# WHY DO YOU NEED A TAXONOMY ANYWAY? AND HOW TO GET STARTED

Have you ever used a search engine, typed in a company or person's name, only to see results and a suggestion that you might be looking for something else instead? You may be relieved that you don't have to try a dozen variants to get at what you were seeking. On the other hand, you may want the version you typed and the suggestions are just an annoyance. Or, you may find that automatic cross-references only exist selectively and you can't depend on all synonyms to be included. There are times when the ability to browse through a list would be welcomed. Perhaps you prefer the option of picking a letter and seeing all the terms that begin with a particular word, as in the phone book index of last names where you can browse the list to find the person you want, whose first name you have forgotten. When the phone book shows you the variant spellings of a name like Crouse (Kraus, Krause, or Krauss) that helps, too, particularly if you have only heard the name spoken. Cross-references like these are a helpful tool when searching.

Taxonomies are really just specialized indexes of terms to guide you to the information you seek. What distinguishes a taxonomy from an index like the phone book, or the index in the back of a book has to do with **controls or limits** on permitted terms coupled with the possibility for **relationships** among the terms in the list. Expanding opportunities for searches on the Web prompt me to share some information about why the concept of taxonomy is important. Librarians have long used taxonomic lists to define how to organize information content to make it easier to retrieve.

In this article, I will define several types of taxonomies, why they are built, and when and where it would be useful for you to be savvy about their existence. I will also address a few basic ideas about how knowledge managers should approach building, maintenance, and application of taxonomy. References to Web sites and articles to expand your understanding are included.

There are many terms that express concepts similar to or related to taxonomy. I may use some of these in the article, or you may encounter other writings that seem to be talking about similar concepts. So that you can understand how I use the terminology in this article, here are some very simple definitions that I use, backed up by the unabridged Random House Dictionary of the English Language:

- Authority Control - Standardized rules governing terms permitted for use as index entries.
- Categorize - To assign index terms from a controlled list to information content.
- Classification Scheme - A preordained structure of words or symbols used to organize information content (or in context outside of information technology, any group of entities, e.g. animals, rocks)
- Controlled Vocabulary - Alphabetic list of terms defined as permissible for a specific indexing project. An example would be Library of Congress Subject Headings, a compendium of documents used by the U. S. Library of Congress to define the terms that may be used to index its collections. It has been adopted as the governing work for indexing the collections of many other libraries of general works. Specialized collections (e.g. National Library of Medicine) often develop their own more specific controlled vocabulary.
- Dictionary - Alphabetic list of words or short phrases with accompanying definitions, word derivations, and pronunciation, usually for a particular language or pair of languages.
- Glossary - Dictionary of specialized terms in a limited field with specialized definitions.

- Index - A list organized in a standardized sequential fashion, defined by its purpose and medium. Types of indexes may include: back-of-the-book, telephone directory, computerized look-up tables (e.g. b-tree, file system), card catalog, meeting roster of attendees, customer list, to name a few. A directory or pointer to specific content entries.
- Ontology - A classification structure emphasizing metaphysical and philosophical relationships. Click here for a treatise on ontology.
- Semantics - The study of meaning as it pertains to terms in a language. Builders of taxonomy, thesaurus or term lists must have semantic context to determine suitability of terms for the intended audience.
- Taxonomy - The science dealing with classification and naming conventions for organisms, which has been extended to non-biological naming systems. The term is currently used to refer to a particular list (e.g. A *topic* taxonomy) used as a device for normalizing the vocabulary applied to indexing a body of content.
- Thesaurus - A controlled vocabulary list, dictionary, glossary or taxonomy that is enhanced with predefined relationships among terms (e.g. Broader, Narrower). In information science, a thesaurus also includes a relationship type that governs which of a group of synonymous terms is permitted as an index term. (e.g. *auto* **use** *automobile*; *car* **use** *automobile*)
- Validation List - A list or table of values permitted in a particular database field.

Philosophers, linguists, and semantic theoreticians give much more depth to the importance and use of the terms listed, plus their many offshoots. ~~For your interest, here are two authoritative Web sites that give more information.~~
~~http://www.geocities.com/pribond/bioinfo/glossary/information.htm~~

~~http://www.jelem.com [for Jessica Milstead thesaurus expert]~~

## WHY BUILD A TAXONOMY?

In the knowledge management field, this is best answered in relationship to the **context** of some body of **content** that you want to make searchable. For example, to organize a collection of documents produced by a working group in an organization (e.g. technical call center reports, or scientific research), it would be best to classify them uniformly to facilitate retrieval. Otherwise, if content can only be retrieved by the language of the author, or by any possible synonymous term, retrieval becomes lengthier and more problematic. Building a list of terms that control the way the documents will be classified or indexed is the first step to bringing efficiency to searching. The net content of these documents will help determine the uniform terminology that is going to be selected, and simultaneously, where cross-references belong. If five authors use the term cell phone in a telecommunications company and one uses wireless phone to mean the same thing, the first term is the obvious choice based on frequency of use. It is also preferable because *wireless phone* may also refer to a cordless phone, which refers to a different type of device in the industry. This suggests the need for the taxonomist to have semantic competency in the specialty the taxonomy will cover.

Deciding what term to use, instead of another, also depends on another form of context, that of the population that will be doing the searching. In building the taxonomy, the professional must think about the context of where the indexable content originates, as well as the target searching audience. Trying to help the searcher find material more efficiently and reliably implies a need for business productivity improvement or outcome. The "why" must be defined partially by understanding the audience, searchers. If the audience is internal knowledge users, you may justify the effort based on timesaving, better quality of work due to accurately finding all relevant materials, or even eliminating replicated work by insuring that a researcher finds the answer to large or small problems without repeating work already documented by others. However, if the audience is external (e.g. customers), justification may be the ability to scale back your total support staff over time by providing an easy-to-use and reliable self-help knowledgebase for common support questions, thus decreasing the need for human intervention.

## HOW IS THE TAXONOMY GOING TO BE APPLIED?

Building a taxonomy is not useful alone. It must be used and applied consistently to be reliable for searchers. If it does not consistently return all the documents on a particular topic, searchers will quickly lose confidence in the tool and will again waste time by looking for alternative ways to insure that their search is comprehensive. Here are some fundamental questions to consider about application before you expend effort to build.

1. Will the taxonomy be visible or invisible? Do you want the searching audience to be aware that there are controls on how the content that they are searching is indexed? Many people who search the Web understand that searching a word like "cell" can return information about prisons, biology, electrochemical batteries, terrorist groups, and boxes in an electronic spreadsheet or table. This is because most search engines index every word in all the content covered by their target domain. An experienced searcher understands that narrowing a search to a particular type of *cell* will eliminate huge amounts of irrelevant citations from the search results. If there is a visible taxonomy and the source technology has reliable methods of applying it across all content, this allows the searcher to preselect the appropriate "cell" term or its approved equivalent from the taxonomy. Giving searchers a selection choice can provide an added measure of confidence in retrieval results.
2. Will the technology automatically cross-reference all synonymous, narrower, or related terms automatically, or prompt the searcher to choose from a list? For example, I may be looking for product information on printers that provide scanning and copying functions. It is important to know in advance whether, searching for *printers* will retrieve all types, including *all-in-one* models. If not, will the system prompt me for other product types to select?
3. Will lists of topics be available for browsing and easy selection of terms for searching, or does the search system have a term or symbolic expandable tree structure, allowing the searcher to dig more deeply into the nomenclature? Northern Light and a number of newer systems use file folders to symbolize broad groupings. Once the searcher chooses a particular folder it expands to further levels of granularity. Searching in any of the narrower groups of topics provides tacit Boolean operations. If you search within a folder labeled "Library Automation" for a keyword "classification," you are searching for all content on *library automation* AND *classification*, in the same document, implying a relationship between the terms. On the other hand, a browsable dictionary list of terms may leave you to choose between searching one term or the other. You would then need to sift through the results of either search term to see if the second term is also present. If browsable lists permit the selection of multiple terms with the provision for adding a Boolean operator you can achieve a similar result to refining your initial search with sub-selections. Knowing that Northern Light had a high quality taxonomic structure applied to search results, which appropriately categorized the results, is important to the skilled searcher. Novice searchers

may be using a taxonomy list without knowing it. When they go to Yahoo, the home page gives a list of topics under Web Site Directory; these topics define the top tier of the taxonomic classification tree. If you click "Health," you will see a new list of about 50 topics each with a count showing the number of citations for that topic. If you choose "Education" with just 60 citations, the content list immediately appears also with six sub-categories for narrowing. If you choose "All Diseases and Conditions" with over 10,000 citations, an alphabet appears to browse more subcategories; in addition to the alphabet about 50 general subcategories is presented. To by-pass subcategories, you can click again to force the selection of the entire topic of over 10,000 disease and condition resources. This is an example of a hierarchical browsable taxonomy that is revealed in stages for the benefit of the searching audience. Through the category tree you will find information about high *blood pressure* under *hypertension*. However, to be automatically directed to this term, you can first do a keyword search for "blood pressure," which leads you to the same content as that classified under *hypertension*, plus other material on *blood pressure* measuring devices.

4. Finally, in your planned system, will a searcher reach results that are automatically hyperlinked to citations and document counts or only to full text? In the previous example, having a Browsable taxonomy that also displayed the number of documents in each folder was very helpful. The searcher might be willing to browse through fifty or a hundred citations but decide to narrow the search if there would be a thousand citations to preview. In a Web environment, search content results are often hyperlinked only to full-text content. In your local application, the relationship between the taxonomy and the content that uses it must be clearly defined.

## METHODS OF BUILDING: HUMAN VS. AUTOMATED

There are numerous software application products and commercial search engines that index a body of electronic content and, through proprietary technology, parse the content into taxonomic entries, and then give terms some form of relationship structure. Yahoo and Alta Vista are two that attempt to cover the Web for the entire corpus of the Worldwide Web content. Other products are intended for a much smaller domain, such as a corporation or academic institution. I have already suggested some reasons for building a taxonomy, advising you to be aware of both context of the content and searching audience. Context is also important when thinking about the technology that will be used to build it and whether substantive human intervention is appropriate and supported.

Top commercial search engines use a mixture of automatic processing to determine the index terms for the content, while also employing numerous professional thesaurus developers to refine and normalize the language to a consistent and reliable standard. This was a hybrid solution. Even if you employ a software application designed for your local body of internal content, you will need the option of being able to change and refine terminology, to add new terms not appearing explicitly in the content, and to remove inappropriate language for your industry. You will also want to define the rules for when a term is used to index material. I once reviewed the automatic indexing of thousands of documents on *project management*. Needless to say, the term "project management" was applied to every document. In the context of the organization and the searchers, this was totally inappropriate.

Scope of content must be analyzed for quantity and diversity. Hundreds, or even a few thousand, documents would benefit little from a totally automated process. With a small amount of content, human intervention is necessary to define subtleties and nuances, which will enable them to organize the material for very precise retrieval. If a small and highly specialized collection has high value, because of the unique knowledge it represents, it deserves skilled human interpretation by people with expertise in the environment of creators and potential searchers. It may be that no automatic processing will be needed until the volume becomes excessive for human indexers. At that point, the foundation taxonomy should already be defined for further expansion and enhancement.

Availability of human and technological resources must always be considered. Cost is important. Going back to reasons for building a taxonomy, you must reexamine your business justifications. Business reasons can justify both the human and technology resources you need to do a credible job. Building taxonomies is clearly a case where there is little advantage for attempting the project without an expert to execute it, or if high quality searching is absent.

## WHAT CATEGORIES OF TERMS ARE CANDIDATES FOR A TAXONOMY?

Most of us think of taxonomy in terms of topical categories but there are many other ways that people in organizations seek content. For example, I may remember talking to a prospect in New Jersey last year. If the Date of all contact calls and the State are both indexed, my search is easy. At most I will need to scan a few hundred entries. While some of the ideas on this list are candidates for simple index entries (not taxonomy controlled), others clearly need vocabulary controls; many also require cross-references. When considering the following list, think about your audience, how they look for things and the types of questions they ask.

- Topics – Subject content is the most common type of taxonomy list because most searches for content are by subject. Topical taxonomy must take into account synonyms, language usage by the target audience, and variants in spelling and form (abbreviations).
- Project or Working Group Names – In a knowledge creating organization, much of the content is produced in the context of a specific group or program. Having a simple way of finding the entire body of information from such an entity is important in most businesses.
- Product Names – Companies that produce products need control over their product names for legal reasons. However, acronyms and code names have a way of becoming pervasive. Although the company name was Digital Equipment Corporation, it built the DECMate word processing system, which increased the popular usage of DEC as both a company name and the name by which customers referred to a whole family of computers, as in "I have a DEC." Neither was correct but in an e-commerce application, it would be important to present a list of possible synonyms or alternatives to the searcher who types "DEC."
- People – When human names are part of the topical index to content, they should be included in the subject list with appropriate cross-references.
- Document or Report Numbers – Standardizing report or document numbering schemes is best addressed when the schema is established. This does not overcome the problem of incorrectly applied standards. It is very common to see even highly regimented government agency publications with nonstandard variants of report number series (e.g. slashes substituted for dashes). It is unlikely that you will develop a taxonomy of report number series but establishing rules for indexing them is important. It may be best to strip all punctuation for index entries and instruct the searcher to do the same when searching, as one is advised to omit commas and prefixes in patent number searches. This is called *normalizing* the data.

- Dates – Always establish rules for date entries and assure that date searching instructions are explicit. (E.g. Will an explicit publication date of *10/2003* be retrieved in a search for *2003*?)
- Authors – It is unusual to have a taxonomy of authors with variant spellings and versions because the amount of material by any one employee is usually small enough that there will be little confusion. However, a Web site that has a lengthy employee or contact name directory as a feature might want to consider some basic finding aids such the telephone directory employs.
- Organization Names – Organization name lists need relationships that account for name changes, mergers and acquisitions, and a format that includes divisions or subsidiaries.
- Geographical Locations – Like organization names, a geographical list requires special cross-references for name changes and the possibility for labels that include the entire hierarchy for a political entity (e.g. Cambridge, Massachusetts, USA)

## WHERE DO YOU START?

Starting is the difficult part but experts recommend beginning small to get a feel for the methodology and effort needed. If you are developing a taxonomy for an enterprise, you may want to begin working with one group, project or product line, which may eventually evolve to a more substantive effort organization-wide. I find it useful to look at the organization chart, job titles and job descriptions of content producers to get a feel for the scope of content that the vocabulary must cover.

Next, look at the departments in your organization, both from the perspective of what they might produce in the way of content and what they might be researching. Gather lists of all products and product components your organization researches, designs, builds, manufactures or sells. If you are part of a service organization, do the same for the services you offer.

To understand the terminology that you have begun to accumulate, seek understanding of the context, and get a sense of the major topics around which you might group the important language in the taxonomy. Finally, seek an understanding of how the language used in one part of the organization might be related to other groups. For example, in a drug company scientists doing pure research will be steeped in compound names, but in production, manufacturing and sales compounds will have become product names.

Having done this internal research, you now have a feel for the subject matter and disciplines represented by content producers and content users. One of my favorite reference tools is Gale's Encyclopedia of Associations. Use it to find professional associations that publish glossaries, thesauri, and maintain Web sites. These will be valuable resources for verifying and adding to the taxonomy you plan to build. Collect term list publications from associations that match your content audiences' interests. You may want to explore the possibility of acquiring through licensing electronic versions as a starting point for your own taxonomy, if the list is relatively small. If it is more than a couple thousand terms, you will spend entirely too much time stripping terms to make it worthwhile to use a published list. It is far easier to add terms to a taxonomy when you begin to see the need for narrower or more specific concepts.

Finally, review the bibliography at the end of this article for other writing that will point you to products you might want to use for building a taxonomy, categorization, and search. Milstead's (JELEM) Web site, cited earlier, has a list of products for building thesauri to consider.

## ONGOING MAINTENANCE

You need a commitment to ongoing maintenance because two types of change are inevitable:

- Language changes
- Organizational interests and business focuses change

New terms come in and others are diminished in importance in the world of commerce and within organizations. The authoritativeness and worth of taxonomic content depends on vigilance, care and feeding. A program of ongoing edits, additions and modifications is crucial to build the trust of your audience in the reliability of search. This means that someone must be in charge - an authoritarian, if you will. It should be part of someone's regular job, a person with knowledge of the industry, good communication practices for gathering user knowledge, and attention to the details and nuances of language. A person trained in information science and indexing is often a fine candidate for this position.

Another suggestion is that the authoritarian or a person involved in Web site content management, invest time in routinely evaluating logs of searches to build and sustain a sense of what people seek, what they find or not. It is also good to have a contact link clearly sited on each search page that enables transmission of both suggestions and search frustrations, a search assistant. Both of these tools will be invaluable to the authoritarian in maintaining the taxonomy and adding cross-references to aid searchers. It is also useful to be able to communicate with users who have a point of view about what would be helpful to them.

## SUMMARY

Taxonomy building, support and maintenance is perhaps one of the most intellectually challenging tasks related to searching infrastructure. It requires a committed person or team if it is to be of any value. Basically, it comes down to a belief in advance effort to insure that when information is needed it can be found in the least amount of time, and with search results that are comprehensive and accurate. The expense is usually justified because the labor to build and maintain such a system is small compared to the number and level of people whose productivity will benefit. If you couple adding productivity to hundreds of employees, with avoiding the repeat of work already done elsewhere in the past and you have a winning reason for doing it.

- Lynda W. Moulton

## RELATED READINGS OF INTEREST

Hapgood, Fred. Skills. Sleuthing Out Data; Categorization software helps search-tool users find what they seek. CIO 05/01/2003, 3p.

McCloskey, Paul. Knowledge Management Building Blocks, Tech Briefing. FCW.COM 04/14/2003, 4p.

Milstead, Jessica. NISO Z39.19: Standard for Structure and Organization ofJELEM Information Retrieval Thesauri Jessica L. Milstead © 1998. 9p.

Rao, Madanmohan. A Decade of KM; a Report on "Real-World Best Practices"Destination KM from American Productivity and Quality Center's 8th KM Conference. Destination KM 06/11/2003, 3p.

Turocy, Pat. No More Information Overload;Companies must consider how they classify data so employees can find it fast. Information Week 12/16/2002, 2p.

Warner, Amy. A Taxonomy Primer. Lexonomy ©2002, 6p.