Tagging for Improved Semantic Interpretation of XML

Vicente Palacios, Juan Lloréns, Sonia Sánchez-Cuadrado and Mónica Marrero
University Carlos III of Madrid
Department of Computer Science

palacios@di.uc3m.es  llorens@inf.uc3m.es  sscuadra@bib.uc3m.es  mmarrero@inf.uc3m.es

ABSTRACT
The development of the Semantic Web depends on agreed and unambiguous knowledge representations, on the availability and accessibility of knowledge, and on retrieval capabilities. The scarce agreement on knowledge representation and the lack of techniques to process semantic structures in web search engines makes it impossible to perform contextualized conceptual retrieval. These limitations imply that users must know beforehand the existence and location of this knowledge to be able to retrieve it. Thus, different ad-hoc knowledge representations and metadata vocabularies, scarcely formalized and agreed on, have been published, hindering the reuse and interoperability. Our proposal has as its main goal the elaboration of heterogeneous semantic schemas by means of a multilevel ontological structure and the alignment with a reference ontology that provides conceptual retrieval and reuse of knowledge.

Categories and Subject Descriptors

General Terms: Standardization, Languages

Keywords: Semantic Web, Interoperability, Knowledge Representation, Metadata Vocabularies.

1. INTRODUCTION
The Semantic Web was originally devised as a solution to improve the interoperability among applications, rather than its use by humans. This circumstance has limited the possibilities of web applications for information retrieval, question-answering and inference-based reasoning. Such limitations have to do with the technology in use, the number and quality of metadata, and the size of the Web [1]. It is our belief that the second point is the main drawback for the spreading of the Semantic Web proposal. This idea is supported by Buscaldi’s findings when he stated that the correct interpretation of semantic elements in schemas (like XML schemas and metadata vocabularies) can be assessed only in 40% of the cases [2]. Some of the problems regard dummy names (e.g. “label1”), tags with unintelligible abbreviations, ambiguous and non-contextualized words like stop words, or even made up words. In short, these problems reflect classical questions in natural language, like polysemy, synonymy or idiomaticity.

The large number of local semantic resources worsens the problem. Although one of the objectives that the Semantic Web Proposal has is to promote a shared terminology, an analysis of the metadata vocabularies shows the lack of consensus. As an example, we can find many vocabularies representing the same terms, like with Zthes [3], Thesaurus’ PSI [4] or SKOS [5].

One of the simplest ideas to obtain a smaller and normalized set of tag names is to perform mappings between the semantic elements provided by different resources. Many solutions to this matching problem have been proposed [6], from areas like databases, information systems, artificial intelligence, semantic web, etc.

For the XML vocabulary formalism, the one-to-one solution has been widely applied in the literature [7][8]. However, this approach is problematic due to scalability, because one-to-one mappings among a numerous set of documents require multiple mappings each time a new document is added [9]; and incomplete mappings due to the degree of specificity within a domain (some of the concepts present in a schema may be absent in another one, either because of their high degree of specificity or because they belong to different domains). Other problems of aligning elements are that they imply a previous acknowledgement of their existence and the understanding of every element’s meaning. This is not a trivial task. On the one hand, the semantics of the elements is rarely explicit in the resource. On the other hand, these documents are designed to exchange data between software applications, but with low usability.

The best solution seems to be the conversion of information into a format which is more easily understandable by software, that is to say, which carries a richer load of codified semantics. This representation conforms a semantic layer applicable to metadata and XML language.

2. THE SEMSE PROJECT
The SEmantic Metadata SEarch proposal (SEMSE) incorporates this semantic layer to the representation of metadata schemas [10].

This layer avoids ambiguous conceptual representations and improves conceptual retrieval techniques based on the semantics and the content of the schema elements. First, we created the semantic structure for information representation and retrieval. Next, the alignment process between specific and reference ontologies is carried out. To achieve these goals, the following activities were defined:

a. Evaluation of the reference ontology. We analyzed existing foundation ontologies and their relevance to the requirements of our proposal. This study aimed at deciding between the development of a new application ontology or the reuse of an existing one. We opted for an intermediate solution with the PROTON ontology (http://proton.semanticweb.org).

b. Evaluation of the mapping method. This activity defined the proper mapping method between the schemas and the reference ontology.

c. Development of the management subsystem for semantic schemas and reference ontology. It also provided the control mechanisms for updating, monitoring and reviewing schemas.

d. Search and selection of the initial set of schemas to be included in the system as the kick off kit. The selection criteria regarded popularity, scope and overlap of elements.

Copyright is held by the author/owner(s).

ESAIR ’10, October 30, 2010, Toronto, Ontario, Canada.
ACM 978-1-4503-0372-9/10/10.

e. Alignment between the schemas and the reference ontology. This activity is in charge of performing the alignment between schemas and the reference ontology.

The solution would be to split the documents in two different views (see Figure 1). The Qualified Scheme is a normalized formalization in RDF of the metadata vocabulary. It also allows including the semantics of each element (‘a’ in the figure), annotated using the hasSemantics property, against the concepts represented in a specific ontology (‘A’ in the figure). The Specific Ontology is a representation specialized in semantics. Formulated in OWL-Full, it includes definitions, synonyms and multilingual support for each element. Adding the information in this document avoids modifying the original resource.

![Figure 1. Artifacts and relations in SEMSE proposal.](image)

These specialized views combine the improved representation with the improved semantics management through the Qualified Scheme and the Specific Ontology, respectively. Once semantically represented, the proposal takes a step forward aiming interoperability among schemas. PROTON establishes a common vocabulary, provides a common semantics and support semantic disambiguation. It avoids the one-to-one mapping and improves the interoperability among schemas. The correspondence between the semantic representation of the schemas and the backbone ontology is carried out by an independent alignment ontology which facilitates the syntactic and semantic extensibility, adaptability and reuse of the concepts’ correspondences.

3. RESULTS

Various foundation ontologies have been evaluated in order to establish the degree of compatibility with our requirements and the cost of use versus the development of a resource of our own. As result, we adopted a mixed approach. In a non exclusive manner, PROTON has been chosen as foundation ontology and it has grown incrementally as new semantic schemas have been included through an application ontology that allowed its extensibility, adaptability and reuse.

In this stage, an automatic process would generate errors because the concepts are not explained in the schemas. Also, concepts may not have an equivalent in other schemes or ontologies, and so a new concept would have to be proposed. The flexible nature of XML, the unexplained concepts in the schemes, and the lack of strict scheme enforcement are some reasons why we require an initial manual alignment approach of concepts between closely accepted and related schemas. We performed a manual alignment between concepts to minimize mapping errors. It’s been identified as a key factor by the impact that a simple error in a schema element has in thousands of semantic documents generated from that schema. This mapping has been resolved using an independent mapping ontology. Therefore, specific and reference ontologies are independent of relationships between concepts, which eases the maintenance, expansion and modification in the ontological base. As a result, six metadata vocabularies to represent personal data have been tested: DC metadata element set, vCard, FOAF, DOAC, DOAP and PIM. The output is six schemas with a total of 289 elements with a higher degree of syntactic and semantic formalization.

The first tests of the proposal in the retrieval area have thrown more precision and recall compared with syntactic and semantic search engines. It also has shown how it’s possible to expand the result set, including semantically related concepts from a search concept and a meaning. The end of the project’s development process will enable a more exhaustive quantitative evaluation.

4. CONCLUSIONS

In this study we analyzed the ontological representation of schemas. The problem is twofold: stating the schemas representation structure to allow disambiguating semantic complexity; and reduce the structural ambiguity. The simplicity to represent semantics in a single document, with scarce information about the semantic and the meaning of the tags, deteriorate the performance of tagging solutions. We propose a simplification in updating semantic resources, reducing dependencies and facilitating extensibility. These are mandatory requisites in systems that try to collect a large number of semantic resources. An integrated search system could thus retrieve different schemes with just one query thanks to the semantic links between their concepts. Therefore, it is not necessary to know the existence and location of such knowledge beforehand.

ACKNOWLEDGMENTS

We acknowledge the National Plan of Scientific Research, Development and Technological Innovation, that has funded this work through the research project TIN2007-67153.

REFERENCES