Tagging for Improved Semantic Interpretation of XML

Vicente Palacios, Juan Lloréns, Sonia Sánchez-Cuadrado, Mónica Marrero
University Carlos III of Madrid, Department of Computer Science
palacios@di.uc3m.es  llorens@inf.uc3m.es  ssccuadra@bib.uc3m.es  mmarrero@inf.uc3m.es

Problem
Many of the existing semantic vocabularies have a very scarce formalization, present structural and semantic ambiguities, and lack of usage recommendations

These limitations imply that users must know beforehand the existence and location of this knowledge to be able to retrieve it

Objective
Elaboration of different views to improve the representation of semantic documents

The approach facilitates the management and retrieval of heterogeneous semantic schemas by means of a multilevel ontological structure and the alignment with a reference ontology

It facilitates conceptual retrieval and reuse of knowledge

SEmantic Metadata SEarch proposal (SEMSE)

1. Evaluation of the reference ontology. We have analyzed different high-level ontologies (WebKB-2, OpenCyc, SIOC, Umbel, Dolce, Sumo and Proton, besides WordNet) in terms of maintenance, interoperability, complexity, and Web compatibility

2. Evaluation of the mapping method

3. Development of the management subsystem for semantic schemas and reference ontology

4. Search and selection of the initial set of metadata vocabularies (with different scope, degree of popularity, stability and formalization)

5. Alignment between the schemas and the reference ontology

Evaluation
Comparison between SEMSE and semantic directories and engines, following the DESMET methodology. Figures in percentages, (SEMSE gets 100%)

Criteria: interoperability, disambiguation, multilingualism, reusability, formalization and conceptual retrieval

Conclusions
- We have disambiguated and formalized 289 elements from vocabularies of metadata, further referenced in several millions of documents in the Web
- We have aligned the elements between the SEMSE reference ontology and the metadata vocabularies, avoiding an all-to-all mapping
- We generate different views of the vocabularies facilitating the maintenance of complex schemas. The original vocabulary is not modified though
- Evaluation: compared to other systems, SEMSE offers advantages in terms of interoperability, disambiguation, multilingualism, reusability, formalization and conceptual retrieval