents a brief formalization of schematic marks.

3.1 Schematic Marks on XML schema

The properties of Dyck's languages were the subject of studies undertaken by J. Berstel (Berstel, 2000). According to the corollary 3.4, for each XML language *L* there is only one reduced XML grammar generating *L*. A reduced grammar does not have any useless non terminal vocabulary. An XML schema does not contain unnecessary tags, so an XML schema does not use unnecessary non terminal vocabulary. Consequently, an XML grammar is necessarily its own reduced grammar. In addition, it means that only one production rule in the XML schema (tag <xsd:element name="myTag" >) is define for a tag in a XML document validated by the XML schema. This proposal makes it possible to introduce the concept of schematic mark.

Definition: A "schematic mark" is a mark on an XML schema that identifies a production rule. Each tag of an XML instance which has the same name was produced by the same production rules.

These marks are used by the RDF rules to identify the production rules which are the tags used to define XML schemas. These marks are specified with the help of the language XPath.

3.1 Schematic Marks and semantic

An OWL grammar is also an XML grammar which can be marked as an XML schema. However, an ambiguous point has to be underlined. In figures 1 and 2, the OWL schema concerns the schema part of the ontology. Indeed, the language OWL has an XML schema in order to validate OWL documents (http://www.w3.org/2007/OWL/wiki/OWL_XML_Schema). In this section we focus on the OWL schema part in an OWL document. The "semantic marks", which are used by the RDF rules to identify the production rules in a XML schema, use the language XPath to identify element of the OWL schema.

OWL uses most of the built-in XML schema data types. References to these data types are by means of the URI reference for the data type, http://www.w3.org/2001/XMLSchema.

4 ENRICHMENT AND POPULATION METHOD

This section describes how the enrichment and the population of an OWL ontology are managed from XML schemas. The method is based on the definition of schematic marks, basic mapping rules and advanced mapping rules (e.g. fig. 2). The first part describes the schema marking in order to annotate the element of XML schemas. It relates to the XML schema but also to the OWL XML schema. The second part presents the mapping step which is composed of the conversion rules, the

ontology enrichment process and the ontology population process.

4.1 Schema marking

An RDF graph is used to annotate each XML schema. These marks are specified to keep all information in a graph that is required during the mapping with the OWL schematic mark step.

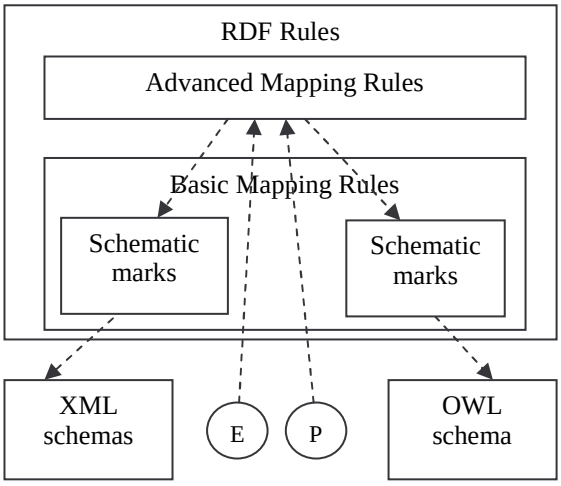

Figure 2: Principle of the RDF Rules definition.

In this figure, the processes (E) and (P) are here to show the relationships established with the RDF rules. In fact, the main objective of the figure is to describe the components of the RDF rules. They are composed of the schematic marks on XML schemas and on an OWL schema. These marks are used to identify elements required for the mapping process.

- Basic mapping rules define the schematic marks on XML schema and OWL schema. The "bmr" name space is used to identify the rule elements. In addition, "bmr:xpath" defines an element in the XML schema as does "bmr:xpathId" with the addition that the latter acts as an identifier. A unique property on an XML element is welcome to define an identifier.
- Advanced mapping rules use basic mapping rules in order to define the mapping between XML schemas and OWL. In addition, these rules allow to define new elements in the OWL ontology for the enrichment process and for the population process. The name space "amr" represents the "advanced mapping rules" elements.

The TriG Syntax (http://www4.wiwiss.fu-berlin.de/bizer/TriG/) is used to ease the explanations of how we employ schematic marks with XML schemas.

4.2 Mapping steps

This step consists in defining a new RDF graph in order to define relationships between XML schema marks and OWL marks.

Conversion rules

The conversion rules consist in defining rules in order to convert properties that are different from the type of the property in the ontology and that cannot be directly copied to the ontology "datatypeProperty". The first kind of conversion is simple because it is the conversion of a simple type into another simple type. The second kind of conversion is complex because it is the conversion of a sub tree from the schema into a simple type into the OWL ontology.

The conversion of complex data can also be found in a semi-structured data. For instance, a date can be defined in a text format and could have to be converted in month, day and year format in the ontology.

Ontology enrichment

The enrichment consists in defining relationships between XML schematic marks and OWL schematic marks. In order to enrich the ontology, the RDF mapping graph has to contain information about new entities in the ontology that have to be. The name space "amr" represents the "advanced mapping rules" elements.

Ontology population

In order to generate the population of the ontology, the RDF mapping has to be defined. To achieve this, relationships have to be created between the RDF graph of the XML schema marks, the RDF graph of the OWL schematic marks and the RDF graph of the conversion rules.

The population of the ontology is an automatic process based on the mapping graphs. To realize this process, we have defined an algorithm that takes into account the type bmr:dpId in order to avoid duplicated instances of the ontology. First, it determines all classes that have to be populated. Second, all "datatypeProperty" of each class is provided to the instances. In the example given in this paper no references are given to the management of restriction on the properties. Some rules can be defined in order to specify which constraints have to be verified. If those rules are not defined then no check on restriction is applied.

4 CONCLUSIONS

We have presented a flexible method to enrich and populate an OWL ontology for the integration of XML data. Basic mapping rules and advanced mapping rules are defined by users and can be reused for other conversions and populations of ontologies. This conversion is the first part of our work. The second part consists in improving the process and in giving some proposition to the user in order to facilitate the mapping. The RDF rules can be used to automatically extract from XML schemas some elements that can be converted in order to help users during the mapping. For instance, a string that contains a date can be detected automatically to guide the user during the conversion.

REFERENCES

Castano, S., Espinosa, S., Ferrara, A., Karkaletsis, V., Kaya, a., Melzer, S., Moller, R., Montanelli S., Petasis, G., 2007. Ontology Dynamics with Multimedia Information: The BOEMIE Evolution Methodology. In *Proc. of International Workshop on Ontology Dynamics (IWOD) ESWC 2007 Workshop* - 7 June - Innsbruck, Austria, 2007.

Cruz, C., Nicolle, C., Ontology Enrichment and Automatic Population From XML Data, 4th ODBIS Workshop on Ontologies-based Techniques for DataBases in Information Systems and Knowledge Systems, Co-located with VLDB 2008, 2008.

Martin, D., Paolucci, M., Wagner, M., 2007, Towards Semantic Annotations of Web Services: OWL-S from the SAWSDL Perspective, In *OWL-S Experiences and Future Developments Workshop at ESWC 2007*, June, Innsbruck, Austria.

García, R., Celma, O., 2005. Semantic Integration and Retrieval of Multimedia Metadata, *Proceedings of 4rd International Semantic Web Conference*, Galway, Ireland.

Do, H.H., Rahm, E., 2002, COMA - A System for Flexible Combination of Schema Matching Approaches, *Proc. 28th Intl. Conference on Very Large Databases (VLDB)*, Hongkong, Aug.

Aumueller, D., Do, H.H., Massmann, S., Rahm, E., 2005, Schema and ontology matching with COMA++, *SIGMOD Conference*.

Thang, H. Q., Nam, V. S., 2008, XML Schema Automatic Matching Solution, In *International journal on Information Systems Science and Engineering*, vo.l 4, number 1.

Ferdinand, M., Zirpins, C., Trastour, D., 2004. Lifting XML Schema to OWL, in: *Koch, Nora and Fraternali, Piero and Wirsing, Martin (Hrsg.): Web Engineering - 4th International Conference, ICWE 2004*, Munich, Germany, July 26-30, 2004, Proceedings, Springer Heidelberg, pp. 354-358.

Bohring, H.; Auer, S.: Mapping XML to OWL Ontologies. *Leipziger Informatik-Tage* (LIT 2005), Sep. 21-23, 2005, *Lecture Notes in Informatics* (LNI).

Rodrigues, T., Rosa, P. and Cardoso, J., 2006, Mapping XML to Exiting OWL ontologies, In *International Conference WWW/Internet 2006*, (Eds) Isaías, Pedro and Nunes, Miguel Baptista and Martínez, Inmaculada J., pp.72-77, ISBN:972-8924-19-4.

Anicic, N., Ivezic, N. and Marjanovic, Z., 2007, *Mapping XML Schema to OWL*, Enterprise Interoperability, Springer London.

Bowers S, Delcambre L., 2000. Representing and Transforming Model-Based Information, In *Proceedings of the Workshop on Semantic Web at ECDL-00*, Lisbon, Portuga.

Berstel, J., Boasson, L., 2000, XML Grammars, *MFCS 2000*: 182-191.

Faatz, A., and Steinmetz, R., 2004. Precision and recall for ontology enrichment. In *Proc. of ECAI-2004 Workshop on Ontology Learning and Population*, Valencia, Spain, Aug.

Ha, Y., Sohn, J. , Cho, Y., 2005. OWLer: a semantic web ontology inference engine, In *Advanced Communication Technology*, 2005, ICACT.

Troncy, R., Celma, O., Little, S., Garcıa, R. and Tsinaraki C., 2007. MPEG-7 based Multimedia Ontologies: Interoperability Support or Interoperability Issue? In 1st International Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies, pages 2–15.