Support of SKOS Vocabularies in the DSpace Digital Repository System

Georgia D. Solomou and Dimitrios A. Koutsomitropoulos

High Performance Systems Laboratory, School of Engineering, University of Patras,
Building B, 26500, Patras-Rio, Greece
{ solomou, kotsomit}@hpclab.ceid.upatras.gr

Abstract. DSpace offers the possibility to characterize its items using a predefined set of keywords, namely a controlled vocabulary. A DSpace compliant controlled vocabulary is represented in a simple XML format. A more interoperable and machine-understandable way, though, for expressing controlled vocabularies is the SKOS data model. SKOS provides a standard way to represent knowledge organization systems using RDF. The support of SKOS in DSpace is implemented through an add-on, provided by the University of Minho. First, we re-implemented this add-on and we finally applied it to the University of Patras live DSpace installation. We then experimented by importing a real SKOS vocabulary: the thesaurus of Greek Terms. As a final step, we tried to tackle with arising problems and to propose solutions, relying on the semantic web techniques.

Keywords: controlled vocabularies, thesauri, SKOS, reasoning, OWL

1 Controlled Vocabularies in DSpace

DSpace supports controlled vocabularies so as to confine the set of keywords that users can use while describing, searching or browsing items. These keywords are organized in a tree (taxonomy) which appears during the search and submission process.

Supported controlled vocabularies are expressed in a simple XML format (aka "DSpace node schema"). According to this schema all information about a term is enclosed in a `<node>` element. Only the expression of a hierarchical (narrower in meaning) relationship is allowed through the use of the `<isComposedBy>` sub-element. Furthermore, by using `<hasNote>` a simple annotation mechanism becomes possible.

The support of multilingual controlled vocabularies is a feature with which we have further enhanced the DSpace system. Each vocabulary's translation is included in a separate file, named by the name of the vocabulary and augmented by the appropriate language code. In order to implement this facility we had to modify the `ControlledVocabularyTag.java`.

2 Support of SKOS in DSpace

The *Simple Knowledge Organization System* (SKOS) [4] is a data model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, taxonomies and other similar types of controlled vocabularies. It is actually a practical application of RDF [5] (and RDFS) and its main objective is to enable easy publication of controlled structured vocabularies for the Semantic Web. SKOS is now a proposed recommendation of W3C, hence a near-to-become standard.

The Odisseia Research at the University of Minho in Portugal has implemented an add-on for version 1.4.2 of DSpace [2] which augments this system with the ability to support controlled vocabularies expressed in SKOS. In particular, the provided add-on makes the following changes:

- Enhances the DSpace inherent node schema so as to support more types of relationships and properties
- Allows for the recognition of thesauri (controlled vocabularies) expressed in the RDF/XML serialization syntax of the SKOS format
- Offers the ability to assign a different vocabulary to each community

The updated version of DSpace node schema allows for the use of *related* and *preferred* (use-instead) terms. The add-on is responsible for correctly rendering this kind of terms in the constructed HTML node tree (taxonomy). It actually affects the transformation process carried out by the `vocabulary2html.xsl` and `vocabularyprune.xsl` files.

In the case of handling controlled vocabularies expressed in SKOS, an additional XSL transformation file is provided (`vocabularySKOS2node.xsl`). The transformation applies to the SKOS vocabulary and takes place a step before the taxonomy construction process (see Figure 1). It actually parses the RDF/XML format in which the SKOS vocabulary is expressed and produces a valid DSpace node schema.

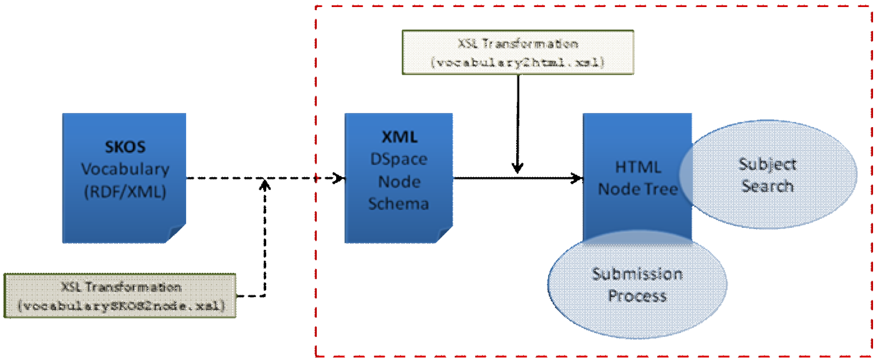

Figure 1. The vocabulary ingestion process

However, during the parsing process we realized that the given SKOS-to-node XSL transformation was behaving problematically (see also section 3.1). The reason was that this transformation was coping with only those narrower terms described also as stand-alone concepts. This issue was fixed by slightly modifying the given XSLT file. Another problematic case – that still remains unresolved – is the appearance of repetitions or the absence of some terms from the constructed taxonomy.

Finally, this add-on augments DSpace with the possibility to assign a different vocabulary to each community. This feature is implemented by relating the unique handle of the community with the file name of a specific vocabulary. The correlation is asserted in the `input-forms.xml` file.

3 Appling the Controlled Vocabulary Add-on to the University of Patras Institution Repository

The University of Patras institutional repository is based on a DSpace – version 1.4.2 – installation. It is used for managing and distributing the university's educational and scientific material. In order for this repository to support SKOS thesauri, we re-implemented and adopted the aforementioned controlled vocabulary add-on. Not all its supplied features were applied though. In particular, the parameterization of the submission metadata fields, so as to handle different vocabularies for each community was excluded. This facility can be implemented in DSpace in another way: by declaring one separate `<form-definition>` for each community inside the `input-forms.xml` file.

The incorporation of a real SKOS thesaurus in the DSpace system has shown that the transformation of a vocabulary from the SKOS format to the simpler DSpace node schema resulted in several problems. A brief description of these problems – mainly reflected in the construction of the taxonomy – is given below where a real SKOS thesaurus is imported in the DSpace installation of the University of Patras.

3.1 The Thesaurus of Greek Terms in DSpace

The National Documentation Center in Greece (EKT) has published the first thesaurus of Greek Terms. This thesaurus is a controlled vocabulary composed of 5227 bilingual (Greek, English) terms covering a broad field of knowledge.

After we had implemented the SKOS format of the EKT thesaurus, we made an attempt to incorporate it in the University of Patras institutional repository. During this process we faced several problems, mainly concerning the construction of the HTML node tree:

- Some terms appeared in the wrong place of the taxonomy (wrong depth level or repetitions of terms)
- A number of terms, although present in the SKOS file, were missing from the tree hierarchy

We found out that the main cause of these problems lies mostly in the following:

- The inability of the provided XSL transformations to handle every possible relationship among terms (e.g. there is no provision for broader in meaning terms).

- The non-exhaustive (but not semantically inconsistent) implementation of the Greek Terms thesaurus in which not every possible relationship is asserted.

As a consequence, a few modifications had to be made in order to fix some of these issues. Actually, our concern was focused on the provided XSL transformation which had to be altered so as to take care for two more kinds of relationships: for both broader and non-asserted ones.

However, a possible and effective solution would be to consider the thesaurus of the Greek Terms thesaurus as an ontology. Besides, the SKOS is by itself defined in the Web Ontology Language (OWL [1]). Such a consideration could allow for a programmatic access to the thesaurus elements. In particular, by exploiting a corresponding API, as for example the OWL API [2] or even the newly appeared SKOS API [7], a simpler way to construct the vocabulary's node tree would be possible. This comes in contrast to the more complex and not always feasible approach offered by the XSL transformation. What is more, the new constructs offered by the updated version of OWL (OWL 2) would allow for the expression of richer semantic conditions in SKOS, as stated in [6].

Another gain in handling the SKOS thesaurus as an OWL ontology, is the possibility to apply OWL reasoners (like FaCT++ and Pellet). OWL reasoners allow for inferencing and thus for the automatic handling of non-asserted relationships. Consequently, an inferenced-based classification and rendering of the thesaurus can be achieved resulting in a complete and consistent tree hierarchy.

We believe that it would be worth investigating the potential of these ideas for controlled vocabularies in DSpace and we are currently working in this area.

References

[1] Bechhofer, S., Harmelen, V.F. , Hendler, J., Horrocks, I., McGuinness, D., Patel-Schneider, P., and Stein, L. (2004). "OWL Web Ontology Language: Reference". W3C Recommendation. http://www.w3.org/TR/owl-ref/

[2] Costa, S., Ferreira, M., and Alice, A., (2007). Controlled-Vocabulary Add-on Patch for DSpace 1.4.2. http://sourceforge.net/tracker/index.php?func=detail&aid=1833347 &group_id=19984&atid=319984.

[3] Horridge, M., Bechhofer, S., Noppens, O., (2007) "Igniting the OWL 1.1 Touch Paper: The OWL API." In Proc. of the OWL Experiences and Directions Workshop, Innsbruck , Austria.

[4] Issaac, A., and Summers, E., (eds) (2009). "SKOS Simple Knowledge Organization System Primer". W3C Proposed Recommendation. http://www.w3.org/TR/2009/WD-skos-primer-20090615/

[5] Klyne, G., and Carroll, J. J., (eds) (2004). "Resource Description Framework (RDF):Concepts and Abstract Syntax". W3C Recommendation. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

[6] SWIGroup HPCLab (2009). SKOS in OWL 2. http://apollo.hpclab.ceid.upatras.gr: 8001/ hpclabwiki/Skos2Owl2

[7] The SKOS API. http://skosapi.sourceforge.net/