ARTICLE EXCERPTED FROM:

ISQ INFORMATION STANDARDS QUARTERLY

WINTER 2012 | VOL 24 | ISSUE 1 | ISSN 1041-0031

SPECIAL EDITION: YEAR IN REVIEW AND STATE OF THE STANDARDS

NISO
How the information world CONNECTS

SP [SPOTLIGHT]

Stella G. Dextre Clarke

Marcia Lei Zeng

STELLA G. DEXTRE CLARKE AND MARCIA LEI ZENG

From ISO 2788 to ISO 25964: The Evolution of Thesaurus Standards towards Interoperability and Data Modeling

STANDARD SPOTLIGHT

The information retrieval thesaurus emerged from pioneering work in the 1960s, and by 1974 the principles and practical guidance for constructing thesauri were enshrined in the international standard ISO 2788 as well as national standards such as ANSI/NISO Z39.19. Successive updates since then have led most recently to the publication of ISO 25964-1, Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval. So what has changed over the years?

In answer to that question, the principles have hardly changed at all. But round about us the world has changed. Technology has changed, and with it the opportunity for extending information retrieval over the whole world's internetworked resources. The new opportunities have led us to re-examine the principles, and discover that in the 1970s we did not articulate them in the clear logical way that is needed for today's computer applications. In particular, we did not then clarify the difference between the *concepts* of a search for information and the *terms* in which we express the query. If this distinction is fudged, human users may not be put out at all, but computers are at risk of floundering. To perform on the Semantic Web, computer software needs an explicit data model that distinguishes between terms and concepts.

In this article we trace the development of the thesaurus standards over the years, looking in particular at how the concept/term distinction is handled and more generally at the changes needed to facilitate interoperability and ease of handling thesaurus data by computers.

Raison d'être of the thesaurus

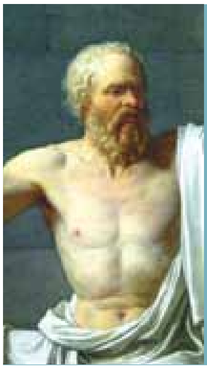
What is a thesaurus all about? The thesaurus is a tool to support subject access to information. Many other tools and approaches have been tried, from classification at one end of the spectrum to full text search at the other, and the thesaurus approach sits somewhere in between.

The classification approach relies on prior development of a scheme of the knowledge in a particular domain (usually reflecting one of the ways a domain carries and passes knowledge from generation to generation) in which each subject or combination of subjects is assigned a unique code. The theory is that if each document in a collection is given the right code according to the rules of the scheme, then anyone searching for a particular subject will find all the relevant documents, just by using the code.

Since conversion of subjects to codes requires some skill, it adds to retrieval costs and is not popular with users who like to express their search needs in ordinary words. This is the argument for full text search, in which users can simply look for occurrences of their search words anywhere in a document collection. The pitfalls of this approach are well known, in particular that a subject may be expressed using many different words and word combinations. An exhaustive search for just one topic typically needs multiple formulations of the query, and even then can fail if the searcher has no insight into the language of the original relevant documents.

This is the rationale for the thesaurus approach: if you can guide people always to use the same terms for the same

BOX 1: The following syllogism will be familiar to students of Aristotelian logic



'man' is a 3-lettered word.

Socrates is a man.

Therefore, Socrates is a 3-lettered word.

The logical flaw is very obvious to a human reader, but a computer can easily be fooled if statements about a term are presented looking like statements about the concept represented by the term.

concepts, and if any particular term can apply to only one concept, then users can search reliably with words, not codes. That's the theory, at any rate. And everything in the thesaurus standards is designed to make the thesaurus work reliably as a guide for choosing the right term for the concept sought. The introduction to the first (1974) edition of the international standard ISO 2788, *Guidelines for the establishment and development of monolingual thesauri*, states this objective: "there is a need for practical methods of representing concepts simply and clearly and of ordering them by clarifying their interrelationships."

Concepts versus terms: the dilemma and the confusion

So if the thesaurus is a guide to help a user choose the right term for a given concept, what are the basic units of its content? Does the thesaurus hold terms or does it hold concepts? This seems a crazy question, for terms and concepts are inextricably linked. All the while a concept is inside our heads, it can be independent of words or language. But as soon as we try to communicate it to another person or to a search system, we have to represent it in some way—usually by words or codes or pictures. The only way a thesaurus can list concepts in alphabetical order is by representing them as terms. Inevitably, the thesaurus contains terms *as well as* the concepts behind the terms. And sometimes, it is hard to tell which is which, as illustrated in Box 1.

Thus although ISO 2788 had a clear objective of organizing *concepts* and their interrelationships, the 1974 edition goes on to recommend: "the hierarchical relation is represented by the references BROADER TERM (BT), representing the relation of a concept being superordinated, and NARROWER TERM (NT), indicating the reciprocal relation." The tags BT, NT, and RT (RELATED TERM) were not invented by ISO 2788 (nor by the contemporaneous American national standard ANSI Z39.19-1974). No, these tags had been used in thesauri throughout the 1960s, especially in the influential *Thesaurus of Engineering and Scientific Terms (TEST)*. However, by perpetuating a convention that signposted relationships between *concepts* with abbreviations suggesting *terms*, the standard allowed confusion to creep in. The most recent (1986) edition of the same standard acknowledges this confusion and explicitly warns the reader "For practical purposes, 'term' and 'concept' are sometimes used interchangeably." This note was an admission that the BT/NT/RT convention was too heavily embedded in practice to change, and so the tags have been retained in standards and continue in widespread use to the present day.

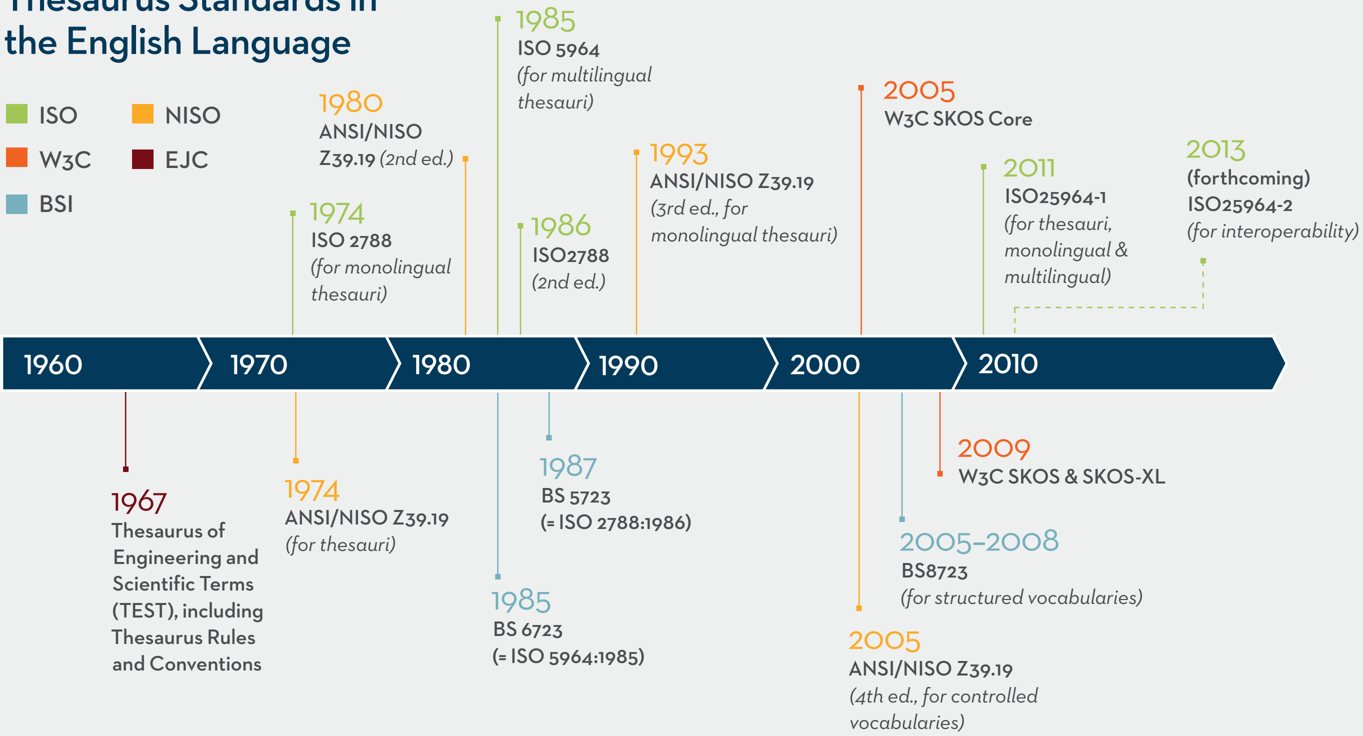
The pressure for clarification and a broader scope

The confusion regarding concepts vs. terms in ISO 2788 could have been dispelled by including a data model. (This same confusion existed in the sister standards ISO 5964, BS 5723, BS 6723, and ANSI/NISO Z39.19. See Box 2 and Figure 1 for brief details of these superseded standards, and page 23 for a description of Z39.19.) But the need for such a model was not fully recognized until the end of the twentieth century. Until then, thesauri had been used mostly in contexts where humans controlled or mediated the search process. Intuitively a human user grasps the difference between a term and a concept, and can interpret search results without confusion. A data model becomes necessary only when a machine needs instruction in how to handle and interpret the data.

CONTINUED »

FIGURE 1.

Timeline of Landmark Thesaurus Standards in the English Language



BOX 2

Landmark Thesaurus Standards, now superseded

TEST AND OTHER PRECURSORS

Pioneering work in the 1960s led to publication of a number of influential thesauri as well as guidelines for thesaurus development, as described in Krooks & Lancaster and Aitchison & Dextre Clarke. Of these, the most influential was the *Thesaurus of Engineering and Scientific Terms (TEST)* in 1967, with its Appendix *Thesaurus Rules and Conventions*. Among the *TEST* conventions still prevalent today is the use of tags BT, NT, and RT to identify relationships between concepts.

ISO 5964 *Guidelines for the Establishment and Development of Multilingual Thesauri*

First published in 1985, it has now been withdrawn, superseded by ISO 25964-1. ISO 5964 was based on the same tacit model as ISO 2788, and suffered from the same lack of clarity in distinguishing between terms and concepts.

ISO 2788 *Guidelines for the Establishment and Development of Monolingual Thesauri*

First edition was published in 1974; the most recent edition (1986) was withdrawn in 2011 when superseded by ISO 25964-1. The intention of ISO 2788 was to deal with concepts, providing guidelines for representing them unambiguously by means of terms. However, there was no explicit data model and the difference between terms and concepts was not articulated clearly.

BS 5723 AND BS 6723

The most recent editions of these British Standards were identical to ISO 2788-1986 and ISO 5964-1985 respectively. They were withdrawn in 2005-2007 when superseded by the first four parts of BS 8723.

CONTINUED »

That need is much more evident in the twenty-first century. The success of the Semantic Web, for example, will depend on computers acting in coordination with each other so that intelligent agents can retrieve and manipulate information from multiple networked resources. If the difference between a term and a concept is not made clear, a computer can easily draw a false inference (see Box 1). The need for machine-to-machine communication and reasoning capability has provided much of the incentive for including a data model in the most recent thesaurus standards.

Semantic manipulation is not the only pressing need. The digital age has encouraged the emergence of many different vocabularies and vocabulary types, often working alongside traditional thesauri. It has also brought a demand for interoperability to underpin activities such as web services; the publishing, aggregation, and exchange of thesaurus data via multiple media and formats; and behind-the-scenes exploitation of controlled vocabularies in navigation, filtering, and expansion of searches across networked repositories. Many of the interoperability needs appear in the recommendations of a Workshop on Electronic Thesauri, organized by NISO on November 4-5, 1999. Following this influential workshop, not only was ANSI/NISO Z39.19 revised, but the new standards BS 8723, SKOS, and ISO 25964 have emerged. Figure 1 shows a chronology of the emergence of the key English-language standards for thesauri.

Towards interoperability: revision of national standards

As a direct outcome of NISO's 1999 Workshop, the 4th revision of the ANSI/NISO standard came out in 2005 Whereas previous editions had dealt only with thesauri, the scope of the revision was expanded to cover various types of controlled vocabularies that may share the same approaches or structures when dealing with common problems (including lists of controlled terms, synonym rings, taxonomies, and thesauri). The new Z39.19 has a section on interoperability, and a revised title: *Guidelines for the Construction, Format, and Management of Monolingual* ***Controlled Vocabularies*** (emphasis added by authors; previous title referred only to "Thesauri").

Like ISO 2788, this version of the standard is fundamentally concept-centered, but still describes the relationships as between "terms". No formal data model is given to clarify the distinction. See for example, "The relationships among terms in a controlled vocabulary are indicated by semantic linking. Semantic linking encompasses various techniques and conventions for indicating the relationships among terms." (ANSI/NISO Z39.19-2005, Section 8.1, Semantic Linking)

Addressing many of the same issues as Z39.19-2005, BS 8723, *Structured vocabularies for information retrieval – Guide*, has five parts, published between 2005 and 2008. As well as covering mono- and multilingual thesauri in depth, it deals more briefly with other vocabulary types (classification schemes, taxonomies, subject heading schemes, ontologies, and name authority lists). And in Part 4 it provides guidance on mapping between vocabularies. The call for a data model is explicitly met in Part 5 (also known as DD 8723-5), together with an XML schema for exchange of whole thesauri or subsets thereof.

The BS 8723 data model does much to dispel the concept/term confusion by establishing separate classes for "concept" and "term". The model clearly shows that hierarchical and associative relationships apply between concepts, whereas equivalence relationships apply between terms. However, the text in other parts of the standard is not always rigorous in articulating the distinction and, like all the forerunner standards, it could not break away from the BT/NT/RT tagging convention.

SKOS data models and the thesaurus standards

While the national and international standards described so far have all dealt fundamentally with the construction of thesauri, the standards of the World Wide Web Consortium (W3C) are concerned instead with Web functions, and in particular those of the Semantic Web. Thus the W3C Recommendation *SKOS (Simple Knowledge Organization Systems)* is designed to support publication of vocabularies such as thesauri on the Web. And at its heart is a data model that explicitly distinguishes between concepts and the labels used to represent concepts.

The SKOS Core data model was released in 2005 as a W3C Working Draft (*SKOS Core Vocabulary Specification*). It clearly emphasized a concept-centric view of vocabulary, where primitive objects are not labels; rather, they are concepts represented by labels. In SKOS the semantic relationships between concepts correspond very closely to the hierarchical and associative relationships recommended in thesaurus standards. They take the form of three standard "properties": skos:broader and skos:narrower for hierarchical links and skos:related for associative (non-hierarchical) links between concepts. The SKOS Core specification was superseded in 2009 by the official W3C Recommendation *SKOS Simple Knowledge Organization System Reference*. In this approved version, the basic SKOS Core data model is supplemented in its Appendix by an eXtension for Labels (SKOS-XL). In addition to all that is conveyed by SKOS Core for relationships between concepts, the extension provides additional support for identifying, describing, and linking lexical entities.

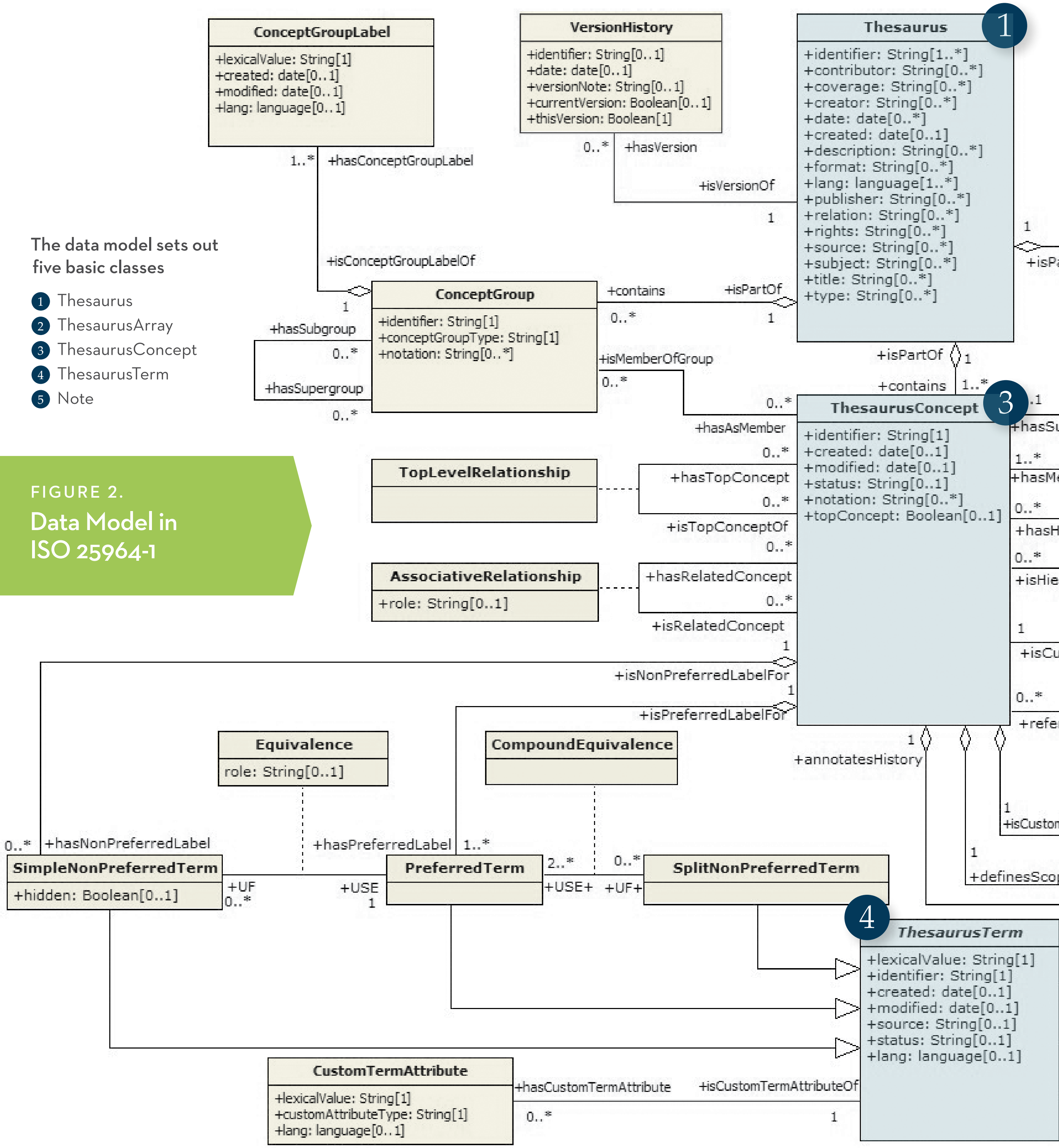
CONTINUED »

The data model sets out five basic classes

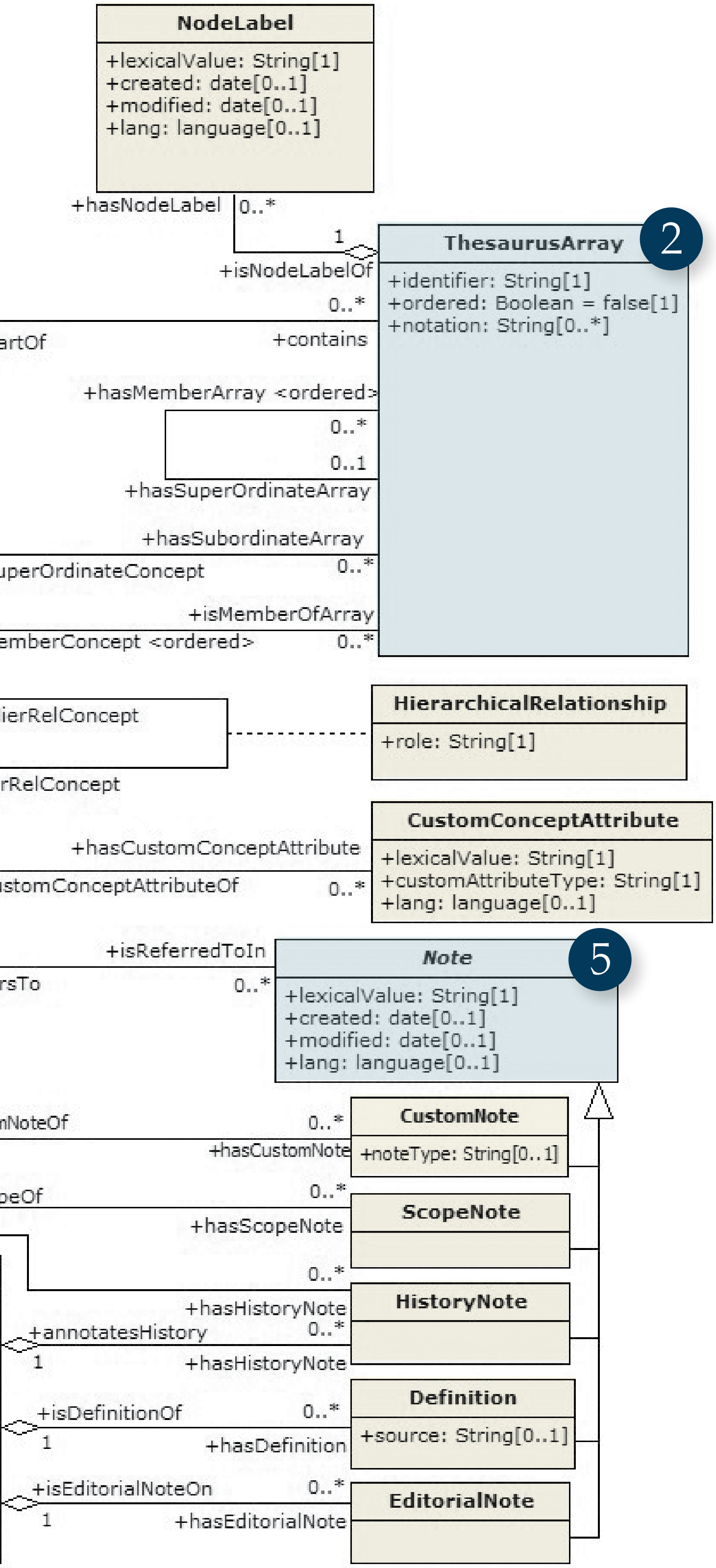
1. Thesaurus
2. ThesaurusArray
3. ThesaurusConcept
4. ThesaurusTerm
5. Note

FIGURE 2.

Data Model in ISO 25964-1



This data model graphic is reproduced from ISO 25964-1:2011 with permission of the American National Standards Institute (ANSI) on behalf of the International Organization for Standardization (ISO). ISO 25964-1 is copyrighted by ISO; no part of this material may be copied or reproduced in any form, electronic retrieval system or otherwise made available on the Internet, a public network, by satellite or otherwise without the prior written consent of ANSI. Copies of this ISO standard may be purchased from ANSI, 25 West 43rd Street, New York, NY 10036, (212) 642-4900, http://webstore.ansi.org.



ISO 25964: Thesauri and interoperability with other vocabularies

The new, two-part international standard has been developed by a working group with members from 15 countries, a chairman from the UK, and a Secretariat run by NISO in the US. The first part, known as ISO 25964-1, *Thesauri for information retrieval*, came out in August 2011. It updates, revises, and replaces ISO 2788 and ISO 5964, as well as some parts of BS 8723. This latest publication has been able to draw on all the previous work, for example the conclusions of NISO's 1999 Workshop and the data model and schema developed in BS 8723.

The scope of ISO 25964-1 includes:

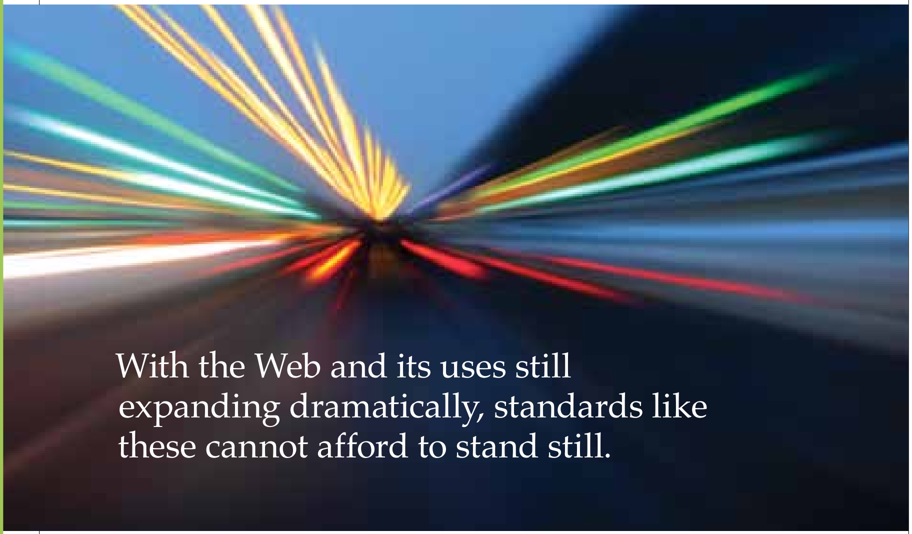
- Thesaurus content and construction, mono- or multi-lingual
- Guidance on applying facet analysis to thesauri
- Guidance on managing thesaurus development and maintenance
- Functional requirements for software to manage thesauri
- A data model and derived XML schema, available free of charge on a site hosted by NISO.

Additional aspects of interoperability (especially guidance on mapping concepts across thesauri and other vocabularies) will soon be covered in Part 2 of the standard.

ISO 25964 is much more rigorous than any of its precursors in distinguishing clearly between terms and concepts. It retains the tags BT, NT and RT (because these have been widely used in thousands of existing thesauri) but clarifies that the relationships they indicate are between concepts, not terms. The text explanation is explicitly confirmed in the data model, shown in Figure 2.

The data model sets out five basic classes, *Thesaurus*, *ThesaurusArray*, *ThesaurusConcept*, *ThesaurusTerm*, and *Note*. Attributes for each class and associations of classes reflect all of the features of thesauri that are recommended in the text. The model is accompanied by clear explanatory notes, for example in Section 15.2.3: "Each concept in the thesaurus is represented by one preferred term per language, and by any number of non-preferred terms. The notation, scope note and broader/narrower/related term relationships apply to the concept as a whole, rather than to its preferred term. A unique identifier can be assigned to each concept." Benefits of adopting the model include easier implementation by computers, consistency enforced in thesaurus construction and mapping, greater interoperability between thesauri and with other vocabularies, and enhanced performance at all stages from design of the thesaurus through development, management, and exchange.

CONTINUED »



With the Web and its uses still expanding dramatically, standards like these cannot afford to stand still.

Continuous and further work

With the Web and its uses still expanding dramatically, standards like these cannot afford to stand still. Part 2 of ISO 25964, covering interoperability between thesauri and other vocabularies, has reached the stage of Draft International Standard, and is the subject of public review and comment through mid-May 2012. The principles and practice of mapping are its prime focus. The scope includes interoperability with classification schemes, taxonomies, subject heading schemes, ontologies, terminologies, name authority lists, and synonym rings. After the feedback is accommodated, an approved standard is expected to emerge later in 2012.

There is more good news. During the past decade, in which the data models for SKOS and ISO 25964-1 were both under development, the teams responsible for them kept up good communication. Both teams drew liberally from the best of the concept-centered intentions of ISO 2788. As a result, the data models are largely compatible, particularly when the SKOS-XL extension is taken into account. At the time of writing, the ISO 25964-1 model has some optional features, not present in SKOS, to allow for capabilities (such as compound equivalence) that are not currently supported by SKOS. However, work is already under way to develop another SKOS extension to provide for these extra features. Alignment is the watchword, avoiding divergence. Already SKOS (supplemented when necessary by SKOS-XL) enables a great many thesauri compliant with ISO 25964 to be published on the World Wide Web, and others will follow. As the Semantic Web evolves, it will be fascinating to see what developments come next. | SP | doi: 10.3789/isqv24n1.2012.04

STELLA G. DEXTRE CLARKE <stella@lukehouse.org> is an independent consultant specializing in the design and implementation of thesauri. She leads the ISO 25964 project. **MARCIA LEI ZENG** <mzeng@kent.edu> is professor at Kent State University. She is the U.S. member of the ISO working group that developed ISO 25964-1, and continues work on Part 2 of the standard.

REFERENCES

ANSI/NISO Z39.19-2005, *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies.* Bethesda, MD: NISO Press. www.niso.org/standards/z39-19-2005/

Aitchison, J., and S.G. Dextre Clarke. The thesaurus: a historical viewpoint, with a look to the future. *Cataloging & Classification Quarterly*, 2004, 37(3/4): 5-21.

BS 8723 (2005-2008), *Structured vocabularies for information retrieval – Guide.* London: British Standards Institution. Parts 1-5.
Part 1: *Definitions, symbols and abbreviations* (2005). Part 2: *Thesauri* (2005). Part 3: *Vocabularies other than thesauri* (2007).
Part 4: *Interoperability between vocabularies* (2007). Part 5: *Exchange formats and protocols for interoperability* (2008).

BS 8723 Schema and Data Model (Part 5 of BS 8723 is also known as DD8723-5). http://schemas.bs8723.org/Model.aspx

ISO 25964-1:2011, *Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval.* Geneva: International Organization for Standards, August 8, 2011.

ISO 25964-1 Schema and Data Model. www.niso.org/schemas/iso25964/#schema

ISO 25964-2. Thesauri and interoperability with other vocabularies Part 2: Interoperability with other vocabularies. [In development; not yet published. Draft online until April 30, 2012 at http://drafts.bsigroup.com/]

Krooks, D.A., and F.W. Lancaster. The Evolution of Guidelines for Thesaurus Construction. *Libri*, 1993, 43(4): 326-342.

Milstead, Jessica. *Report on the Workshop on Electronic Thesauri November 4-5, 1999.* National Information Standards Organization, 1999. www.niso.org/news/events/niso/past/thesau99/thes99rprt.html

W3C Working Draft, *SKOS Core Vocabulary Specification.* W3C, May 10, 2005. www.w3.org/TR/2005/WD-swbp-skos-core-spec-20050510/

W3C Recommendation, *SKOS Simple Knowledge Organization System Reference.* W3C, August 18, 2009. www.w3.org/TR/skos-reference/

W3C Recommendation. *SKOS eXtension for Labels (SKOS-XL).* In: SKOS Simple Knowledge Organization System Reference, Appendix B. W3C, August 18, 2009. www.w3.org/TR/2009/REC-skos-reference-20090818/#xl