Back of the Book Indexing and Thesaurus Creation: A Comparison

BY HEATHER HEDDEN

I have worked as a back-of-the book indexer, a periodical database indexer, and a controlled vocabulary/thesaurus editor. These three areas of work are inter-related, and the skills required are similar. Book and periodical indexing are related, of course, because they are both indexing activities (reading texts, discerning the main ideas, and providing index terms that point to the text). Periodical indexing and controlled vocabulary work are related, because periodical indexing makes use of index terms managed in controlled vocabularies, and the same person may perform both tasks. As for the relationship between back-of-the-book indexing and creating controlled vocabularies, that may not seem as obvious. In fact, the actual work of designing and creating a thesaurus type of controlled vocabulary does have more similarities to back-of-the-book indexing than it does to periodical indexing.

What is a Thesaurus?

For some background, a thesaurus (as a knowledge organization system, not a writer's type of dictionary) is a specific kind of "controlled vocabulary." A controlled vocabulary is, as its name implies, a controlled list of terms. Only terms on the list may be used for indexing, so a controlled vocabulary (also called a "taxonomy") ensures consistent indexing. The same concept will then always be indexed with the same term. A controlled vocabulary is particularly useful when indexing a series of volumes or issues of periodicals over time, lest the indexer forget what term to use for the same concept. It is also useful when multiple indexers are indexing the same project, because each indexer may otherwise come up with different terms for the same concept. Unless the controlled vocabulary is small enough to browse through within a single page, it typically includes *See* cross-reference terms within it to help the indexer find the approved index term.

A thesaurus is a structured kind of controlled vocabulary, where, in addition to having the *See* references, approved index terms are also related to each other. Terms can be related to each other in one of two ways: (1) hierarchically as broader and narrower (a.k.a. generic/specific, parent child, superordinate and subordinate), or (2) associatively as merely related, like *See also* terms. For a freelance indexer needing to create one's own controlled vocabulary for a large project, a thesaurus-type of controlled vocabulary probably has more structured detail than necessary, but a thesaurus is quite common to support the indexing of periodical literature and is essential for operations that index multiple periodicals as part of the same index (e.g., PsychINFO, ERIC, Medline, EBSCO, Reader's Guide to Periodical Literature, etc.) What is interesting about a thesaurus is that the same thesaurus is used by both the indexer (as an aide in consistent indexing) and by the end-user information seekers, as the periodical index.

As with any controlled vocabulary, terms in a thesaurus are listed alphabetically. Under each term is then listed any relationships or information, including: Broader Term (BT), Narrower Term (NT), the See reference terms (technically called "nonpreferred terms" or "nondescriptors") that it is used from (UF), and possibly even a scope note (SN). The nonpreferred terms (See reference terms) are interspersed alphabetically, just like the See references with an index, and they are followed by *See* or *USE* and the preferred term that the thesaurus user should use. Following is an example of a term in a thesaurus:

Recreation facilities
UF: Recreation centers
BT: Public buildings & facilities
NT: Amusement parks
NT: Athletic facilities
NT: Bowling alleys
NT: Golf courses
NT: Stadiums & arenas

Heather Hedden is a taxonomy consultant with Project Performance Corporation, an instructor of the continuing education online course "Taxonomies & Controlled Vocabularies" through Simmons College Graduate School of Library and Information Science, author of the book *The Accidental Taxonomist* (Information Today, 2010), and founder/past manager of the Taxonomies & Controlled Vocabularies SIG of ASI. She will be presenting a pre-conference workshop "Taxonomy & Thesaurus Creation" at the 2012 conference in San Diego.

Points of comparison

Back of the book indexing and thesaurus creation are similar because back of the book indexing is not just "indexing" it is also "index creation," and the design and creation of an index is not too different from the design and creation of a thesaurus. Both activities center on creating index terms or entries for concepts (called entries in indexes and terms in thesauri), arranging and structuring them, and dealing with their variants/synonymous terms (multiple entry points). This is different from most periodical indexing, because periodical indexing usually relies on the use of pre-existing terms in a controlled vocabulary or thesaurus.

Following is a comparison of the methods to create common features for both for both back-of-the-book indexing and thesaurus creation in three areas:

1. multiple points of entry
2. hierarchical structure
3. indication of related concepts

1. Multiple Points of Entry

The indexer or thesaurus creator (sometimes called "taxonomist" for want of a one-word name) creates multiple terms with the same meaning for a single concept as multiple entry points, for the purpose of directing the users, who are prone to use various terms that mean the same thing, to the same location in the text. The types of variants are the same for both book indexes and thesauri. They can be synonyms, near synonyms, phrase inversions (such as digital photography and photography, digital), abbreviations or acronyms and their spelled out forms, slang vs. formal language, etc.

In back-of-the-book indexing, there are two methods for creating multiple entry points:

1) Double-posts, whereby two or more index entries of the same meaning are added to the index, with none designated as preferable to the other, and all have the same locators pointing to the same points of text. (This is the preferred method when the entries do not have subentries.)

2) *See* cross-references, whereby additional index entries of the same meaning are added to the index, but they point to a single favored entry, which is the only one with the locators. (This is the preferred method when the entry has subentries.)

In thesaurus construction, on the other hand, there is basically just one method for creating multiple entry points, which is akin to the index *See* reference. There is nothing like double-posts (different terms for the same concept having equal standing) in a thesaurus. The *See* reference terms are called "nonpreferred" terms, and they each direct the user by pointing (whether by the word *See* or *Use* or simply by a hyperlink) to the "preferred" term that is used in indexing. Furthermore, while a *See* reference in an index only points in one direction, in a thesaurus this is considered a bidirectional reciprocal relationship. So, in additional to "nonpreferred term *Use* preferred term," we also have "preferred term *Used* from nonpreferred term."

Another feature of a thesaurus, polyhierarchy, while not exactly another "point of entry," serves as an additional way for users to navigate to the specific term and content they want. This is explained in more detail in the next section on hierarchical structure.

2. Hierarchical Structure

Both indexes and thesauri have some form of hierarchical structure between terms/entries to guide users from more generic concepts to more precise topics. If a term has too many locators/references, it needs to be broken out by creating multiple corresponding subordinate concept entries/terms. In indexes these are called subentries, and in thesauri these are called Narrower Terms (or sometimes "subtopics"). In both situations, indexing may be done using the main entry/Broader Term for a general treatment of the subject and also with its subentries/Narrower Terms for specific aspects or sub-topics.

Despite their similarities, though, subentries in an index and Narrower Terms or subtopics in a thesaurus are not the same and do not function identically. Index subentries serve as subdivisions or more specific aspects of the main entry. The options for subentries are limitless. For example, the main entry of a country name, such as Russia, could have subentries of periods of history, relations with other countries, politics, economics, ethnic groups, religions, etc. By contrast, Narrower Terms in a thesaurus are limited to (1) specific examples or instances of a more generic term, or (2) parts of a whole. In the case of a term of Russia, the only Narrower Terms would be geographic subentities, such as its subregions, republics (states) or cities, for example, Siberia or Moscow.

Just because both a subentry displays indented under a main entry, and a thesaurus Narrower Term displays indented under a Broader Term does not mean they are the same thing. As an *aspect* of a main entry, index subentries must be related to the main entry and can be mere adjectives or prepositional phrases. Narrower terms in a thesaurus, on the other hand, are standard terms themselves which can stand on their own, so must be nouns or noun phrases as all thesaurus terms are. In fact, "Narrower Term" is only relative to another term; "Narrower Term" and its recip-

rocal "Broader Term" designate *relationships* between terms and not types of terms. The fact that a main entry and a subentry in an index can be interchanged or "flipped" is indicative of the fact that it is their mere pairing that is significant rather than any intrinsic hierarchical (parent-child) relationship. You cannot switch the position of a Narrower Term and Broader Term in a thesaurus.

Index subentries and thesaurus Narrower Terms lend hierarchical structure that is evident when the user skims the entire index or thesaurus. In an index, a single sub-level of subentries is most common, although sub-subentries, bringing the total number of levels to three, are also frequently seen. Deeper sub-entries, however, should be avoided. In a thesaurus, however, hierarchical levels to the depth of four or five are not uncommon, nor are a problem. What is especially interesting about a thesaurus structure is that it can be generated in different displays: hierarchy for each top-level term (as is done in an index), hierarchy for *every* term listed alphabetically, or the immediate Narrower Term and Broader Term indicated for a selected term but not the full hierarchy. Finally, a thesaurus may even have terms with such broad meaning, such as *Countries*, that they are not even used for indexing, but just for grouping Narrower Terms. All terms in an index, by contrast, are used for indexing.

As mentioned previously, in some cases in a thesaurus, a specific Narrower Term may have more than one Broader Term, a feature called "polyhierarchy." This helps users who are browsing the thesaurus hierarchically and may start from a different Broader Term. The feature of polyhierarchy is not the same as the re-use of the same subentries under different main entries in an index. A subentry has a specific meaning in relationship to the main entry it falls under and restricts. A thesaurus term, even as a Narrower Term to another term, is completely independent in meaning from the Narrower Term. The same term that appears under more than one Broader Term has the exact same meaning and thus points to the same content in both cases.

3. Indication of Related Concepts

A third feature of both book indexes and thesauri is to have methods to indicate related topics of possible interest to the user. In an index, the designation is *See also*, and in a thesaurus it is a relationship called *Related Term* (RT). In both cases, they function similarly. They can appear at different levels of the hierarchy, and it tends to be a judgment call of the indexer or thesaurus creator as to when creating such a reference would be helpful without having so many that they get in the way of the

ease of use of the index/thesaurus.

The situations in which the Related Term relationship can be created in a thesaurus, however, are more limited than when the See also reference can be created in an index. As mentioned previously, there are strict rules about when the hierarchical (Narrower Term/Broader Term) relationship can be created in a thesaurus, as the Narrower Term must be a specific type or instance of a Broader Term or an integral part of a whole Broader Term. By extension, if two terms in a thesaurus have a broader-narrower relationship, then they must be arranged in such a hierarchy and consequently cannot be Related Terms. For example, in a book index, you could have Cancer *See also* Melanoma (a kind of cancer). In a thesaurus, however, your only option is to have Cancer NT (Narrower Term) Melanoma; Cancer RT (Related Term) Melanoma would be incorrect, no matter if it's in addition to or instead of the other, hierarchical relationship.

Another difference between *See also* references in an index and Related Terms in a thesaurus is that the *See also* reference is not necessarily always reciprocal, whereas the Related Term always is. In an index, if one term points as *See also* to a second term, usually, *but not always*, there is also a *See also* reference from the second term back to the first. For example, the indexer might choose to have Education *See also* Teachers, but not to have Teachers *See also* Education (with the reasoning that people looking up Education might be interested in but not think to also look up Teachers, but people looking up Teachers would know to look up Education). In a thesaurus where Related Term relationships are used, it is mandatory to have the relationship at both terms pointing in both directions.

Skills and Process Comparison

The skills used by book indexers and by thesaurus creators are similar. Both involve content analysis and term creation. Indexing requires perhaps greater specific content analysis, since the indexer must read the entire text and index to specific passages to identify topics and names that are important, whereas thesaurus creators only look at sample texts. Thesaurus creators, on the other hand, must do analysis that is more broad-based. They often need to examine other existing thesauri on the subject. Thesaurus construction also requires more study of the "audience" or users, who will use the thesaurus and considers how it will be used, something that book indexers almost never do.

Thesaurus creation *may* also demand more subject-specific knowledge than does book indexing. Book indexers usually require subject expertise only for technical and scholarly subjects. Generally, book indexers can let the book be their guide for terminology and structure. The thesaurus creator, on the other hand, does not have a single work to rely on, so must consult multiple sources. These multiple sources may cover the subject area unevenly or even conflict with each other, so that is where subject-matter knowledge can be helpful.

The process of book indexing and the process of thesaurus creation may be similar depending on the approach taken. Depending on the book and the indexer, back-of-the-book indexers may take one of either two approaches: (1) after a cursory review of the table of contents, read and index page by page from the beginning, or (2) skim the book and write down common themes, topics, and names, as likely index terms, considering even which will be broad topics with many subentries, and then go back and begin indexing. Thesaurus creation is similar to the second book indexing approach, but then additional work is put into refining the terms, and the indexing stage is not necessarily even performed by the thesaurus creator.

Style and Practice Comparison

Finally, there are a few points of comparison regarding editorial style and "best practices" in creating indexes vs. thesauri. The wording of the terms is usually the same in both cases, being both concise, yet unambiguous. It's easier to skim a list of terms that are 1-3 words in length, than those that are 4-5 words long, although proper nouns may need to be longer. Countable nouns should be made plural in both indexes and thesauri

There are differences in style, though. A book index is designed for alphabetical lookup only. Therefore, the first word of an entry should be a "keyword" that is likely to be looked up. Consequently phrase terms are often inverted (noun, adjective) so that the important word comes first. Thesauri, on the other hand, are designed to be looked up both alphabetically or hierarchically, and possibly in additional variations of each. Furthermore, thesauri are usually accessed electronically, not in print, which means that terms can be searched for and not just browsed. Therefore, the standard practice is *not* to invert the words within preferred terms (although nonpreferred terms may be inverted).

There are no strict rules about capitalization for either indexes or thesauri. However, it is more common to see lower-case entries (when not proper nouns) in book indexes, and it is somewhat more common to have title case for terms in thesauri.

Finally, there is the issue of balance in the structural design. In a book index, some main entries have no subentries, some have a few, and some have so many that they go on for more than a column of a page and then may have additional sub-subentries. There is no need to strive for a balanced looking structure of an index, since an index should simply represent the reality of the subject scope of its book. In thesaurus design, however, some degree of structural balance is an issue. Having terms that lack any broader or Narrower Terms, called "orphan terms," is often avoided. The "top terms" of a thesaurus, those hierarchically on top with no Broader Term ideally should have some degree of balance with respect to number of Narrower Terms under each.

In conclusion, due to the number of similarities between back-of-the-book indexing and thesaurus creation, a person with interests and skills in one area might consider work in the other area. Due to the number of differences between back-of-the-book indexing and thesaurus creation, though, anyone transitioning from one activity to the other needs to study the other field first and not make assumptions about any apparent similarities. ●