Questions and Answers on Current Developments Inspired by the Thesaurus Tradition

Stella G. Dextre Clarke* and Judi Vernau**

*Luke House, West Hendred, Wantage, UK, OX12 8RR, <stella@lukehouse.org>
**Metataxis Ltd, 8 Thurleigh Avenue, London, UK, SW12 8AW, <judi.vernau@metataxis.com>

Stella Dextre Clarke has until recently been an independent consultant specializing in the design of thesauri and other types of knowledge organization systems. She is probably best known for her work on the national and international standards BS 8723 (Structured Vocabularies for Information Retrieval) and ISO 25964 (Thesauri and Interoperability with other Vocabularies). Her work on standards and on taxonomy development was recognized in 2006 when she won the Tony Kent Strix Award for outstanding achievement in information retrieval. Nowadays she is active as Vice-chair of ISKO-UK and Vice-President of ISKO.

Judi Vernau is a Director of Metataxis Limited, a consultancy specialising in information architecture and information management. She is a qualified librarian who made a move into reference publishing and the electronic structuring and categorisation of content in the 1980s. More recently she has worked on the development of information architectures to support enterprise information management, including the use of ontologies to support knowledge sharing and automatic categorisation. She has taught information architecture at London City University and Victoria University Wellington, and is currently Chair of ISKO-UK.

Dextre Clarke, Stella G. and Judi Vernau. 2016. "Questions and Answers on Current Developments Inspired by the Thesaurus Tradition." *Knowledge Organization* 43 no. 3: 203-209. 5 references.

Received 18 February 2016; Accepted 21 February 2016

1.0 Introduction

In this special issue of *Knowledge Organization*, papers such as those by Kempf and Neubert, and by Tudhope and Binding, point to a confident future for thesauri across the webby world of linked data. But there is less certainty in the corporate world, where public and private sector bodies need to manage often immense volumes of knowledge, information and other content items behind their firewalls. This is the "enterprise space," where White points to a need for knowledge organization (KO) tools but Hjørland expresses doubts about the efficacy of thesauri and especially about their associative relationships. The doubts are unsurprising, given the long-standing experimental difficulty of controlling multiple inter-related variables tightly enough to prove effectiveness. (For references see Dextre Clarke 2016).

Stella Dextre Clarke and Judi Vernau have each worked extensively in that enterprise space, helping clients to find sustainable, cost-effective solutions that are built on sound KO principles. Consultancy has given them the opportunity to apply theoretical results to the realities of practice. Judi (JV) has become increasingly aware of the limitations of the traditional thesaurus, hence the inspiration for the Debate held by the UK Chapter of the International Society for Knowledge Organization (ISKO) in February 2015. Now retired from active consultancy, Stella (SDC) picks up on some points that have arisen during the Debate, and asks Judi about her recent experiences with thesauri and other types of knowledge organization Systems (KOSs).

2.0 Questions and answers

SDC: Q1. During the Debate last February several speakers, yourself included, expressed frustrations about the standards for thesauri – in particular ISO 25964. What is it about them that you dislike?

JV: One impetus for the debate was my perceived need to get some clarity about KO terminology, in particular the words that we use to describe the semantic structures that might be found as part of a knowledge organization system. ISO 25964 defines a thesaurus as a "controlled and structured vocabulary in which concepts are represented by terms, organized so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms and quasi-synonyms" (International Organization for Standardization 2011, 12). Else-

where in the standard it is made clear that the purpose of the terms is to support information retrieval via the mechanism of matching user search terms to thesaurus terms applied to content. There are rules for the form of the terms (for example, plural for countable nouns) and for the types of relationship that are allowed, particularly the generic, partitive or instantial hierarchical relationships. This seems very useful and clear, but many of the KOSs being constructed now do not conform to these rules for a variety of practical reasons. They may be flat lists of terms, hierarchical browse structures, hierarchical tagging structures, ontologies or other structures; there may be a single list or a number of different vocabularies used within one application, organization or subject domain; but in my view these structures should not be referred to as thesauri when they do not conform to the rules and guidelines set out in the standard. So it is not the standard itself that is frustrating, but rather the fact that people (clients, practitioners and academics) often use the word "thesaurus" to describe these other kinds of KOSs.

SDC comment: While I certainly agree that the loose way people use terms like "thesaurus," "taxonomy" and "ontology" is highly confusing, I have given up hope of a solution. Even the LIS profession seems to pay zero attention to definitions in standards such ISO 25964, ANSI Z39.19, ISO 5127, or to glossaries provided by several of our colleagues on their web pages. ISKO has a glossary project under way, and I shall be very surprised if it resolves any of the widespread differences in opinion. Let's move on to a more productive question!

SDC: Q2. So do you gain any benefits from the existence of a standard?

JV: Yes indeed. It's always useful to have an authoritative source to show clients so they don't think we're making these rules up! But seriously, it's a very useful reference tool, and one that I use a lot in training. It provides a language that practitioners, systems developers and clients can understand, and many helpful examples. Facet analysis is crucial, and the standard relationships ensure hospitality. My criticisms would be that in the first place it doesn't discuss the use of concept maps (sometimes also referred to as domain models), which is where I always start when developing a KOS, if only to ensure that everyone involved is agreed on the scope and the key facets; in the second place it doesn't look at how the thesaurus relates to a metadata scheme. This may sound like an obvious relationship, but if you need to call out particular facets to support parametric search and filtering, it is good to know that from the beginning, for example.

SDC: Q3. I have often heard you declare dissatisfaction with the standard thesaural relationships, and I gather that for some of your clients you have been building vocabularies with different inter-concept relationships. Tell me more.

JV: I think two rather different scenarios could illustrate some of this work. The first is the situation where members of staff are tagging important content to support retrieval: recent examples I have worked on are a set of web pages for a support centre, content components (text and images) for a publisher, and documents in a government knowledgebase. For all of these, the information objects are being tagged by people who know the subject area but are not skilled in, or particularly interested in, the theory and practice of KO. The KOSs which provide the tags for them to choose from must be easily comprehended and navigated, which tends to result in a restricted number of top terms, with a much smaller set of total terms which are therefore organized in a looser structure than that recommended by the ISO standard. The UK Integrated Public Sector Vocabulary[1] was a good example of this in that under "information management," some of the narrower terms were "controlled vocabularies," "data analysis," "data security," "disclosure" and "information architecture."

This first scenario paints a relatively simple picture. At the other end of the semantic spectrum I am seeing increasing interest in the use of ontologies. For example, I developed a KOS for the New Zealand Department of Conservation (DOC) which is an ontology in as much as it is a representation of the domain using a strict class model with class attributes and specific relationships between classes (or their sub-classes). Its purpose is to support not just the automatic categorization of documents and records to apply subject and disposition categories, but also in due course to be a source of controlled vocabularies for all DOC systems as well as being a knowledgebase in its own right. For the ontology we identified eleven classes, with a total of 56 sub-classes (see table 1), and built specific relationships between them (see tables 2 and 3) so that a user can find out which species are predators on the blue duck, for example, where it is found in New Zealand (NZ), what its preferred habitat is, etc. The relationships as well as the terms are used to support the automatic categorization.

SDC: Q4. How do the specialized relationships get implemented in your clients' information retrieval software/interface?

JV: These types of custom relationships can have a number of purposes, for example to support data-driven creation of websites (the BBC wildlife site, which I wasn't involved with, is a good public example of this. A full explanation is provided by Howard and Oliver (2011), and the site itself is still viewable at http://www.bbc.co.uk/nature/wildlife, to help publishers create new products, to create a knowledgebase in its own right, or to support the

Class	Sub-class
Asset	Consumable; Equipment; Fleet; IT system; Structure
Authority	Certification; Code of practice; Legislation; Mandate; Permission, licence or pass
Classification	Asset classification; Authority classification; Event classification; Information classification; Party classification; Place classification; Species classification
Earth	Climate; Habitat; Landform; Matter
Event	Business event; Conservation event; DOC internal project; Named campaign; Natural event; Recreational event; Transaction
Function	DOC core business; DOC support business; Third party activity
Information	Data set; Document; Image; Sound; Video; Web page
Party	Group; Organization; Person
Place	DOC management area; NZ boundary; NZ named feature; NZ protected area; World region
Product	Merchandise; Named experience; Publication; Service
Species	Animal; Bacteria; Fungus; Macroalgae; Plant; Protozoon; Virus

Table 1. DOC classes and sub-classes.

Class or Sub-class	Relationship to/from		Class or Sub-class
[All]	Related to >	< Related to	[All]
Asset	Is facility of >	< Has facility	Asset
Asset, Place	Is component of >	< Has component	Named experience
Authority	Authorizes >	< Is authorized by	Authority, Party
Authority	Has object >	< Is object of	Species
Authority	Has subject >	< Is subject of	Party, Place, Species
Authority	Regulates >	< Is regulated by	Event, Function
Authority, Publication	Has format >	< Is format of	Information resource
Classification	Is classification of >	< Has classification	Asset, Authority, Earth, Party, Place, Species
Code of practice	Is guidance for >	< Is guided by	Function
Equipment, Consumable, Information	Is used by >	< Uses	Party, Function
Equipment, Consumable, IT system, Structure	Is sited in >	< Is site of	Structure
Event	Is part of >	< Has part	Event
Function	Has object >	< Is object of	Party role, Party type, Person, Species
Function	Is performed by >	< Performs activity	Party
Group	Is part of >	< Has part	Organization
Information resource	Has subject >	< Is subject of	[All]
Landform, Habitat	Is habitat of >	< Has habitat	Species
Legislation	Mandates >	< Is mandated by	Classification, NZ protected area, Function
Mandate	Mandates >	< Is mandated by	Function, Certification, Classification, Organization
Party	Creates >	< Is created by	Information resource
Party	Has function >	< Is function of	Function
Party, Species, Asset	Is involved in event >	< Involves	Event
Person, Group	Is member of >	< Has member	Group, Organization
Place	Has climate >	< Is climate of	Climate
Place	Has landform >	< Is landform of	Landform
Place	Is covered by >	< Has spatial coverage	Authority
Place	Is location of >	< Is located in	Asset
Place, Landform	Has matter >	< Is matter of	Matter
Service	Service provided by >	< Provides service of	Party
Species, Party	Is predator of >	< Has predator	Species

Table 2. DOC class relationships.

Relationship to/from	
Has abbreviation >	< Is abbreviation of
Has common name >	< Is common name of
Has Destination ID >	< Is Destination ID of
Has LSID >	< Is LSID of
Has Maori name >	< Is Maori name of
Has NZOR concept ID >	< Is NZOR concept ID for
Has scientific name >	< Is scientific name of
Has short name >	< Is short name of
Is AMIS ID of >	< Has AMIS ID
Is evidence term for >	< Has evidence term
Maps to BPS >	< Is BPS term for
Use For >	< Use

Table 3. DOC equivalence relationships.

automatic categorization of content. I mentioned that the NZ Department of Conservation is using the ontology to tag documents with the relevant retention and disposal schedule. This was all part of an enterprise content management (ECM) programme using Oracle's ECM cloud solution over OpenText and Microsoft SharePoint, together with SmartLogic's Semaphore. DOC is maintaining the ontology inhouse, and expects to use it as a knowledgebase in due course. Users can browse the ontology using the graphical interface, which has proved very popular as an easier way to understand the semantic structures.

SDC: Q5. What benefits have ensued?
JV: It is difficult to unpick the benefits of the ontology from the benefits of having an enterprise-wide ECM solution, since the two were implemented as part of the same programme, but certainly DOC would not have been able to apply the subject tags and retention schedules automatically without it. Users now have considerably better access to documents and records, and the initiative has been a real success story which other government and commercial organizations in New Zealand are looking to emulate.

SDC: Q6. How do you determine what types of relationships a given client will need and value?
JV: Of course it comes down to what they are trying to do with the KOS, or think they are trying to do. Several clients have asked for a thesaurus or a taxonomy without being at all clear how they expected it to function! So there are always three first questions to be answered, which correspond well to Rosenfeld and Morville's (2006) context, content and users:

1. What are you trying to achieve? Is the organization trying to categorize content for re-use (i.e. component content management), for retrieval, for dynamic content delivery, to create a knowledgebase, or for some other purpose? Will the tags be applied by subject experts, by all staff, or automatically? What sort of technology is available to support findability and the development of and maintenance of the KOS itself? And of course what's the budget?
2. What is the scope of the content domain and what types of information are involved? What level of content granularity is required? How complex is the domain? I have certainly found that some domains lend themselves to facet analysis and class models more easily than others, and it would be very interesting to explore exactly why this might be so.
3. How do the users want to look for information, and what kinds of functionality and semantic structures will support that? Does the KOS provide support for retrieval, or does it (also) provide information in itself. For example, as I mentioned above, at DOC it will be possible to identify predators of the blue duck, and also what mechanisms can be used for dealing with them.

Knowing what will work best for the client obviously depends on a good understanding of all three of these aspects, but persuading them to adopt a more sophisticated solution like the DOC ontology also depends on helping your client to visualize what that solution might look like (see Figure 1). In the case of DOC we created large posters which showed a number of scenarios involving connections between important entities, and shared them with senior staff, as well as displaying them on the wall in a major office thoroughfare. Staff loved being able to trace these connections, and even added their own.

By the way, many people advocate buying in existing taxonomies (and the standard recommends reviewing what's out there as a first step). I always advise caution around this point: going back to Rosenfeld and Morville's matrix (2006, 24), how often is your context, content and user group the same as someone else's? I have built or reviewed taxonomies/ontologies for at least five different bodies of the UK National Health Service (NHS), and have come to the conclusion that although it is useful to be able to see how others in your domain have done it, there may be surprisingly little overlap between their KOS and what's appropriate for you.

SDC: Q7. How confident are you that clients will be able to maintain the vocabularies you've built for them?
JV: This is always a worry. I believe in providing detailed documentation explaining not just the structure of the vocabulary as it stands, but also the processes for maintenance, and we always build some level of training into any engagement, but it is a particular skill to build and maintain these structures, and while staff may be willing to learn and work at this, there is a big risk that the relationships will not be accurately maintained, resulting in lapses of logic, and a much less useful artefact. There should be a much greater awareness of the importance of KO built into all areas of study, and more opportunities for people who are interested in the field to build up good skills and experience based on sound principles.

SDC: Q8. In organizations that index their content with a controlled vocabulary, what balance do you find between manual and automatic indexing?

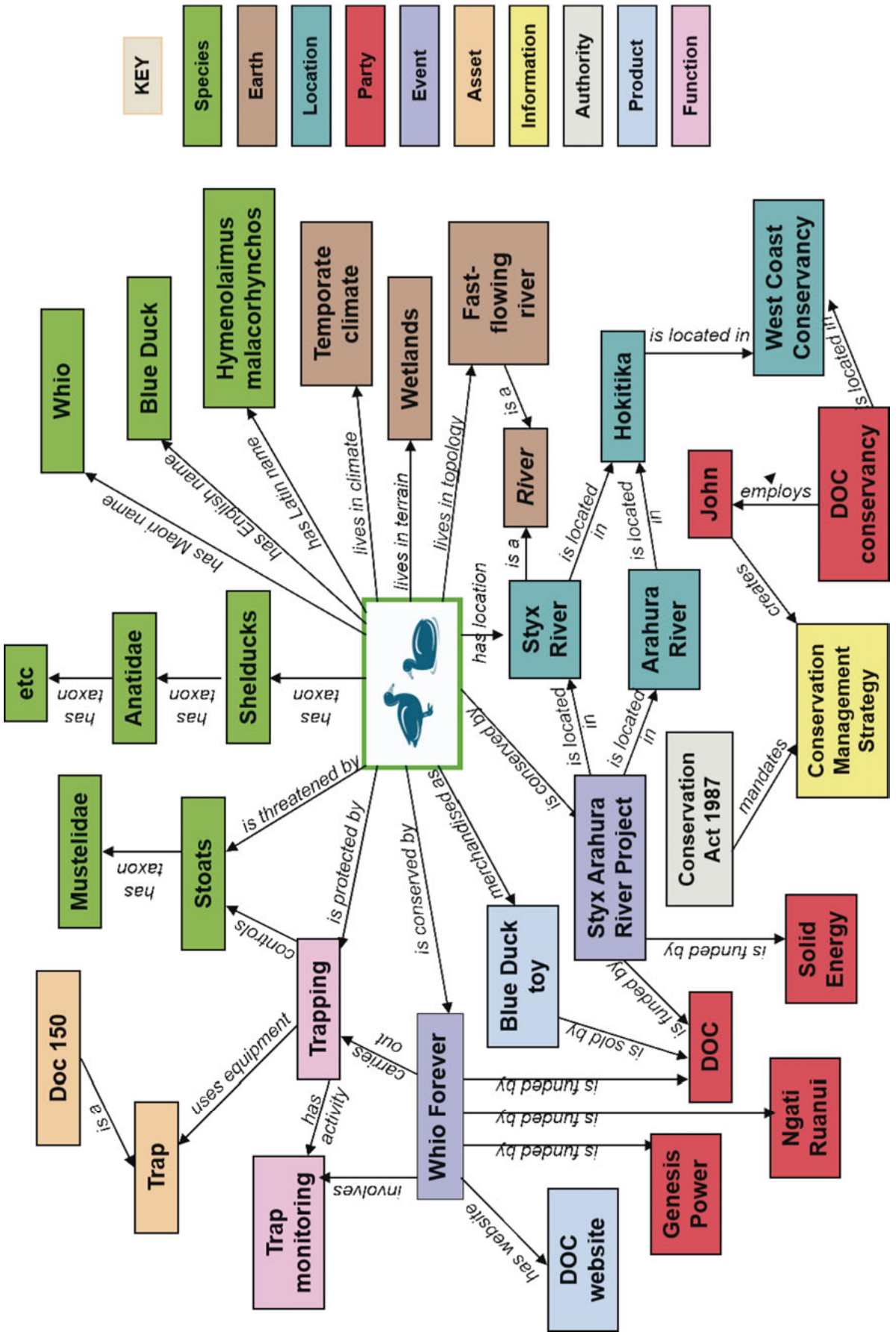

Figure 1. Early examples of entities and relationships.

JV: There is a growing recognition of the value of automatic categorization for enterprise content, as can be readily witnessed by the increase in the number of companies offering this functionality. But automatic categorization still needs to be underpinned by an appropriate KOS, and that KOS still needs continuing maintenance.

Where I have seen automatic categorization work most efficiently is across high volumes of enterprise content. I have yet to see it work well for much smaller knowledgebases and websites, but those kinds of content tend to need very precise tagging to be properly helpful to the end user (whether an inhouse editor or an external searcher). Indeed, I can think of particular sets of content where the wrong tag would be at best annoying and at worst potentially dangerous.

SDC: Q9. What research helps you with the design and development of KOSs?
JV: I found the recent edition of *Knowledge Organization* (42, no. 8) fascinating for its discussion of domain analysis, which has such a large overlap with concept mapping. More research into how domains relate to each other, and the impact of cultural differences, would be useful.

In this special issue Hjørland (2016) makes the point that "the bundling of RTs" [related terms] has "never been properly argued in theoretical or empirical research in information science," and it seems time to ask how useful this generic "bundling" is, when sufficient relationships between pieces of content can be perhaps created by matching metadata or by natural language processing (NLP) techniques, as well as by creating more explicit relationships. It would be very useful to have more insight into the relative merits of these approaches. Any research which provides input into the continuing discussion about return on investment would be welcome.

SDC: Q10. What trends do you see for the continued or evolved adoption of KOSs such as thesauri in the public and/or private sectors?
JV: My sense is that KOSs are becoming increasingly important as organizations try to make sense of the huge amounts of internal content they are amassing, and publishers (formal or "accidental") want to exploit their content by publishing it in more targeted ways. I am increasingly encountering "ontologies," some of which I would call thesauri or taxonomies since they are not actually built on class models with specific relationships between classes, they are just rendered into OWL. But they show an increasing awareness of what's possible.

I think that organizations will lose their squeamishness about the automatic categorization of enterprise documents and records, because they will have no alternative, but only if the categorization is built on a robust KOS, and if the software providers are sensible about pricing! I also think that techniques for automatically categorizing smaller, more focussed sets of content are improving all the time, so it will be interesting to see if they can provide the kind of quality of tagging that a knowledgebase or content for re-use require.

I keep thinking that if we are rigorous in our construction of classes, we will find more and more possibilities for interoperability at least within a given domain to support initiatives around the semantic web and big data, as well as local programmes, but I think we're some way away from that. With the need to share information to support efficiencies as well as joined up thinking, it will become increasingly important to understand why that is, and whether it needs to be so in future.

3.0 Conclusions

The future for thesauri and other KOSs is probably multi-faceted (or at least offers scope for experiment in multiple directions). The work described by Judi Vernau illustrates an expanding opportunity to develop hybrid vocabularies in response to particular situations. Such hybrids may arise from marrying ontology features with those of thesauri and sometimes classification schemes.

Likewise there is an opportunity and a need for teachers to equip graduates with a thorough understanding as well as practical skills in the design and implementation of KOS and metadata schemas. The needs include a good grasp and intuition for the principles, diagnostic capabilities, knowledge of evaluation methods, and ability to work with users and IT (information technology) specialists as well as managers.

There remains considerable scope for KO research into inter-concept relationships—their benefits and how they differ between domains—and into how autocategorization methods might need to differ between different domains and types of collection.

Despite all the endeavours of the KO community, proving the benefits or cost-effectiveness of a thesaurus or any type of KOS is still a challenge. User satisfaction is probably the most feasible measure we have. That leaves plenty of scope for developing more objective and quantitative evaluation techniques.

Note

1. The UK Integrated Public Sector Vocabulary (IPSV) could be described as a thesaurus-taxonomy hybrid. It was at one time accessible on the Internet, but is now visible in outline only at http://id.esd.org.uk/list/subjects. It follows many of the recommendations in the thesaurus standards, but hierarchical relationships

are designed for convenient display and do not comply with the standards.

References

Dextre Clarke, Stella G. 2016. "Origins and Trajectory of the Long Thesaurus Debate." *Knowledge Organization* 43 no. 3: 138-44.

Hjørland, Birger. 2016. "Does the Traditional Thesaurus Have a Place in Modern Information Retrieval?" *Knowledge Organization* 43 no. 3: 145-59.

Howard, James and Oliver, Silver. 2011. "Enhancing the BBC's News and Sports Coverage with an Ontology-Driven Information Architecture." Slides and audio recording. http://www.iskouk.org/content/enhancing-bbcs-news-and-sports-coverage-ontology-driven-information-architecture

International Organization for Standardization. 2011. ISO 25964-1: *Information and Documentation—Thesauri and interoperability with other vocabularies - Part 1: Thesauri for information retrieval*. Geneva: International Organization for Standardization.

Rosenfeld, Louis and Peter Morville. 2006. *Information Architecture for the World Wide Web*. Sebastopol: O'Reilly.