SKOS and the Ontogenesis of Vocabularies

Joseph T. Tennis
The University of British Columbia
Tel: +01 604 822 24321
Fax:+01 604 822 6006
jtennis@interchange.ubc.ca

Abstract:

The paper suggests extensions to SKOS Core to make explicit where concepts in a knowledge organization system have changed from one version of the system to another.

Keywords:

controlled vocabularies, SKOS, versioning, version tracking, vocabulary ontogeny, information retrieval, RDF.

1. Introduction

The Simple Knowledge Organization System (SKOS) Core Guide is a working draft of the W3C. It outlines methods for "expressing basic structure and content of concept schemes (thesauri, classification schemes, subject heading lists, taxonomies, terminologies, glossaries and other types of controlled vocabulary)" in RDF [1]. It outlines two mechanisms for concept scheme revision: (a) notes and (b) OWL versioning. As it stands an editor of a concept scheme can make notes or declare in OWL that more than one version exists. This paper adds to the SKOS Core by introducing a tracking system for changes in concept schemes. We call this tracking system *vocabulary ontogeny*. Ontogeny is a biological term for the development of an organism during its lifetime. Here we use the ontogeny metaphor to describe how vocabularies change over their lifetime. Our purpose here is to create a conceptual mechanism that will track these changes and in so doing enhance information retrieval and prevent document loss through versioning.

1.1 Vocabulary

In order to illustrate vocabulary ontogeny, we use the metadata thesaurus [2, 3, 4] used for the Dublin Core Online Conference Proceedings. This vocabulary is revised each year in order to faithfully represent the content of the proceedings. It has been revised three times to date (2002-2004). However, none of the documents indexed with the older versions are re-indexed with the revised version of the vocabulary. Each year then, is indexed using its own expanded version of the vocabulary.

1.2 Retrieval Problem

Because the metadata thesaurus undergoes constant revision, it is unstable and cannot provide fixed relationships between indexing terms (concepts) and the entire collection of Dublin Core Online Conference Proceedings. For example, a paper indexed in 2002 will not be re-indexed with the revised index terms with the papers for 2004. However, the purpose of a controlled vocabulary is to collocate documents on the same subject. In order to accomplish this task, a secondary mechanism is required. We need a mechanism to express relationships of similarities and dissimilarities across the different versions. This mechanism would chart the development (ontogeny) of the metadata thesaurus and in doing so provide a structure for identifying similar and dissimilar terms across all versions of the thesaurus.

2.SKOS

According to the deprecated guide [5] SKOS stands for Simple Knowledge Organisation System. Perhaps a better way to think of it—according to its purpose—is as the Schema for Knowledge Organization Systems. The SKOS Core Guide [1] suggests how to track revisions and versions. The Guide is in *Editor's Working Draft* form, so the suggestions it presents stand as first thoughts on the matter and not final recommendations. We will use the SKOS Core

Guide's suggestions as starting points to address our example of vocabulary ontogeny. It outlines two suggestions for tracking revisions: (a) notes and (b) OWL versioning. We outline both suggestions below.

2.1 Notes in SKOS

The Guide [1] offers two types of thesaurus editor notes:

skos: historyNote
skos:changeNote

The historyNote is a note for the users of the concept scheme. The historyNote documents a significant change to the meaning, form, and or state of a concept. SKOS does not provide an example for this note.

The changeNote serves both the editor and indexers using the thesaurus. It is a private note, not intended for users, that documents "fine-grained changes to the concept for the purposes of administration and management" [1]. The example given is a change in labeling a concept from "laptop computers" to "notebook computers" [1].

2.2 OWL Versioning in SKOS

In order to signal a change from one version to another, SKOS suggests using OWL, Web Ontology Language [6] in concert with Dublin Core Terms [7] to accomplish two functions:

a) identify versions of concept schemes; and
b) identify one-to-one changes of concepts between schemes.

The second function of OWL Versioning does not account for a change in the concept, except where one concept (e.g. bananas) wholly replaces another concept (e.g. plantains). [1] This one-for-one act of substitution does not always happen. Editors often refine or lump together concepts in concept schemes. Currently, OWL Versioning in SKOS does not account for this refinement, lumping and other transformations of concepts (and their relationships) between different versions of concept schemes [8]. If more than a simple one-to-one relationship can be expressed, then thesauri could continue to evolve according to the literature of the DCMI conferences, while retaining the power of pulling together kinds of documents and *similar* documents, and still excluding *dissimilar* documents from search and retrieval. If SKOS incorporated mechanisms for making the ontogeny of vocabularies explicit, like the evolution of terms in the metadata thesaurus for the Dublin Core Online Conference Proceedings, then it would exploit the structured nature of revisions in order to facilitate retrieval. The next section outlines what structures will make explicit kinds, similar, and dissimilar concepts in concept schemes using the metadata thesauri [2, 3, 4] as examples.

3 Metadata Thesauri 2002-04

The DC2002 Terms list [2], generated by Bradley Allen, is a flat list of terms. It served as a pilot project for the Siderean interface [9, 10] to the DC2003 Conference Proceedings. We added hierarchical structure to this list with concepts from the literature of the 2003 Conference to develop the DC2003 Metadata Thesaurus [3]. Consequently, the relationship structure of the DC2002 Terms list changed dramatically when it migrated to the DC2003 Metadata Thesaurus. See the following example:

DC2002 Terms
a) applications
b) web services

DC2003 Metadata Thesaurus
a) Applications
NT Web Services

In 2002, the relationship between "Applications" and "Web Services" is *associative*—they were related by virtue of being at the same level of specificity within the domain of metadata research. However, in 2003 the relationship between the two concepts became *hierarchical* with "Web Services" represented as narrower in meaning than "Applications."

Another change from 2002 to 2003 is the lumping together of terms. For example: "metadata harvesting" and "Open Archives Initiative."

DC2002 Terms:
a) metadata harvesting
b) Open Archives Initiative

DC2003 Metadata Thesaurus
(a) Open Archives Initiative Protocol for Metadata Harvesting

Here we can see how two terms are lumped together to form one concept—focusing the meaning from a general account of harvesting and a general discussion of Open Archives Initiative to the specific Protocol for Metadata Harvesting sponsored by the Open Archives Initiative.

Finally, there are examples of refining concepts in the transition from 2003 to 2004 thesaurus.

DC2003 Metadata Thesaurus:
a) Cultural Heritage
[no other concepts]

DC2004 Metadata Thesaurus
a) Cultural Heritage
NT Sekisui-zu

From this example, it is clear that an indexer can be more specific about Cultural Heritage in the 2004 version.

As seen in the examples, when a concept scheme, in this case a term list turned thesaurus, changes over time, editors refine, lump and reconfigure concepts according to new relationships. The extension to SKOS Core Guide suggested here accounts for these phenomena. The extension not only account for vocabulary ontogeny, but also exploits that ontogeny for the purposes of retrieval.

4. Extending SKOS: Lumping, Refining, and Relationship Changes

In this next section, we outline how SKOS Core might handle the three types of problems encountered in revision of the metadata thesaurus discussed above. These suggestions are basic and are provided in order to start the conversation and not to finish it. Thesauri, as types of concept schemes, are complicated structures. We have not reviewed all the possible changes that could take place when revising them. To that end, we will limit ourselves to three types of changes: lumping, refining, and relationship changes. We will also discuss how identifying these changes in a vocabulary ontogeny will allow searchers to identify kinds, similar, and dissimilar documents.

4.1 Relationship Changes

In the example above where the concept scheme moved from a term list to a thesaurus, we saw how the relationship between two terms changed from being *associative* to *hierarchical*. The former is a relationship of loose definition [11, p. 60-61], where terms are associated conceptually. Aitchison et al. describe it as a relationship that is neither hierarchical nor equivalent—making it a bit of a catchall. The hierarchical relationship is one that shows superordination and subordination [11, p. 54] of concepts – where one is broader and the other narrower.

To illustrate a change in relationship structure in SKOS, we suggest that an explicit statement about the old relationship and a new relationship be made. It might be done like this (using a modified N3\Turtle [12]):

DC2003
skos:Concept “Web Services”
skos:wasRelated “Applications”
skos:narrower “Applications”

Since the relationship is a resource, it can be referenced in RDF/XML. This basic structure also allows for more detailed and descriptive statements about the kind of relationship. For example, there are a number of types of associative relationships [11] and an editor might express these as refinements where necessary.

4.2 Lumping

Where two concepts are lumped together into a single concept, we suggest SKOS make an explicit statement that what were once two concepts are now one. For example:

DC2002
skos:Concept “metadata harvesting”
skos:Concept “Open Archives Initiative”

DC2003
skos:ConceptLump “Open Archives Initiative Protocol for Metadata Harvesting”
skos:ConceptLumpTrace “metadata harvesting”
skos:ConceptLumpTrace “Open Archives Initiative”

Here we have a trace in the new version of the change. This conforms with the current suggestions of OWL versioning outlined in SKOS Core Guide [1]. It is assumed for this paper that it is not desirable to express lumping in DC2002.

4.3 Refining

Where an editor refines one concept by adding another subordinate concept, we suggest SKOS make an explicit statement stating that where once there was one concept there is now more than one.

skos: Concept “Cultural Heritage”
skos:ConceptRefinement “Sekisui-zu”

From these examples, and from the suggestions here about SKOS extensions, it is possible to see how making these changes between versions of concept schemes explicit an editor can aid retrieval. The searcher or a machine can follow the changes in concepts from version to version. Furthermore, these changes can be exploited by crawling through these changes and making sense of them. These changes can be used to describe similar and dissimilar documents for retrieval.

5. Summary

This paper has suggested three extensions to the SKOS Core Guide [1] all under the name vocabulary

ontogeny. We have proposed making explicit some of the changes between versions of concepts schemes by stating where concepts have been refined, lumped together, or their relationship structure has changed. We posit that making this explicit through SKOS Core will enhance information retrieval by making explicit these changes in the display of the retrieved set.

By extending SKOS in this way, we can put into place mechanisms that will exploit not inhibit the evolution of knowledge organization systems and their purpose—retrieval—on the Web..

Acknowledgements

This conceptual work of this paper relies primarily on the applied and empirical work of Bradley P. Allen and Siderean Software. I am grateful to Stuart Sutton for commenting on an earlier draft of this paper.

References

1. A. Miles and D. Brickley. SKOS Core Guide. Available: http://www.w3.org/2004/02/skos/core/guide/.
2. B. P. Allen. DC2002 Terms Available: http://purl.oclc.org/METADATARESEARCH/dc2002_termsrdf
3. B. P. Allen and J. T. Tennis. DC2003 Metadata Thesaurus Available: http://purl.oclc.org/METADATARESEARCH/dc2003_thesaurus.rdf
4. DC2004 Metadata Thesaurus Available: http://purl.oclc.org/METADATARESEARCH/dcconf_thesaurus.rdf
5. A. Miles and D. Brickley. SKOS-Core 1.0 Guide Available: http://www.w3.org/2001/sw/Europe/reports/thes/1.0/guide/
6. D. McGuinness and F. van Harmelen. OWL Web Ontology Language Overview Available: http://www.w3.org/TR/owl-features/
7. DCMI Usage Board. DCMI Metadata Terms Available: http://dublincore.org/documents/dcmi-terms/
8. S. R. Ranganathan. *Prolegomena to library classification* 3rd edition. Asia Publishing House, Bombay. 1967, pp. 351-359.
9. DC 2003Online Conference Proceedings Available: http://www.siderean.com/dc2003/search.jsp
10. Siderean Software website: http://www.siderean.com
11. J. Aitchison, A. Gilchrist, D. Bawden. *Thesaurus construction and use: a practical manual*. 4th edition. Fitzroy Dearborn Publishers, Chicago. 2000.
12. D. Beckett. *Turtle – Terse RDF Triple Language*. Available: http://www.ilrt.bris.ac.uk/discovery/2004/01/turtle/