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

EDITOR’S SUMMARY
The international standard for thesaurus structure, ISO 259641:2011 – Thesauri for Information Retrieval, published August 2011, presents an extensive data model more advanced than previous versions. The standard facilitates presenting a knowledge organization system in machine compatible SKOS (simple knowledge organization system) format. The model lays out the concepts and terminology applicable to single thesauri, establishing the important difference between concepts and terms. It provides strategies to go beyond simple hierarchical relationships and related concepts, clarifying the nature of relationships between concepts. The new standard addresses several details in thesaurus construction and use for information retrieval, including compound equivalence and node labels. The updated standard supports concept groups or microthesauri and provides for term notes and file versions. Mapping concepts between thesauri will be covered in Part 2 of the standard.

KEYWORDS
thesauri
standards
index language construction
data models
SKOS
information retrieval
standards developing organizations

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. It is intended to provide a rigorous presentation of the entities and relationships that will not only clarify and standardize the varying and conflicting interpretations that exist, but which can also be implemented consistently in automated systems. The SKOS (simple knowledge organization system) format is designed to present KOS data in a format that is suitable for machine inferencing and particularly for use in the Semantic Web. This standard 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 these other features.

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 that 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.

**Compound Equivalence**

A more complex case 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 case 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 attaching a scope note to it – 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 modelled, and the traditional symbols such as BT/NT are retained for consistency with current practice, although these designations 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 role can be used to distinguish the three types of hierarchical relationship – generic (kind of), partitive (part of) andinstantial (instance of) – and even to subdivide these types further if required, but in a way that allows the distinctions to be ignored by systems that 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, that is, 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 are a special case because the concepts have proper names that 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.

**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 link 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. A concept can also have a Boolean (true/false) attribute to indicate whether it is a “top concept.” This feature 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 could be found, but as this search 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 all relationships 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 labels are an important and helpful feature for navigation, browsing and selection of terms when hierarchical displays of thesauri are presented in a human interface, and many existing systems do not handle these array displays 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. 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.

**Concept Groups**

Many thesauri group concepts into subsets, often discipline based, called “themes,” “microthesauri,”
“domains” 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” or “Classaurus.”

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 to the particular needs of special applications. 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.

Version History
There is provision for attaching a version history to a thesaurus, recording the various versions that have been made available and, for each, showing what distinguishes that version from others and whether it is still current. Dates of creation and modification can also be attached to each concept and each term.

Coming Soon – Part 2: Mapping
The model given in ISO 25964 is for a single thesaurus. It may be multilingual, but the structure of concepts does not differ among 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.

The data model in diagrammatic form is publicly available on the website for the ISO25964 project, at 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.

Obtaining the Standard
The full ISO 25964 standard may be purchased directly from ISO in Switzerland (www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=53657) in pdf or paper format or from national standards organizations such as ANSI (http://webstore.ansi.org/RecordDetail.aspx?sku=ISO+25964-1:2011) in downloadable pdf format only.

Acknowledgement
I am grateful to Stella Dextre Clarke, leader of the ISO 25964 project, for helpful comments on a draft of this article. ■