

# Taxonomies and the Information User

PEOPLE APPROACH INFORMATION FROM DIFFERENT PERSPECTIVES, AND A WELL-DESIGNED KNOWLEDGE ORGANIZATION SYSTEM IS VITAL TO ENSURING ALL OF THEM CAN FIND THE INFORMATION THEY NEED.

BY HEATHER HEDDEN

**S**pecial libraries serve the information needs of a diverse range of clients, including researchers, administrators, executives, partners, contractors, vendors, clients, investors, regulators, educators, and the general public. Aside from determining the kind of content to make available to each group, librarians must also decide how the content should be organized, displayed, and searched.

How do we serve these different information users and their diverse needs? A librarian could, of course, interview members of each group and tailor information services to the groups' needs. But the number and types of potential information users have grown far

beyond what a librarian can personally serve, and as more and more content providers have begun targeting information consumers, many users have come to assume that they should seek information themselves. Without a librarian as a personal intermediary for every information user, the means of delivering information must be adaptable to a variety of information needs.

Making an information system flexible and adaptable to different information users is not simply a matter of technology but, more fundamentally, of information design and especially taxonomy and indexing. Even before the advent of the Web, flexible, self-service information systems existed. Consider the printed alphabetical index, such as is

found at the back of a book. The same topic is indexed multiple ways, because different readers will approach each topic from a slightly different perspective. Even if users think about the same concept, they may use different words for it, so double posts or *See also...* cross-references are used in both book and periodical indexes.

For example, in a cookbook, a recipe for a Mexican-style bean salad would be indexed under the entries for salads, beans, Mexican dishes, and possibly side dishes. This allows different people—a person who wants to fix a Mexican salad, another who wants to fix a bean salad, and so on—to find this recipe.

In online media, a combination of



**HEATHER HEDDEN** is an independent taxonomy consultant, a continuing education instructor at the Simmons College Graduate School of Library and Information Science, and the author of *The Accidental Taxonomist*. She previously worked as a controlled vocabulary editor at Gale and as a taxonomy consultant with Earley & Associates. She is a member of SLA's Taxonomy Division and the SLA Boston Chapter. She can be reached at [heather@hedden.net](mailto:heather@hedden.net).

information design (based on index and thesaurus creation principles) and information technology helps people find the information they want on their own. Instead of limiting users to a linear, browseable alphabetical index, as is the case with printed products, online media can include hyperlinks that allow topics to be arranged hierarchically, so users can “drill down” from broader to more specific topics and quickly jump to related terms. Additionally, relational databases allow users to select their choice of simultaneous multiple criteria or “facets” of a query to narrow in on the most precise information sources. Whether it’s called a taxonomy, a thesaurus, or something else, a knowledge organization system can, if well designed, help a variety of users find the information they want.

### Taxonomy Types and Definitions

Controlled vocabularies, thesauri, taxonomies and the like may collectively be called knowledge organization systems, but this designation is rarely used outside of library and information science courses and scholarly writings. While there are defined types of knowledge organization systems (as explained in the following paragraphs), there is no agreed-upon term to cover them all.

**Controlled vocabulary.** A controlled vocabulary may cover any kind of knowledge organization system, with the possible exclusion of highly structured ontologies. At a minimum, a controlled vocabulary is a restricted list of words or terms for some specialized purpose—usually indexing, labeling or categorizing. The list of terms is “controlled” because only terms from the list may be used for the subject area covered. It is also controlled because, if used by more than one person, there is control over who adds terms to the list and when and how they may do it.

Most controlled vocabularies have a *See also...* or *Use...* type of cross-reference system that directs the user from one or more “non-preferred” terms to the preferred term. Controlled vocabularies may also have other relation-

ships among their terms. A controlled vocabulary may be called an authority file, especially if authoritative sources are provided for each term and/or the terms are proper nouns.

**Taxonomy.** A taxonomy, in its narrower and original meaning, is a hierarchical classification system. The word comes from the Greek *taxis*, meaning arrangement or order, and *nomos*, meaning law or science. A taxonomy is thus a kind of controlled vocabulary where each term has a designated broader term (unless it is the top-level term) and one or more narrower terms (unless it is a bottom-level term), and all the terms are connected in one large, hierarchical structure.

What constitutes a taxonomy could be either a single hierarchy or a limited set of hierarchies. This structure is often referred to as a tree, with a trunk, some main branches, and several smaller branches off the main branches. Another way to describe a taxonomy is by its characteristic “nested categories.” The expression “drill down” is often used to refer to how a user navigates through the trunk and into the branches. Hierarchical taxonomies may or may not make use of non-preferred terms.

**Thesaurus.** A thesaurus is a kind of dictionary that contains synonyms or alternate expressions for each entry. A thesaurus for information management and retrieval shares this characteristic of listing similar terms for each controlled vocabulary entry. The difference is that in a dictionary type of thesaurus, all the associated terms could potentially be used in place of the entry, depending on the specific context; in an information retrieval thesaurus, on the other hand, the associated terms are designed for use in all contexts within the domain of content covered. The synonyms or near-synonyms must, therefore, be suitably equivalent under all circumstances.

An information retrieval thesaurus must clearly specify which terms can be used as synonyms, which are more specific (narrower terms), which are broader, and which are merely related.

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. Several national and international standards provide guidance on creating such thesauri, including ANSI/NISO Z39.19, ISO 2788 and ISO 2788 (to be replaced in 2011 by ISO 25964), and BS 8723.

A thesaurus could be thought of as having the features of a taxonomy with the addition of associative relationships, thus allowing for a greater degree of structural complexity. This is basically true, except for the extent of the hierarchy. While all terms must belong to a limited number of hierarchies within a hierarchical taxonomy, this is not a strict requirement for a thesaurus. In addition, non-hierarchical related-term relationships may also exist in strictly hierarchical taxonomies, so the distinctions between thesauri and taxonomies can be blurred.

Although most thesaurus terms will have a broader and/or a narrower term, such relationships are not necessarily required for every term. In a thesaurus, the focus is often more on the individual terms than on the top-down structure. A thesaurus might include multiple small hierarchies comprising just two or three terms, but have no overarching “tree” structure typical of a hierarchical taxonomy. The greater detail and information contained within a thesaurus help the user find the most appropriate term more easily.

To best serve the information user, an information professional needs to select and develop the right type of taxonomy for a given content/information resource. For example, where proper nouns are significant, a simple controlled vocabulary without hierarchical or associative relationships between terms would suffice. A hierarchical taxonomy is appropriate in situations where the hierarchical classification of terms (such as in product categories) is natural for the content and intuitive for the intended users. A thesaurus is especially useful for a relatively large controlled vocabulary that involves human indexing and/

## Taxonomies help bring users and content together, and the many structures and displays of taxonomies help serve different users' needs.

or supports a list of terms that the end user can browse.

### Varied Terms for Varied Users

The use of non-preferred terms is what makes a taxonomy especially valuable in helping information users approach content through different routes. As in a book index, the same concept can have multiple, essentially synonymous variants that reflect the diverse ways in which different information users might think of or describe a concept. The taxonomist designates the concepts to be included in the taxonomy and then identifies, for each concept, terms that mean essentially the same thing. The taxonomist chooses one term to be displayed as the preferred term and designates the others as non-preferred terms. (In cases where the taxonomy is never actually displayed to the user but serves to match searches, it is not necessary to designate a preferred term.)

Non-preferred terms help make content accessible to different groups of people who think of and/or describe the same concept differently. Examples of such groups are as follows:

- Subject area experts and laypersons;
- People with library training and those without;
- People who use different languages or dialects; and
- People with various cultural, religious, political, or socio-economic perspectives and biases.

Non-preferred terms can be thought of as synonyms, but they include more than just synonyms. The relationship between a non-preferred term and a preferred term is one of equivalence, but this does not necessarily mean that the terms have to be "equal" or synonymous. First, it is the *concepts*, not the terms themselves, that are equivalent.

Second, the two terms merely need to be sufficiently similar with respect to the nature of the content being indexed—that is, too much redundancy, ambiguity and confusion would result from trying to maintain distinct terms and thus, for the purpose of indexing the content, they should be treated as the same.

Non-preferred terms can be near-synonyms, acronyms, spelling variations, foreign language terms, phrase inversions, or even antonyms or broader terms. Examples of synonyms are *doctor* and *physician*, *car* and *automobile*, and *film*, *motion picture*, and *movie*. Near-synonyms include *junior high school* and *middle school*, and *foreign policy* and *foreign relations*. The decision to designate a word as a non-preferred term rather than an additional preferred term will depend on the nature of the content and the people who will use that content.

Spelling variations can be non-preferred terms if they are accepted dictionary variants or British/U.S. variations, such as *defense* and *defence*. Foreign language terms can be non-preferred terms in an English-only taxonomy if they are sometimes used in English, such as *Islamic law* and *Sharia*, or if they are proper nouns, such as *French Academy of Sciences* and *Académie des sciences*. Phrase inversions that involve putting a noun before an adjective, such as *photography*, *digital*, are acceptable and helpful non-preferred terms when an alphabetical sorting of the taxonomy or thesaurus is provided.

Antonyms can be non-preferred terms in cases where both terms describe essentially the same issue, such as *literacy* and *illiteracy*. Finally, a narrower term (such as *wind power*) can serve as a non-preferred term for a broader term (*alternative energy*) because the latter can always be used for the former.

Narrower non-preferred terms should only be used in the absence of supplemental keyword searching, which can make use of concepts that are not in the taxonomy.

### Hierarchies and Facets

The way in which a taxonomy is structured and organized can also help it serve different users. Hierarchical taxonomies have become common because hypertext supports them well, but hierarchies are not suited to all kinds of taxonomies or all kinds of users. A hierarchy assumes a common understanding of how concepts are organized and relate to each other, but not everyone shares the same perspective. For example, some people classify industries according to their SIC or NAICS codes (natural resource producers, manufacturers, transportation services, wholesalers, retailers, and services), whereas other people look at industries as vertical fields or professions (such as health care, technology, energy, consumer goods, etc.).

One way that a hierarchical taxonomy can serve different information users with different ideas about a hierarchy is by including polyhierarchies. In a polyhierarchy, the same term has more than one broader term, meaning that it appears in more than one hierarchy and thus can be found under more than one broad category. For example, *whales* could be listed under both *mammals* and *ocean life*, while *Egypt* could be found under both *Africa* and the *Middle East*.

When creating a polyhierarchy, it is important to keep in mind that a term's meaning and usage must be identical in both locations in the hierarchy. The objective is to provide two paths to the same result. Thus, including *paint brushes* under both *home improvement supplies* and *artists' supplies* is incorrect, as the different locations imply different kinds of paint brushes linked to different content.

Although polyhierarchies make hierarchical taxonomies more useful to different audiences, a hierarchical tax-

onomy should not be so full of polyhierarchies that its overall structure is lost. A dominating hierarchical structure is still needed so that information users understand the organizing principle and can navigate more quickly to the desired term instead of getting lost among numerous hierarchical relationship paths.

A taxonomy organized into facets provides another method for people with different perspectives to find the information they want, especially about more complex and detailed topics. For example, if a user is interested in identifying the marketing strategies of the leading laptop manufacturers in Europe in the late 1990s, a facet could exist for each of the following:

- Business activity (with the selected term marketing);
- Industry (computer industry);
- Product type (laptops);
- Geographic region (Europe); and
- Time period (1990s).

Not only do facets support more complex queries, but users can choose any facet in any order. Some users might think along the lines of business activity > industry > place > time, whereas others might think of place > industry > time > business activity. Result sets become further limited at each stage, so facets also allow information users to limit or broaden their search depending on how many or few results they want.

Hierarchies and facets are not mutually exclusive. The controlled vocabulary within a given facet may be arranged in a hierarchy, and hierarchical taxonomies may be structured so that their top terms reflect facets (although it's up to the user whether true faceted searching is actually performed).

### Investing Time and Effort

Taxonomies help bring users and content together, and the many structures and displays of taxonomies (alphabetical, hierarchical, faceted, and various

combinations thereof) help serve different users' needs. Non-preferred terms, polyhierarchies, and non-hierarchical associative relationships between terms are the key features of taxonomies that allow them to serve the different information needs of varied users.

Implementing a taxonomy, especially as part of an enterprise content, document, or knowledge management strategy, requires time and effort. A well-designed taxonomy should be intuitively easy to use and should not require users to have any special training or review any documentation. (It is always a good idea, however, to provide documentation for the minority of users who will read it.)

The fact that good taxonomies are easy to use and, thus, appear simple to users does not mean they are effortless to create. A considerable amount of thought and attention must be invested when designing a taxonomy to save users time and effort later. **SLA**

**Find the right job in your industry.**

## SLA CAREER CENTER

- Post Resumes
- Useful Resources
- Search for Jobs
- Seminars
- Career Advice

**www.sla.org/careers**