Versioning Concept Schemes for Persistent Retrieval

by Joseph T. Tennis

Things change. Words change, meaning changes and use changes both words and meaning. In information access systems this means concept schemes such as thesauri or classification schemes change. They always have. Concept schemes that have survived have evolved over time, moving from one version, often called an *edition*, to the next. If we want to manage how words and meanings – and as a consequence use – change in an effective manner, and if we want to be able to search across versions of concept schemes, we have to track these changes. This paper explores how we might expand SKOS, a World Wide Web Consortium (W3C) draft recommendation in order to do that kind of tracking.

The *Simple Knowledge Organization System (SKOS) Core Guide* is sponsored by the Semantic Web Best Practices and Deployment Working Group. The second draft, edited by Alistair Miles and Dan Brickley, was issued in November 2005. SKOS is a "model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, other types of controlled vocabulary and also concept schemes embedded in glossaries and terminologies" in RDF. How SKOS handles version in concept schemes is an open issue. The current draft guide suggests using OWL and DCTERMS as mechanisms for concept scheme revision.

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, thereby enabling persistent retrieval.

Joseph T. Tennis is assistant professor in the School of Library, Archival and Information Studies, University of British Columbia. He can be reached at jtennis<at>interchange.ubc.ca

Changes in Concept Schemes

In order to illustrate vocabulary ontogeny, we use the *Metadata Thesaurus*, a type of concept scheme (Allen, 2002; Allen and Tennis, 2003, 2004, 2005) 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 (2002-2005). However, none of the documents indexed with the older versions are re-indexed with the revised version of the vocabulary. Each year is indexed using a different version of the vocabulary.

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. This would also, as alluded to above, make explicit how words change, how meaning changes and how use changes both words and meaning.

SKOS

SKOS officially stands for Simple Knowledge Organization System. Perhaps a better way to think of it, however, is as the Schema for Knowledge Organization Systems, since its purpose as described in the guide is to provide "a simple yet powerful framework for expressing knowledge organization systems in a machine-understandable way." The guide suggests how to track revisions and versions, but 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.

Notes in SKOS. The *SKOS Guide*, citing the draft *British Standard for Structured Vocabularies for Information Retrieval (BSI, 2004)*, offers two types of thesaurus notes useful for recording changes in versions of concept schemes:

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. The guide provides two examples:

1. A change in the placement of the concept: "*Pears* was previously listed as a narrower concept under *vegetables* instead of *fruits*"
2. A change in the labeling of concepts: "Introduced 1999; prior to that use *laptop computers* for the concept *notebook computers*"

The changeNote serves both the editor and indexers using the thesaurus. It is a private note, not intended for users, which documents "fine-grained changes to the concept for the purposes of administration and management." There are two examples given: one is a change where the concept moved from one part of the hierarchical structure to another. The second is a change in labeling, where the concept was relabeled from *Laptop computers* to *Notebook computers*.

OWL Versioning in SKOS. In order to signal a change of a concepts scheme from one version to another, SKOS suggests using OWL, Web Ontology Language (see *OWL Overview* edited by McGuinness and van Harmelen) in concert with Dublin Core Terms (DCMI Usage Board, 2005) to accomplish two functions:

1. Identify versions of concept schemes
2. Identify one-to-one changes of concepts between schemes

The second function of OWL Versioning as included in *OWL Overview* does not account for a change in the concept, except where one concept (for example, "bananas") wholly replaces another concept (for example, "plantains"). This one-for-one act of substitution does not always happen. Editors often move, refine or lump together concepts in concept schemes. Currently, OWL Versioning in SKOS does not account for this refinement, lumping or other transformation of concepts (and their relationships) between different versions of concept schemes. 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 pulling together *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 as examples.

Metadata Thesauri 2002-04

The *DC2002 Terms* list is a vocabulary with some broader and narrower term relationships. It served as a pilot project for the Siderean interface (Siderean, 2006) to the *DC2003 Conference Proceedings*. We added more hierarchical structure to this list with concepts from the literature of the 2003 Conference to develop the *DC2003 Metadata Thesaurus*. Consequently, the relationship structure of the *DC2002 Terms* list changed dramatically when it migrated to the *DC2003 Metadata Thesaurus*. The same types of revisions happened between 2003 and 2004. See the following examples:

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 tran-

sition 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 these examples, when a concept scheme changes over time, editors refine, lump and reconfigure concepts according to new relationships. Thesaurus construction methodologies and classification theory have all accounted for changes in concept schemes, yet they do not preserve structures in order to search and retrieve across versions.

The extension to *SKOS Core Guide* suggested here accounts for these phenomena. The extension not only accounts for vocabulary ontogeny, but also exploits that ontogeny for the purposes of retrieval.

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.

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 (Aitchison, Gilchrist and Bawden, 2000, 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 (Aitchison, Gilchrist and Bawden, 2000, 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 as below following the model provided by Turtle (the terse RDF Triple Language as defined by Beckett (2004)):

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 (Aitchison, Gilchrist and Bawden, 2000), and an editor might express these as refinements where necessary.

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 to the current suggestions of OWL versioning outlined in *SKOS Core Guide* (Miles and Brickley, 2005). It is assumed for this paper that it is not desirable to express lumping in DC2002.

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 persistent retrieval. The searcher or a machine can follow the changes in concepts from version to version. Furthermore, these changes can be exploited through query expansion methods and tools. They can be used to describe similar and dissimilar documents, regardless of the version of the concept scheme, for persistent retrieval.

Summary

This paper has suggested three extensions to the *SKOS Core Guide* (Miles and Brickley, 2005), 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 or 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. This will thereby enable persistent retrieval across different versions of a concept scheme.

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. By extending SKOS we can make the way in which words change and how meaning changes in concept schemes both human and machine readable.

Acknowledgements

A version of this paper was presented at the 2005 International Dublin Core Conference on Metadata Applications. The conceptual work of this paper relies primarily on the applied and empirical work of Bradley P. Allen and Siderean Software.

For Further Reading

Readings on indexing languages

Aitchison, J., Gilchrist, A., & Bawden, D. (2000). *Thesaurus construction and use: A practical manual.* 4th ed. Chicago: Fitzroy Dearborn.

Ranganthan, S. R. (1967). *Prolegomena to library classification.* 3rd ed. Bombay: Asia Publishing House.

Soergel, D. (1974). *Indexing languages and thesauri: Construction and maintenance.* Los Angeles: Melville Pub. Co.

DC Term Lists (Metadata Thesauri)

Allen, B. P. (2002). *DC2002 Terms.* Retrieved April 16, 2006, from http://purl.oclc.org/dcpapers/dc2002terms

Allen, B. P. & Tennis, J. T. (2003). *DC2003 metadata thesaurus.* Retrieved April 16, 2006, from http://purl.oclc.org/dcpapers/dc2003thesaurus

Allen, B. P. & Tennis, J. T. (2004). *DC2004 metadata thesaurus.* Retrieved April 16, 2006, from http://purl.oclc.org/dcpapers/dc2004thesaurus

Allen, B. P. & Tennis, J. T. (2005). *DC2005 metadata thesaurus.* Retrieved April 16, 2006, from http://purl.oclc.org/dcpapers/dc2005thesaurus

Other Reading

Beckett, D (2004). *Turtle – Terse RDF Triple Language.* Retrieved April 16, 2006, from www.ilrt.bris.ac.uk/discovery/2004/01/turtle/

British Standards Institution. (2004). *BS8723, Structured vocabularies for information retrieval.* Public draft. British Standards Institution.

DCMI Usage Board. (2005). *DCMI metadata terms.* Retrieved April 16, 2006, from http://dublincore.org/documents/dcmi-terms/

McGuinness, D. L. & van Harmelen, F (Eds). (2004, February 10). *OWL Web Ontology Language Overview.* World Wide Web (W3C) Web Ontology Working Group. Retrieved April 16, 2006, from http://www.w3.org/TR/owl-features/

Miles, A., & Brickley, D. (Eds.) (2005, November 5). *SKOS Core Guide: W3C working draft 2 November 11, 2005.* World Wide Web (W3C) Semantic Web Best Practices and Deployment Working Group. Retrieved April 13, 2006, from www.w3.org/TR/swbp-skos-core-guide/.

Siderean Software website. Retrieved April 16, 2006, from www.siderean.com