

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.

