Controlled vocabularies, thesauri, and taxonomies

Heather Hedden

Controlled vocabularies, thesauri, and taxonomies comprise a field that is closely related to indexing. Some indexers already do work in these areas, and others could find themselves called to do such work soon. Therefore, it is important for indexers to be familiar with these tools/methods for organizing information. This brief article defines the concepts more fully.

Controlled vocabularies

The term 'controlled vocabularies' covers the full range of these tools for organizing information retrieval. At a minimum, a controlled vocabulary is a restricted list of words or terms used for indexing or categorizing. It is controlled because only terms from the list may be used for the subject area covered by the controlled vocabulary. It is also controlled because, if it is used by more than one person, there is control over who adds terms or how terms can be added to the list. The list could grow, but only under defined policies. Most controlled vocabularies have the additional feature of *See*-type cross-references pointing from a non-preferred term to the preferred term. The objectives of a controlled vocabulary are to ensure consistency in indexing, tagging, or categorizing and to guide the user to where the desired information is.

A controlled vocabulary has many uses in indexing. An indexer can create a simple controlled vocabulary for his or her own individual use if working on a large project, such as more than one volume or a series of articles. Controlled vocabularies are especially useful for providing indexing consistency between several indexers contributing to the same index. This is particularly the case for periodical or database indexing, or web page meta-tag keyword indexing. Sometimes controlled vocabularies are referred to as 'authority files,' especially if they contain just named entities.

We may criticize search engines for their deficiencies in picking up just *any* words within documents, but if the search engine is configured to retrieve documents based on what is in a document's keyword field, and keywords have been assigned by indexers taking them from a controlled vocabulary, then good results can be achieved. Not as accurate as human indexing, but still better than simple free-text search, is the use of search engines and controlled vocabularies in conjunction with automated indexing or auto-categorization. The controlled vocabulary's synonyms or variants associated with each term enable document retrieval even when the words entered by the indexer into a search box do not exactly match any text in the relevant document. If used behind the scenes with a search engine and never displaying a browsable list for the user, the distinction between preferred term and non-preferred term is moot. Instead you simply have a set of synonyms for each concept with no one term being seen as the preferred term. This type of controlled vocabulary is therefore called a 'synonym ring.'

Thesauri

The classic meaning of a thesaurus is a kind of dictionary that contains synonyms or alternative expressions for each term, and possibly even antonyms. A literature retrieval thesaurus shares this characteristic of listing similar terms at each controlled vocabulary term entry. The difference is that in a dictionary-thesaurus all the associated terms *could potentially* be used in place of the term entry depending upon the specific context, which the user needs to consider in each case. But in certain contexts some of these terms are not appropriate. The literature retrieval thesaurus, on the other hand, is designed to be used for all contexts, regardless of a specific term usage or document. The synonyms or near synonyms must therefore be suitably equivalent in all circumstances. A literature retrieval thesaurus must clearly specify which terms can be used as synonyms (called 'used from'), which are more specific (narrower terms), which are broader terms, and which are related terms.

A thesaurus, therefore, is a more structured type of controlled vocabulary that provides information about each term and its relationships to other terms within the same thesaurus. National and international standards have been developed to provide guidance on creating such thesauri, including ISO 2788, ISO 5964, ANSI/NISO Z39.19, and most recently updated BS 8723. The standards explain in great detail the types of relationships that fall into three types: hierarchical (broader term/narrower term), associative (related term), and equivalence (use/used from). A thesaurus also includes scope notes to clarify usage of some or all terms. The greater detail and information contained within a thesaurus compared with a simple controlled vocabulary aids the user (whether the indexer or the literature searcher) in finding the most appropriate term more easily than in a simple, unstructured controlled vocabulary. A thesaurus structure is especially useful for a relatively large controlled vocabulary that involves human indexing and/or supports a display that the end-user (searcher) can browse.

Heather Hedden

Taxonomies

The word taxonomy means the science of classifying things, and traditionally the classification of plants and animals. It is becoming a popular term now for any hierarchical classification or categorization system. Thus, we no longer speak of 'taxonomy' as a science but rather 'a taxonomy' (plural: taxonomies) as a kind of controlled vocabulary that has a hierarchy (broader term/narrower terms), but not necessarily the related-term relationships and other requirements of a standard thesaurus.

Recently the term taxonomy has also become popular as the term for *any* kind of controlled vocabulary, whether a structured thesaurus, a simple synonym ring, or anything in between. This is especially the case in the corporate world, where one might speak of 'enterprise taxonomies.' It's simpler to have a one-word term for the concept of controlled vocabularies, especially when speaking of the people involved: 'taxonomists' instead of 'controlled vocabulary creators/editors.'

Conclusions

The individual who creates and edits controlled vocabularies should simply adopt the language of the client, whether it is controlled vocabulary, thesaurus, or taxonomy, while also determining what the specific requirements are, independent of what it is called.

Creating controlled vocabularies is related to indexing with respect to the objectives. Furthermore, the skills needed for indexing are some of the same skills needed for developing controlled vocabularies. These include deciding how best to word a concept; creating a broader term/narrower term hierarchy, like subentries for main entries; creating used-for terms, like *See* references; creating related-term relationships, like *See also* relationships; and deciding what is important and likely to be looked up.

Taxonomies/controlled vocabularies is a growing field, largely due to the growing volume of information, especially in electronic form. Controlled vocabularies used to be created just for library literature retrieval databases, but more and more companies are finding that their growing repositories of internal data require categorization to aid in retrieval. While totally inappropriate for back-of-the-book indexing, automatic indexing may be the only alternative when dealing with tens or hundreds of thousands of web pages or other documents, especially if they are constantly changing or growing in volume, and controlled vocabularies can be very useful.

I intend in the next issue of *The Indexer* to provide a comparative evaluation of three software programs that freelancers can use to create and edit thesauri or any controlled vocabularies. In the meantime, indexers can learn more about this field by visiting the website of the new Taxonomies & Controlled Vocabularies Special Interest Group, www.taxonomies-sig.org.

References

About Taxonomies & Controlled Vocabularies (2007) *Taxonomies & Controlled Vocabularies Special Interest Group*. www.taxonomies-sig.org/about.htm (accessed 13 January 2008).

Agee, V. (2008) Controlling our own vocabulary: a primer for indexers working in the world of taxonomy. *Key Words* **16**(1), 30–1.

Leise, F. Fast, K. and Steckel, M.(2002) What is a controlled vocabulary? *Boxes and Arrows* [online] www.boxesandarrows.com/view/what_is_a_controlled_vocabulary_ (accessed 13 January 2008).

Heather Hedden *is an information taxonomist with Viziant Corporation and teaches an online workshop in creating taxonomies and controlled vocabularies through the continuing education program of Simmons College Graduate School of Library and Information Science. She is also manager of the Taxonomies & Controlled Vocabularies Special Interest Group of the American Society of Indexers. Email:* heather@hedden.net

Letter

Who are we indexing for exactly?

I was mortified in looking in the 'Indexes censured' section of the October 2007 *Indexer* to see an index I had done. Finally, I thought, a review of an index I had done, after 30+ years of indexing, and it's a rotten one. I'm sure we all turn to the reviews pages before anything else when *The Indexer* arrives, rather like someone looking, in a biography/autobiography, for their own name before reading the actual text.

Of course, the fault lies not with 'Indexes reviewed', which makes it clear that it passes no judgment of its own on the views of reviewers, or particularly with the publisher of the original review. (Although just maybe, it would be good if a review, if controversial, could be sent to the original publisher/author of the book in question before being published, so that at least the other side of the argument could be put. This happens often with some Letters to the Editor, particularly in journals. Any controversial letters commenting on a previous paper nearly always have a reply from the author or editor below the letter with a response as to why something was said or done.)

However, I was so mortified, and then annoyed, that I thought to put pen to paper and explain why the dear reviewer in my eyes was wrong.

The book in question is *The Economist style guide*. The reviewer's comments that got my hackles up were as follows:

> There is an index of limited usefulness (and no help to me) when . . . I decided to check whether *The Economist* prescribed roman type or italics for the Latin expression 'ad hoc' [the reviewer chose roman]. I found the answer, but only after several minutes' hunting.

In a panic I rushed to view my index and to see why the reviewer had had difficulty. The reader's note says: 'Spellings or meanings of particular words or phrases are given in the A–Z section of Part 1 and are not repeated here'. The reason for this is that the publishers had asked me not to repeat in the index all the head words in the alphabetical section (unless there were more throughout the text that would justify it), or the words or phrases in the many lists that were included in this book. Otherwise, the index would have been enormous, probably bigger than the text itself!

So to the 'ad hoc' problem. The phrase was not in the index under A (see reader's note above as to why), but it was easily found under *Latin, words and phrases*, and *italics, foreign words and phrases*. Interestingly, the reviewer would have had a problem in deciding whether to use roman or italic. He chose roman, and this would have been right. 'Ad hoc', under I for italics, came in a list of phrases of foreign words that have become so 'familiar that they have become anglicized and so should be in roman'. But he could have chosen italics because under L for *Latin, words and phrases*, 'ad hoc' is in italics. So the author and copy-editor were at fault here and should have spotted the problem. On reflection, it might have been a good idea to have had an entry under *fonts* or *typeface*.

Did the reviewer have difficulty because he couldn't find the phrase 'ad hoc' in the index itself or because he couldn't think of the headings it might be under, i.e. *Latin* or *italics*? So back to my title – who are we making indexes for? As an indexer I was thinking of terms that the phrase 'ad hoc' might be looked under. As a publisher, you are looking for the cheapest possible option – so not too long and not too expensive to produce. As a reader, and I know I have done this myself, you look for the word you are having problems with, and get annoyed when it's not there.

As it happens, the editor did suggest that, with hindsight, it would have been better to have done this index like the actual *Economist* in-house style book, i.e. indexing every word that comes in the book. But that, of course, would have been a concordance, not an index so no way forward there.

The answers:

For the indexer: Pay us more, give us more room, and the reader will get a perfect index.

For the publisher: Resign yourself to more expense, but be pleased that your reader is happy.

For the reader: Get a life. If at first you don't succeed, try again and all will be revealed in the text. Become educated in the way of the index. Don't make comments in a public place about something you have no understanding of, nor care about in the scheme of things, so that you don't know why 'ad hoc' wasn't in the index in the first place.

Michèle Clarke, indexer, with tongue in cheek

Addendum

Robert Fugmann has asked us to make the following correction to his letter printed in the October 2007 issue of *The Indexer* (**45**(4), 268). The last sentence in the second paragraph should have read:

> Furthermore, the subheadings in the 'informative' index are detailed in a way that is common for the annotations for entire books in library practice.