Feature: The Great Debate, 19 February 2015, ISKO UK

Lecture Theatre, British Dental Association, 64 Wimpole Street, London, W1G 8YS

This house believes that the traditional thesaurus has no place in modern information retrieval.

Once upon a time, the thesaurus was venerated. It marked a breakthrough in the retrieval of very specific needles of information hidden in large haystacks. Some of the veneration rubbed off on to the trained information professionals, who alone mastered the occult art of using it to concoct effective search strategies. All this was in the time before we had a computer on every desk, when a collection of 10,000 articles was considered large, and long before the Google era.

But now, who has the patience to consult a complicated thesaurus? Only a dedicated few. Has the thesaurus passed its sell-by date? And even its use-by date? These questions, and more, were tossed around at the Great Debate by a community of enthusiasts. While some limitations of the old-fashioned (?) thesaurus were noted, it still received a happy vote of confidence at the end.

- Judi Vernau (2015) First speaker for the proposition
- Vanda Broughton (2015) First speaker for the opposition
- Helen Lippell (2015) Second speaker for the proposition
- Leonard Will (2015) Second speaker for the opposition
- Cross-examination of expert witnesses
- Martin White (2015) Questions and discussion from the floor

This lively discussion was moderated by Martin White, our Chairman for the day. The thesaurus scored an overwhelming victory! For a blow by blow, listen to the accompanying audio file.

1.0 Event Report

A Summary of the Debate by Judi Vernau

I should immediately say that I was the lead proposer of the motion, so you could be forgiven for supposing that any write-up of the key points from both sides would be totally biased. And you may be right, but in fact all the speakers seemed to be in broad agreement about the need for robust models, controlled vocabularies and relevant semantic relationships—it seemed to me that the bigger question was "what do we mean by a traditional thesaurus."

A word first about the format: the debate was chaired (very ably) by Martin White of Intranetfocus; we had two speakers for the motion, and two against, plus two expert witnesses. After contributions from all of these contributors, the debate was thrown open to the floor, and we had a higher than usual number of contributions, all of which were thought-provoking. In the feedback after the event, most people were very positive about the format as a change from the standard presentation of papers, but some would have liked the speakers (particularly the proposers) to have been more provocative. In the event, I think we were all trying to be too true to our real beliefs, which is understandable but makes for a less exciting debate! I hope we'll use the format again in due course, and if so, we'll encourage the speakers to take a more 'extreme' stance.

As the lead proposer, I started off by defining my view of the 'traditional thesaurus', by which I meant the thesaurus as defined in ISO 25964. It was my aim to show that the rules and relationships defined in the standard are on the one hand too narrow for today's requirements and on the other hand too rigid, since following standard thesaural relationships means that one inevitably ends up with a very large number of terms, many of which become too broad to be useful (e.g., Risk management BT Management). In the modern organization it is usually the case that tags are applied to corporate con-

tent (for example) either by members of staff (who are not trained or even necessarily interested in classification) or by automatic categorization systems. If the former, creating huge vocabularies will make it very difficult for individuals to apply tags; if the latter, you may find that looser structures work as well or even better in applying tags automatically. So the 'pure' BT/NT relationships required by the standard are not always appropriate in creating a tagging structure.

Vanda Broughton, the lead speaker for the opposition, began by insisting that the concept of the thesaurus is alive and well and essential to information management. She presented some humorous slides involving dinosaurs and Captain Hook (not on the same slide!), and emphasized that a thesaurus is a 'central processing unit' and 'one manifestation of an underlying conceptual model' which allows you to identify, control and relate concepts together. Making the case for a thesaurus is making a case for other manifestations of knowledge systems. Shirky and others argue that there is no hierarchy of knowledge, only links, but if you want to manage terminology and if you want to use it to find stuff, you need to impose some ordered structure. So if you argue that a thesaurus, or other knowledge organization structure is artificial, that's OK, because it's just a means to an end. The point of a thesaurus is that it teaches us to take a critical approach: it makes us think about the nature of concepts, about relationships, and about useful labels. But you could use any other information tool that you can mention, such as a domain model, a taxonomy or an ontology – these are all steps on the journey. The basic theory underpins all of these things. The careful critical examination of concepts and relationships is fundamental to all technical solutions: the principles provide an underlying rationale as a discipline.

Helen Lippell, the seconder for the motion, emphasized that thesaurus projects can be expensive, structures can be unwieldy, and if done properly, thesaurus maintenance will add considerably to the cost. There may be organizations that where such investment is appropriate, but there's a big risk that it will be disproportionate to end-user value (which can be hard to define in any case).

Helen's first job was to help construct a large thesaurus and after she left the project it mushroomed out of control, which is another risk. That building and maintenance of such large, unwieldy structures harks back to the age of the intermediary when information professionals were responsible for finding information, and for creating and managing the semantics to support it. Now these things are used by end users they need to be more system and user aligned. It is clearly good to capture terminology and relationships, but they need to be focused on how people use them, and there could be a risk of getting too divorced from user requirements.

In many digital products, search is the predominant functionality for finding information, and Helen naturally does not believe that search alone is sufficient, but she wondered if the thesaurus is too much of a heavy implement. Human intelligence and knowledge can compensate for the relative 'dumbness' of a search algorithm, but other semantic tools are also useful: the addition of synonyms, autosuggest, related queries, seeing the applied tags. These approaches may not by completely ideal, but extra development time needed can be minimal, and the ongoing maintenance time will be less. It is also easier to define return on investment against the application of these smaller components. Saving on resources needed to create and apply these semantic tools is very important (Helen cited her own experience in the media sector).

You could say that there are cheap and basic solutions at one end of the spectrum, and full-on semantic triple stores with millions of RDF statements at the other. If we use a UK supermarket analogy, the basic solution might be Aldi, and the deluxe version might be Waitrose, both of which squeeze out the middle range - thesauri (Tescos etc)

So we need to be confident in our abilities to use semantic structures to solve user needs, and not be tied down to one approach.

Leonard Will drew on biblical references to give examples of (very!) early examples of thesauri, citing the naming of light as day and of dark as night, and Adam naming the beasts and fowls (thus making him the first taxonomist).

Leonard believes that the word taxonomy should be restricted to its biological use, and not used generically, as it frequently is. A thesaurus defines concepts, labels them and links them. Links in the form of broader/narrower relations (as well as scope notes) help to define context. Associative relationships are also useful, but they don't define. Concepts are units of thought, and this underlies KO schemes. Labels are needed to support discourse but definitions of concepts are also needed. Leonard also mentioned the importance of the librarian acting as intermediary; computer systems can do this, but they need aids – semantic structures - to support this. For example, Wikipedia uses these kinds of disambiguation tools and allows you to search broader and narrower terms. This serves two roles, that of map and gazetteer – you can see what the term is but you also have a navigation aid to take you to a more specific term, if required. You can see the subject distribution, and it allows you to see more than you thought you wanted, supporting a wider view of the subject. Leonard showed a slide with the data model for a thesaurus from the standard, which inter alia shows that the thesaurus can be migrated into an ontology. He also touched on metadata schemas and linked data: the

argument that it's too much work to develop a thesaurus ignores the fact that we can use existing thesauri. Co-operative efforts should reduce the effort needed.

He said that controlled vocabularies and other semantic tools are really synonyms for thesauri, all of which have a core principle which is the centre of modern information retrieval.

We then had two expert witnesses, Alan Flett and Phil Carlisle.

Alan made the point that over 5 years of working for SmartLogic he has never been asked to build an ISO standard compliant thesaurus, and in doing a discovery piece for a new project, he does not tend to discover existing thesauri. SmartLogic do use Agrovoc for testing purposes, but otherwise they don't work with ISO-compliant thesauri. They do naturally work with concepts, relationships and labels, but Alan thinks the standard advocates a rather precious and overly restrictive use of terminology. What do clients mean when they ask for software that conforms to the standard? They probably just mean the use of BT/NT etc, so more about how you label relationships rather than the nature of those relationships. Alan is usually involved in modelling, and developing facets and bespoke relationships, and has worked on some big vocabularies, but applying the standard would be impractical: the scale of model would be too big.

For Alan the methodology is usually responsive to the situation, reacting to what's there, and what the users want, and his work is usually focused on autocategorization. He commented that as regards interoperability, you wouldn't use a thesaurus: you would be looking at other mechanisms, and in any case in his experience findability is the big driver rather than interoperability.

Phil Carlisle said that the Data Standards Unit at English Heritage is keen on interoperability, and has therefore developed a common vocabulary to support this, which has also been made available to local government. They have built a piece of software to help support it. Phil does not think that any thesaurus is completely ISO compliant, because flexibility and pragmatism is always needed.

There was originally one national preferred term in the thesaurus, but you need the richness of user-generated terms and lots of synonyms. They are trying to move towards indexing with the concept, so that different communities will see / or have available different terms. Judi commented that if there is no preferred term and they're all equal, this is contrary to the standard. Phil said that you could say the ID becomes the preferred term, with language and dialect variants.

Helen commented that not everyone has the luxury of such formal structures, since there is a constant need to be pragmatic.

Phil agreed that thesaurus development is extremely resource intensive, but thought that there's a benefit to others via linked data.

Alan was concerned about the imprecise use of language: are we talking about thesauri or other things? An ontology is not a thesaurus.

Phil agreed whole-heartedly with this, and thought that the debate was much too friendly and in agreement!

Discussion from the Floor

[Apologies for not naming all the people who spoke: it wasn't always possible to identify them.]

Sarah Saunders (Electric Lane) took issue with the idea that the commercial world doesn't benefit from a thesaurus. With images the thesaurus is key for supporting findability. The accurate and unambiguous results are important. The big problem is that software often doesn't handle it well.

Widad Mustafa El Hadi (Univ of Lille) commented that software is not up to it: we need more sophisticated tools.

There was discussion around the amount of time it takes to develop a thesaurus: organizations know they need vocabularies, but you have to take baby steps sometimes. And a thesaurus has to be maintained in order to be useful.

Linked data: are we getting to a point where you end up with half a dozen knowledge stores globally? There are large thesauri in some fields such as the AAT, and these are growing in number. Linked data will make this easier. But they must grow according to people's needs, and local needs can be very specific. Maybe one way forward is linked data which allow you to share vocabularies.

If people can't find what they want, you might as well not have the information. Projects are underway to link thesauri together. It's an enormous task with no automatic way to do it: it has to be a manual task, which obviously does need a lot of effort.

The traditional thesaurus has hierarchies which can be huge, but it produces silos. You can relate them of course, but a thesaurus does not say how it can be related. We need more semantic relationships such as you get in an ontology – you could try to build it all into a thesaurus, but it's much more efficient in an ontology. The thesaurus concepts are the underpinning, but not enough on their own.

There was discussion on the limitations of the debate question. The question is really what is the place of the thesaurus, or places. What other things do you need as well? Maybe we could look at the contexts in which you might use other tools. So the question is how best to use a thesaurus.

The multi-lingual area is another place where the thesaurus comes into its own, where you need to look at the nuances of difference between a concept in one language and in another, because they. may not be exactly equivalent.

Instability: who curates the thesaurus? But this could be an unstable relationship if there's only one person who understands and someone takes over who interprets things differently. Need to have clear rules and be consistent and stick to the standard. But on the other hand terminology changes and it's hard to deal with that: you need human experts who can understand that this is that.

ISKO UK has recently established a repository called 'ISKO Media', holding the multimedia files associated with each ISKO UK event, and needs to find a way to tag the different artefacts: how should we do that? It's only a small collection of content, so the effort of preparing a controlled vocabulary may be out of proportion. Or should we just use free indexing and see how that looks after a while?

One of the huge weaknesses is that people just often don't understand a thesaurus because it's too huge and complex. You have to be in a situation where you have the tools to use it, but the end users won't understand it.

We need software that really knows how to handle terminology. Is open source the answer to this? There can be a problem with interfaces not explaining what you're getting, and the kinds of relationships in thesauri can be limited, and not made explicit. How can non-professionals choose the correct term: why is this term suitable or not suitable? You should be able to understand the term according to the context or by following the links. The traditional thesaurus doesn't make these relationships explicit in the way an ontology would do. Where you don't know the right word, you may not find what you want by browsing a traditional thesaurus, and the words you need to guide you might need to be more descriptive. What you might need is a classification scheme. Again, the point is made that organizations frequently don't have the skills or the time.

A thesaurus generally provides a viewpoint so sharing is not necessarily workable: an example was given of thesauri for three organizations in the same area where overlap was only 15%!

At this point, Martin asked each of the main participants to give a summary of their view in the light of the foregoing discussion.

Leonard: There is general agreement that the principles of a thesaurus are accepted, but less agreement on whether this must be done rigorously or sloppily. User friendly shouldn't mean sloppy. We haven't really talked about retrieval software, but if we do proper facet analysis, combining these things in retrieval to deal with complex queries, the software should be able to cope.

Vanda: It's all about semantics – we've been a little loose in our response to the question, but we have all interpreted it, and perhaps understand the thesaurus in different ways, but we are reaching the same conclusions.

Judi: We seem to be using all kinds of words to refer to a thesaurus, like ontology or information architecture or controlled vocabulary, and these are clearly not synonyms. Are we saying that we stand by the ISO standard? Clearly not because we're defining it in different way. For example, the URI advocated by English Heritage can't follow thesaurus standards. But we are talking about rigour and models, whether we're building a thesaurus or whatever the KOS is, and we do all agree with that. So it depends how you interpret the motion.

Helen: There is broad agreement, but the problem is back to front – we haven't really discussed whether these systems meet user needs. Different users need different things at different times.

Martin then asked what we really mean by the motion. He commented that search is the only solitary activity that we do, and when you try to do it as a group, you find how differently people approach the task. He asked people to vote by clapping harder for the side they support, and the motion was duly lost.

Thesaurus Debate Needs to Move On

Comment by Stella Dextre Clarke

Surprise, surprise - last Thursday's debate on this proposition was a pushover for the opposition. To defeat any argument of the form "XXX has no place in YYY", all you have to provide is one counter-example.

Just for starters:

- The UK Data Archive, powered by the HASSET thesaurus
- The FAO's AGRIS database, searchable using AGROVOC, and
- EUROVOC, used for searching publications of the EU institutions and others

were among 11 such examples that Leonard Will managed to cram on to one slide. He could have gone on to cite dozens more cases where a thesaurus provides sophisticated and indispensable search capabilities.

The "expert witness" Philip Carlisle backed him up by describing the nine vocabularies and related services that English Heritage built and maintains for the heritage community. Contributions from the floor drew attention to the power of a thesaurus to cross language boundaries, not to mention image searching, where indexing with a controlled vocabulary still outperforms all the other methods.

But simply overthrowing the proposition misses the point – the role of the thesaurus in modern information retrieval has shrunk from what it once was. The high development and maintenance costs of an extensive controlled vocabulary deter most potential implementers. Most users simply do not want to know about such a complicated-looking beast, and so the shy thesaurus needs to perform discreetly but cost-effectively behind the scenes. Given a discerning team of developers, curators, IT support staff and indexers, this sophisticated tool can and should function interoperably alongside statistical algorithms, NLP techniques, data mining, clustering, latent semantic indexing. linked data, etc. Networking and collaboration, not rivalry, are the future.

As the professional body that has grown up around classification, indexing, use of thesauri and other knowledge organization systems, ISKO has a mandate to mark out that future. Follow-up activities could usefully explore:

- The contexts in which the thesaurus is or is not a useful tool;
- how to choose between a thesaurus and another type of knowledge organization system;
- how to integrate a thesaurus with the other components of a modern information retrieval system;
- how to adapt a standard thesaurus to the needs of special contexts; and,
- features of the software needed for thesaurus management.

The knowledge organizer with a grasp of these topics is ideally placed to develop the hybrid vocabulary structures (e.g. a layer of thesaurus model hooked on to upper level ontologies and coated with taxonomy features) needed in today's networked environments.

Posted by Stella Dextre Clarke

Comment

Posted by Stella Dextre Clarke on behalf of Birger Hjørland

> "First I will congratulate with this fine initiative!
> We really need in information science to consider what we are doing and the basic premises on which we are acting.
> I'll provide a few comments here, but I would find it better and more satisfactory to provide comments in our journal *Knowledge Organization*. Therefore here only some short points:
> In my recent paper: "Are relations in thesauri "context-free, definitional, and true in all possible worlds"?" http://onlinelibrary.wiley.com/doi/10.1002/asi.23253/abstract I criticize the claim that "paradigmatic relationships are those that are context-free, definitional, and true in all possible worlds" and that paradigmatic relations are the kinds of semantic relations used in thesauri and other knowledge organization systems. In other words: I see problems in some common norms in standards and understandings of relations in thesauri.
> In another paper in press in Knowledge Organization "Theories are knowledge organizing systems (KOS)", I consider the relations between thesauri and ontologies and argue that "it does not follow that thesauri would not improve, if these characteristics from ontologies were adapted. The question is why thesauri are limited to the relatively few kinds of semantic relations (and therefore tend to bundle different relationships)? As far as I know, there has never been put forward arguments or research demonstrating the functionality of such a bundling. The set of relations used in thesauri have to my knowledge never been theoretically motivated! (They may be intuitively motivated by the need of searchers in online databases to increase 'recall' and 'precision's but this function has never been properly examined and for me it seems unlikely that a broader set of specified semantic relations should not provide better results)."
> There is much more to say about controlled vocabularies in general and their challenge from Google-like systems that need to be explored by our community. But my attitude tend to support the claim "that the traditional thesaurus has no place in modern information retrieval".
> Let us continue this important debate!"

Reference

Hjørland, Birger. 2014. Are Relations in Thesauri "Context-Free, Definitional, and True in all Possible Worlds? *Journal of the Association for Information Science and Technology*. DOI: 10.1002/asi.23253