Leonard Will - Willpower Information, Enfield, UK

The ISO 25964 Data Model for the Structure of an Information Retrieval Thesaurus

Abstract
International standard ISO 25964-1:2011 – Thesauri for information retrieval, includes a detailed data model for thesaurus structure. It is intended to provide a rigorous presentation of the elements and relationships which will not only clarify and standardise the varying and conflicting interpretations which exist but which can also be implemented consistently in automated systems. It makes a clear distinction between concepts and the terms which are used to label them, and includes other features that may be present in a thesaurus, such as compound equivalence, arrays and node labels, concept groups, notes and version history.

Introduction
The recently-published international standard, ISO 25964-1:2011 – Thesauri for information retrieval, presents a data model for thesaurus structure which is more extensive than any published previously. The model is shown in Figure 1, using UML conventions (Object Management Group, 2007). Previous thesaurus standards are summarised by Dextre Clarke and Zeng (2011), so they will not be discussed here other than to note that they did not provide a model for the structure of thesaurus data. The new model is intended to provide a rigorous presentation of the entities and relationships which will not only clarify and standardise the varying and conflicting interpretations which exist but which can also be implemented consistently in automated systems. The SKOS (Simple Knowledge Organization System) format (W3C, 2009) is designed to present KOS data in a format that is suitable for machine inferencing and particularly for use in the Semantic Web. This is largely compatible with the ISO model, but does not yet implement all its features. Discussions are continuing on possible extensions to SKOS to cover arrays, concept groups, compound equivalence and version history.

Structure Based on Concepts, not Terms
The model is based on the understanding that thesauri show the relationships between concepts – units of thought – and distinguishes these from the terms that are used to label these concepts. These terms may be in one or more languages and one term per language is chosen as a preferred term for each concept. One or more additional terms for the same concept may be recorded in the thesaurus, as non-preferred terms. This linkage of multiple terms to the same concept is another way of expressing the traditional equivalence relationship between terms normally indicated by the tags USE/USE FOR, although the model does also show that relationship, for compatibility with existing systems. It additionally provides a “role” attribute which allows the nature of the relationship to be specified if desired, for example that the relationship between a preferred and non-preferred term may be abbreviation/full form, formal/informal, obsolete/current or scientific/popular. It was thought unnecessarily complicated to provide for such relationships between one non-preferred term and another.

Multilingual Thesauri
It is sometimes argued that some concepts which occur in one language do not occur in another, so that both language versions cannot use the same structure. This argument is invalid, because although a language may not have a specific term to represent a concept, the concept can still exist. In many cases a term is sufficient to identify a concept unambiguously, but the general principle is that concepts are not defined by the terms used to label them but rather by their place in a hierarchy and by scope notes which specify the criteria for distinguishing them from their parent concepts. For this reason the model specifies that each concept should have an “identifier” which is independent of the terms
which may be used to label it. If a language does not have a convenient label for a concept, it may be necessary to use a compound term, a "loan term" from another language, or the term which most nearly matches the concept, with a note to specify how its meaning is restricted or extended for use within the thesaurus.

**Fig. 1: The ISO 25964 data model**

Boxes in this diagram indicate “classes” in UML terminology. The diamond symbol represents an “aggregation”, i.e. a “has-a” relationship. The triangle symbol indicates “generalization”, i.e. an “is-a” relationship, linking a general class with more specific types of that class, which inherit all the attributes of the parent class. Normal associations (without a diamond) indicate a relationship where each instance of one class is linked to an instance of another class.

Symbols such as 1, 0..* and 1..* indicate “one”, “zero or more”, or “one or more occurrences” respectively.
Compound Equivalence

A more complex case than the simple USE/USE FOR or “preferred/non-preferred” term relationship is that of compound equivalence, where a compound concept, such as coal mining, does not exist in the thesaurus but has to be expressed as a combination of two or more simpler concepts which are there. This is shown symbolically as:

coal mining
USE+ coal
USE+ mining

with reciprocals such as “coal UF+ coal mining”. Because the complex concept is not in the thesaurus, there is no provision for recording its attributes or giving it a scope note – it has to be interpreted from the scopes of the component concepts. As a thesaurus is normally
used for post-coordinate indexing, the indexer would assign the two terms coal and mining to a document without expressing any relationship between them. A searcher would be expected to construct a search statement combining these terms with a Boolean AND operator. In the terminology of set theory, coal mining applies to the “intersection” of the set of documents that deal with coal and the set of documents that deal with mining. On the other hand a compound concept may apply to the “union” of two or more sets of documents rather than their intersection. Although ISO 25964 does not specifically deal with this case, it is generally better for the thesaurus builder to add such a compound to the thesaurus, showing its components as narrower concepts, rather than expressing it as a compound non-preferred term. For example, rather than

fossil fuels
USE+ coal
USE+ natural gas
USE+ petroleum

it is better to have

fossil fuels
NT coal
NT natural gas
NT petroleum

Hierarchical Relationships and Transitivity

Hierarchical relationships between concepts are included in the ISO25964 model, and the traditional symbols such as BT/NT are retained for consistency with current practice, although these are to be interpreted as meaning “broader concept/narrower concept” rather than “broader term/narrower term”. There is provision for each relationship to be specified by an optional “role”. This can be used to distinguish the three types of hierarchical relationship: generic, partitive and instantial, (symbolised as BTG/NTG, BTP/NTP and BTI/NTI) and even to subdivide these further if required, but in a way which allows the distinctions to be ignored by systems which do not use them.

The first level of distinction is important in automated systems and for compatibility with ontologies, where it is necessary to recognize whether a relationship is transitive or not, i.e. whether the relationship holds between concepts which are related hierarchically but where one is not the direct child of the other. A hierarchical chain in which all the relationships are generic/specific will maintain transitivity, but if it is mixed with whole/part relationships it will not.

For this reason, among others, the standard recommends that partitive relationships should normally be used only in a few specific cases: disciplines or fields of discourse, geographical locations, systems and organs of the body and hierarchical social structures. The first of these could be interpreted as generic in any case – is physics a “kind” or a “part” of science? Geographical locations is a special case because the concepts have proper names which label individual instances rather than classes, so that a generic relationship is not possible. This is different from the instantial relationship, which is used to show that an instance is a member of a class, so that we have

countries
NTI India

but

India
NTP Karnataka
Top concepts
Each concept can have a pointer linking it to the concept at the top of any hierarchy in which it occurs. These top concepts can be facet names, for example, and this can facilitate browsing by clearly indicating which facet a concept is in. It can also be used for validation, because hierarchical relationships are valid only if the two concepts are in the same facet. In addition, a concept can have a Boolean (true/false) attribute to indicate whether it is a “top concept”. This can be useful in producing a list of top level concepts from which to start browsing.

These links and attributes are strictly speaking redundant, because top concepts could be identified by navigating up the hierarchy until no more broader concepts can be found, but as this would use substantial processing resources it will generally be more efficient to store the information rather than determining it every time it is needed.

Associative Relationships
Similarly, associative relationships can optionally specify the nature of the relationship, such as cause/effect, process/product, or person/discipline, while allowing these all to be treated as the catch-all “related concept” (RT/RT) when necessary. This allows a thesaurus to come closer to the approach taken in ontologies, where the nature of every relationship is specified.

Arrays and Node Labels
Groups of sibling concepts, which have a common parent concept, may be organized into arrays, introduced by node labels. These are an important and helpful feature for navigation, browsing and selection of terms when hierarchical displays of thesauri are presented to users, and many existing systems do not handle these well. The order in which concepts are displayed within an array may be different from the alphabetical order of preferred terms, perhaps following some inherent sequence such as number, size or age. This is indicated by an attribute of the array entity called “ordered”, which may be true or false. Node labels, which normally contain a characteristic of division (such as “by age” in the node label “people by age”) do not represent concepts and do not have hierarchical or associative relationships with concepts. They are not preferred or non-preferred terms, although the limitations of some thesaurus software force them to be treated as such.

Arrays may be nested, so that any array may contain subordinate arrays as well as individual concepts. For example this extract from the Art and Architecture Thesaurus (2012) shows arrays introduced by the node labels <nails by form: head type> and <nails by form: point type> as subordinate arrays within the array with the label <nails by form>.

nails
  – <nails by form>
    – – cut nails
    – – helical nails
    – – hook nails
    – – <nails by form: head type>
      – – – double-headed nails
      – – – flat-head nails
      – – – headless nails
    – – <nails by form: point type>
      – – – barbed nails
      – – – blunt nails
      – – – chisel nails
Concept Groups
Many thesauri group concepts into subsets, often discipline based, called “themes” (Eionet, 2011), “microthesauri” (Unesco, 2012), “domains” (Eurovoc, 2012) or “groups”. The box in the model called “concept group” provides for such groups. The concepts within such a group may or may not have any hierarchical or associative relationship with each other, and may be drawn from distinct hierarchies or facets of the thesaurus, such as activities, people, places or things. Concept groups may be nested, and may have a scheme of notation distinct from that used for concepts or arrays, thus providing the possibility of a classified arrangement which complements the generic hierarchy of the thesaurus itself, as in a “Thesaurofacet” (Aitchison, 1970) or “Classaurus” (Devadason, 1985).

Notes and Attributes
The model provides for notes of various types to be associated with concepts and terms, as well as allowing the addition of custom notes to cater for the particular needs of special applications. Any note may contain references to other concepts, and the model allows these to be stored as links rather than text, to avoid the common problem of references not being updated when changes are made. In addition, many of the boxes in the model include several attributes, and where possible these have been drawn from other standard schemes; many of the attributes of the thesaurus as a whole, for example, are those of the Dublin Core Metadata Initiative (2012).

Version history
There is provision for attaching a version history to a thesaurus, recording the various versions which have been made available, and for each of these showing what distinguishes that version from others and whether it is still current. The attribute “this version” in each of the entries in the list of versions is set to “true” for the version in hand. Dates of creation and modification can also be attached to each concept and each term.

Part 2, Mapping, Coming Soon
The model given in ISO 25964 is for a single thesaurus. It may be multilingual, but the structure of concepts does not differ between languages. Mapping, or the creation of relationships between two or more thesauri or other types of knowledge organization schemes, will be discussed in Part 2 of the standard, currently in draft. To extend the model to cover such mapping would require models for each scheme to be shown side-by-side, with relationships between the concepts of one and the concepts of the other.

Acknowledgements
I am grateful to Stella Dextre Clarke, leader of the ISO 25964 project, for helpful comments on an early draft of this article.

Permission to reproduce extracts from ISO 25964 is granted by BSI. ISO standards can be obtained from the ISO store at http://www.iso.org/iso/store.htm and British Standards can be obtained in PDF or hard copy formats from the BSI online shop: www.bsigroup.com/Shop or by contacting BSI Customer Services for hardcopies only: Tel: +44 (0)20 8996 9001, Email: cservices@bsigroup.com. ISO standards are also normally available from national standards bodies.

The data model in diagrammatic form is publicly available on the web site for the ISO25964 project, at <http://www.niso.org/schemas/iso25964/>. An XML schema intended for use when exchanging thesauri in whole or in part has been derived from the data model and is on the same site together with related documentation and a test document illustrating how a typical thesaurus conforming to the ISO 25964 data model can be serialized in an XML format.
References


