



Controlled vocabularies vs. full text indexing

A discussion paper by Lesley Mackenzie-Robb

Vantaggio Ltd

www.vantaggio-learn.com

©. Vantaggio Ltd 2010

The debate over the respective merits and advantages of controlled vocabularies versus full text indexing is a loaded one. Keyword search functionality invites the information seeker to type their own words into a search field and implies a search engine which automatically indexes a document's text (e.g., document titles, abstract or description). Browse searching offers the information seeker the opportunity to browse pre-determined information categories, often presented as a hierarchy of menus, and implies an engine which is based on the use of manually applied descriptors, otherwise known as controlled vocabularies. The ultimate goal, whichever method is used to manage information, is to make information easily discoverable, providing the user with accurate returns. In the modern organisation, the ability to store and retrieve information – or knowledge – accurately is a key driver for innovation and competitive edge. The question is, which method of information management is better? Which is easier for the information seeker and more likely to yield more accurate and relevant results? Which method is less likely to yield zero results, or completely inappropriate returns?

This is actually more complicated than it sounds. What I this short discussion paper aims to do is open the lid just a tad and hopefully generate a clearer understanding of the influencing issues.

On the face of it, you would think that using a controlled vocabulary – that is a pre-determined list of terms relevant to a particular domain, which can be hierarchically structured and contain lots of interesting relationships – would be the method most likely to lead to accurate search results. After all, the user selects a category from a list to reveal all contents which have been indexed to that term. Ah but....

Subject indexing or tagging is the process of constructing a representation of a resource that is being tagged. Savoy (2005), for example, suggests that prescribing a uniform and invariable choice of vocabulary terms actually helps in achieving consistency. In tagging a document or whatever resource, you are making a statement about it. According to Vob (2007), this entails a process of

conceptual analysis and translation. It is consequently a subjective process and there is no guarantee of achieving high levels of consistency between human subject indexers. In the simplest terms, the filename that you might give to a document one year may hold absolutely no resonance in your memory when seeking that same document in another year. Human memory is notoriously fickle (Loftus, 2002).

Of course, it is possible for an information collection to be searchable using both browse and keywords. The browse option in this case provides the information seeker with structured guidance – pathways and directions – and insight into the terminology and contents of the collection. Keyword searching is a rapid discovery method of choice when the information seeker knows what they are looking for, and is confident of how to describe it. In an admittedly small-scale study of searching behaviour by library visitors, Tang (2002: as cited in Jansen, Booth & Smith, 2009) found that people's searching was either resource-orientated (e.g., tourist guides on Sicily) or query-orientated (how do you know when pumpkins are ready to pick?). This finding helps us to see how browse searching powered by controlled vocabularies can be useful in one type of seeking, while free text searching powered by both controlled vocabularies and full text indexing can be useful to another.

Much research indicates that people prefer to use key word searching rather than browsing based on pre-determined categories (Shirky, 2005; Waller, 2010). Vob (2007) even suggests that full text indexing has now become the standard for web searching, highly influenced by Google's PageRank algorithm. Others suggest that Google is also responsible for the popularity of keyword searching: bearing in mind that Google's market share in the UK was just short of 90% in 2008 (Waller, 2010), that is hardly surprising. If that is the case, the question still remains: in keyword searching, are the outcomes likely to be better if one applies controlled vocabularies to content, or is it sufficient to rely on text indexing? Is one or the other more likely to satisfy the information seeker?

Before looking at the evidence, it is worth stressing two key factors in this domain of information management. These are, firstly, the notion of language as symbols of meaning and secondly, the notion of language as behavioural interaction. Now this latter might sound a little strange: interacting with what? A computer database? No, using your own knowledge of linguistic symbols to interact with the linguistic symbols of others. Searching, in this sense, is a form of discourse. Fortunately, there is a huge literature providing all sorts of useful and relevant insights into human discourse from the perspective of discourse as action.

Unfortunately, as far as this particular domain is concerned, there is a dearth of solid, empirical evidence-based research. Much of the research on data classification, for instance, comes from computer science rather than the social sciences (Vob, 2007). This is somewhat intriguing when one considers that information seeking and indeed content classification are both forms of behaviour which, in their own right, have meaning, motivation and influencing factors. Mai (2008) is one of the few researchers who places the design, construction and use of controlled vocabularies firmly in an analysis and understanding of the context of their function. That is, the vocabulary is itself an expression of a human activity or concept. Taking this approach a step further, the vocabulary is a product of language which Whorf (1940) famously described as the "shaper of ideas, the programme and guide for the individual's mental activity". Taking this notion yet another step further, it can also be argued that there is a world of difference between the spoken word – the spontaneous

thought enquiry translated into a form of words – and the word that is written in a document. It is this aspect of searching behaviour – searching discourse in action – which has been neglected. Yet it is the one most likely to give us the answers to our questions.

On the other hand, there is a considerable amount of apparently influential but heavily opinionated viewpoints. Shirky (2005) uses dubious arguments and examples to debunk the existence of what he calls ontologies which, according to Shirky, demonstrate high levels of bias amongst other flaws. In this context, the “bias” to which he refers is so relative as to be completely meaningless. Gergen (1974) argued that “labelling biases pervade our literature”: he was referring to the Social Psychology literature, but the statement is true of any human endeavour. There is no such thing as an objective observer or participant. A vocabulary is the product of its context, time and socio-cultural norms.

One study by Savoy (2005) evaluated search outcomes in a comparison between free-text indexing and manually assigned classification terms. Previous studies have shown that the use of controlled vocabularies improves search precision, but that these improvements are not statistically significant. Savoy’s study confirmed this, but also concluded that the best precision results were always consistently achieved when both manually assigned classification terms and automatic text-word indexing are combined.

Similarly, Waller (2010) investigated user search logs, drawn from a 4 year period, generated by people using the online catalogue of a large public library. Waller suggests that people tend to use key word searching because it offers the line of least effort, while a study by Jansen, Booth and Smith (2009) noted that people tend to use nouns in their keyword searching. In other words, when searching for information, people use the strategy of least cognitive demand. This is consistent with psychological models of problem solving and decision making, both of which have been linked to information seeking as explanatory frameworks (Jansen, Booth & Smith, 2009).

Waller’s study found that 20% of people undertook only one search and that one in five of those who got zero hits on their first search abandoned any further attempt. In other words, people are impatient when searching. She also found that “keyword anywhere” searching was most likely to result in more than 1000 hits. Also, 66% of searchers do not go beyond the first page of results. More impatience. 10% of all failed searches involved misspellings. What all this shows is that keyword searching can be a frustrating business. People are generally not “trained seekers of information”. Organisations do not typically put their staff through a training course on how to search for information on intranets and the internet. Perhaps they should.

One interesting paradigm seeks to explain search behaviour in the context of a learning process framework (Jansen, Booth & Smith, 2009). Their notion is that searchers will select a search strategy based on their mental model of the information requirement. That is, how much they know about what they are looking for, and how well they know how to use any available system to search. Or, put another way, how well they are trained and experienced in information seeking. In this case, the user’s knowledge of the search engine combined with their knowledge of the topic of search will determine strategy. If, as Marchionini (2006: as cited in Jansen et al, 2009) proposes, learning is the development of new knowledge, then is information searching a learning process? If it is, Jansen et al argue that such a learning theory can be used to predict online searching behaviour, and therefore

to inform system design. Information seeking probably is a learning process in the experiential learning tradition, but I think it is far more complex than this.

It would seem, in conclusion, that the safest option is to offer the information seeker a choice of both search methods. Savoy (2005) found that using both automatic text indexing and controlled vocabularies was most likely to result in accurate search outcomes. We can also offer two very practical suggestions to help maximise search precision and satisfying outcomes. First, to address the incidence of misspellings, it would be useful to build in a “Did you mean...” functionality. Secondly, vocabularies which are to be used as browsing structures – partially or wholly – should not display any more than 10 sub-menu items. More than this will lead to increased cognitive load on the user, and confusion. Most importantly, design strategies for search and discovery should take account of searching behaviours and the discourse used by searchers. And a little training might go a long way.

References

Gergen, K. (1973). Social Psychology as history. *Journal of Personality and Social Psychology*, (26), 2, pp 309 - 320

Jansen, B., Booth, D., and Smith, B. (2009). Using taxonomy of cognitive learning to model online searching. *Information Processing and Management*, 45, pp 643 – 663

Lotus, E. (2002). Memory faults and fixes. *Issues in Science and Technology*, pp 41 - 50

Mai, J. (2008). Actors, domains, and constraints in the design and construction of controlled vocabularies. *Knowledge Organization*, 35, (1), pp 16 – 29

Savoy, J. (2004). Bibliographic database access using free-text and controlled vocabulary: an evaluation. *Information Processing and Management*, 41, pp 873 – 890

Shirky, C. (2005). Ontology is overrated: categories, links and tags. [Online] http://shirky.com/writings/herecomeseverybody/ontology_overrated.html. Accessed 7 July 2010.

Vob, J. (2007). Tagging, Folksonomy & Co – Renaissance of manual indexing. *Proceedings of the 10th International Symposium for Information Science*, Cologne.

Waller, V. (2010). Accessing the collection of a large public library: an analysis of OPAC use. *Library and Information Science Research Electronic Journal*, (20), 1

Whorf, B. (1940). Science and linguistics. *Technology Review*, (42), 6