Proposed subsession for NKOS 2007 Workshop

Project ISO NP 25964: Structured vocabularies for information retrieval

ISO 2788 and ISO 5964, the international standards for monolingual and multilingual thesauri respectively dated 1986 and 1985, are very much in need of revision. A proposal to revise them was recently approved by the relevant subcommittee, ISO TC46/SC9. The work will be based on BS 8723, a five part standard of which Parts 1 and 2 were published in 2005, Parts 3 and 4 are scheduled for publication in 2007, and Part 5 is still in draft.

This subsession will address aspects of the whole revision project. It is conceived as a panel session starting with a brief overview from the project leader. Then there are three presentations of 15 minutes, plus 5 minutes each for specific questions. At the end we have 20 minutes for questions to any or all of the panel, and discussion of issues from the workshop participants.

Overview of ISO NP 25964 *10 mins*

Stella G Dextre Clarke, *Project leader, ISO NP 25964*

This presentation will briefly describe the five parts of BS 8723 which will form the basis for developing ISO 25964. It will bring out the changes of scope and other differences between the existing international standards and those now being planned, for example:

- updating the basics of monolingual and multilingual thesaurus construction
- adding advice about software for thesaurus management
- guidelines for other types of vocabulary, e.g. classification schemes and taxonomies
- guidelines on mapping between vocabularies
- formats and protocols for data exchange

One of the substantial innovations in BS 8723 is its emphasis on interoperability, which needs international adoption to become fully effective over the networks. Procedures for and progress towards adapting and adopting BS 8723 as an international standard will be described.

A data model and XML schema for BS 8723-5 *15-20 mins*

Nicolas Cochard, *Porism Ltd*

Data model

To ensure that the needs of BS8723-2 are understood and anticipated, the BS8723-5 working group started the development of the standard by creating a data model of a Thesaurus. This was achieved using the syntax of UML class diagrams.

Two models were defined: BS8723-5-Core and BS8723-5-Full.

BS8723-5-Core

The BS8723-5-Core model provides a presentation that looks simple initially and allows novices to get a foothold into using the standard, without having to overcome a huge comprehension barrier.
This model contains a representation of

- Thesaurus concepts
- Traditional hierarchical and associative relationships
- Scope notes
- Preferred and non-preferred

BS8723-5-Full

The BS8723-5-Full model provides a presentation of all the features required by a Thesaurus as specified by the BS8723-2 document.
This model contains a representation of

- Thesaurus concepts
- Traditional hierarchical and associative relationships and the ability to define additional types of relationships
- Scope notes, history notes, definitions, editorial notes
- Preferred, non-preferred and compound non-preferred terms
- Arrays and node labels
- And more…

XML Schema

From the each data model was derived an XML Schema. The BS8723-5-Full XML Schema was implemented as an "extension" of the BS8723-5-Core XML Schema. This ensures that an XML file in the BS8723-5-Core format is valid also in the BS8723-5-Full format.
Short extracts of Thesauri were implemented in XML to test that the basic features of a Thesaurus were supported by these XML Schemas. The tests implemented insure that alphabetical, hierarchical and classified display could be generated from these XML files via an automatic process.

Presentation

My presentation will introduce the NKOS audience to

- the choices made in the two UML class diagrams
- the difficulties encountered during the creation of the diagrams
- the various points of discussion which are still subject of controversy
- and the various choices made when deriving the XML Schema from the UML class diagram

From a thesaurus standard to a general knowledge organization standard?!

15-20 mins

Daniel Kless, *Airbus*

It is more than 30 years ago that the guidelines ISO 2788 and 5964 were developed for the design of thesauri – basically to meet the demands of bibliographic databases and libraries. The last revision of the standards dates around 20 years back. Information technology has changed the usage of thesauri – a development that has motivated many changes in BS 8723, the first successor of ISO 2788 and 5964.

Not only has the world of thesauri and libraries matured. Ever since there have been defined a variety of structured vocabularies, thesauri being just one of them: ontologies, taxonomies, classification schemes, topic maps – just to name some of them. For these types of vocabularies there hardly exist rules for the construction of the vocabulary content comparable to those for thesauri. There are standards for the formal description of some vocabulary types at most, e.g. SKOS, Topic Maps, RDF.

The guidelines in Parts 1 and 2 of BS 8723 and its predecessors are, unfortunately, not simply applicable on vocabularies other than thesauri. The degree to which this is reasonable has not been analysed. Thus, applications using structured vocabularies other than thesauri lack guidance for the construction of the vocabulary content.

The developments in BS 8723 – particularly those in the forthcoming Part 3, "Vocabularies other than thesauri" – try to catch up that knowledge gap. However, these vocabulary types are treated in much less detail than thesauri. Part 3 seems rather a detour from thesauri than a standard for other vocabularies. BS 8723 will basically remain a thesaurus standard, particularly in terms of its rules for construction.

The further development of BS 8723 as an ISO standard (ISO 25964) is a chance to continue the transformation of a once thesaurus-only standard to a truly general knowledge organization standard. The most important reasons that encourage such strategy are:

- While some of the rules from the thesaurus standard will have to be modified, a significant number of rules can be expected to apply directly to other types of structured vocabularies. So it makes sense to keep them in a single standard.
- It is easier to develop rules for different structured vocabularies if being put in contrast to thesauri. The thesaurus standards are based on decades of extensive experience and include also knowledge about "what is relevant to cover".
- It is highly useful to give general guidance in choosing the right type of structured vocabulary before the structured vocabularies are detailed.
- Bringing together various disciplines avoids the reinvention of knowledge and strengthens "knowledge structuring" as a professional discipline.
- A higher differentiation of knowledge structures / vocabularies can be expected resulting in more efficient and purpose-oriented development of vocabularies.

The resulting and certainly greatest benefit from a true “multi-vocabulary” standard is the interest of industry and many other disciplines than library science. Thus, the relevance of the standard will be increased. Examples for potential application areas of a common standard are:

- Skill catalogs in competence management
- Visualizations of the organizational structure
- The directory structures in computer file systems (particularly *shared* folders)
- Categorizations (typology) of files in a Document Management System
- Description of knowledge assets in knowledge management tools
- Categorization systems (typology) of music or picture archives
- Knowledge maps
- Corporate Encyclopaedias
- Vocabularies for search expansion in search engines
- The navigation structure, labelling system and the metadata on web sites
- The Table of content and / or the index of a book or complex document

Crosswalks and the USA Perspective *15-20 mins*

Emily Fayen (MuseGlobal), *Project Lead for ANSI/NISO Z39.19*; Marjorie Hlava *(Access Innovations, Inc.)*

In 2003, the USA ANSI / NISO began a major revision of Z39.19, the 1993 Standard for Monolingual Thesauri. In late 2005, the revised standard was published as Z39.19 Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies

In order to make the 1993 version relevant in today’s internet and active search and retrieval environment many changes were needed. . Whereas the original considered only thesauri, this revision is expanded to many types of controlled vocabularies. The enhanced range of the standard includes detailed explanations of important concepts and principles. The revised standard takes into account changes in information technology, the various ways that users search or browse, and the many types of content they find. The scope of the revised standard is significantly increased to address the broader needs of information producing organizations as well as new and different types of content and distribution. There is also attention paid to the display needs in the new internet environment.

One continuing challenge in the scope is insuring compatibility and crosswalks, not only between controlled vocabulary standards such as the current guidelines ISO 2788 and 5964, the new BSI 8723 but between other groups developing guidelines and standards in this area such as the W3C with SKOS and OWL and governments world wide developing and mandating taxonomies for their agencies and information offerings. If we can maintain such compatibility then controlled vocabularies based on one standard may be implemented on systems based on another. This would also increase reuse and mapping interoperability between controlled vocabularies. A clearing house to keep track of all the initiatives and suggested standards, a means to allow input from and to those initiatives, and publishing of best practices or lessons learned from implementations would be a worthwhile activity.

This paper will attempt to outline the Z39.19 process, the establishment and effect of current crosswalk initiatives and results - such as SKOS and the government initiatives currently underway.

Discussion between panel and workshop participants *20 mins*