A Taxonomy Primer

by Amy J. Warner, Ph.D.

Taxonomies…thesauri…classification systems…synonym rings. We've heard all of these terms in the context of the Web. As Web sites expand, the task of organizing them has become increasingly problematic and complex.

All of the terms mentioned above are controlled vocabularies. That means that they are organized lists of words and phrases, or notation systems, that are used to initially tag content, and then to find it through navigation or search.

Unfortunately, a great deal of disagreement exists as to the individual definitions of each of the terms I mentioned; we spend too much of our valuable time misunderstanding each other. The terminology is in flux; hopefully, at some point, specific definitions will be standardized.

Until then, we need to focus on the possibilities of various types and levels of controlled vocabulary. Frankly, the specific terminology isn't nearly as important as knowing what each example can accomplish, and be used most effectively.

I have three purposes in this introduction to the world of controlled vocabularies:

1. Describe where vocabulary control fits into the information architecture of a Web site.
2. Describe the basic steps in controlling vocabulary and how these steps map to the terminology.
3. Provide some basic guidelines and recommendations for creating controlled vocabularies and leveraging them effectively.

Where Do Controlled Vocabularies Fit Into the Information Architecture of a Web site?

There are two basic places in information architecture where the form of labels and search terms can be controlled and then often organized in some way, usually hierarchically:

- In the navigation scheme, which should use unambiguous labels and where the primary organization is usually hierarchical.
- In the search system, where search terms are selected and organized for tagging content items and searching for them, now usually through a content management system and a search engine.
- Where Do I Get the Labels or Terms for My Navigation Scheme or Search Vocabulary?

There are several options for obtaining labels or terms; which ones you use depend both on what you want to do and what resources you have at your disposal. Basically, your options fall into three categories — buy/borrow, revise, or build from scratch.

There are literally hundreds, if not thousands, of controlled vocabularies floating around, some in electronic format and others still only in print. The vast majority of the vocabularies in existence are probably in-house lists. Unless you a) have some insider knowledge, and b) can convince their owners to share them with you, you probably won’t be able to get your hands on them.

At the end of this column you’ll find a list of useful links, including several to Web sites that collect links to vocabularies that are available in electronic format. While some vocabularies are in the public domain, many are only available as development resources if you either license or buy them.

Although it seems attractive to obtain an existing resource for fee or for free, my experience has been that many of these are more suited to libraries and classical document collections (i.e., books, journal articles, etc.). They are often fairly broad in scope and the terms are therefore not very deep or specific in extremely specialized areas. Of course, there are always exceptions; it pays to look at these to see if there is something that you can use.

In some cases, this type of resource can be better utilized if it is either combined with another, and/or revised. You might take an existing vocabulary and use only the part that interests you. For example, say that you are a PC manufacturer and want to develop a controlled vocabulary for your Web site. You might choose an existing computer science oriented vocabulary, and take only the part that contains the terms relevant to PC engineering. Since these will be more general than you want, you could then add to this base to create a tailored vocabulary for your site. Be sure to consider any copyright issues that might apply to borrowing terms from other sources.

The last option, build your own, is the one I encounter most frequently. Frankly, after you go through all the trouble of extracting just the terms you want from an existing vocabulary and then adding to that structure, it is probably just as easy to start from scratch. In this case, you use all resources at your disposal to create a stock of unstructured terms that you then organize as you wish.

There are lots of sources for word stock, including internal vocabularies and lists of terms, labels and terms from other Web sites related in subject matter to yours, and the previously mentioned electronic controlled vocabularies. The copyright caveat applies here as well.

Sometimes, the only option if you are in a new field and there are no other resources available, is to use your existing content items from which to choose your terms. This is the most labor-intensive option, and in my view, often the last resort.

I Have My Terms, Now What Do I Do?

It depends on your goal. You are either developing a navigation scheme with labels or a search vocabulary, or sometimes both. The key is to decide how much “control” you want, and then to take the steps to achieve that. There are basically three levels of control. From simplest to most complex, they are:

1. Control of synonyms or terms considered equivalent.

2. Arrangement of terms into one or more hierarchies, proceeding from general to specific.
3. Determining other associative or related term relationships among terms or labels.

In developing a navigation scheme, the first option, synonym control, means that you will choose the best, most consistently clear and unambiguous labels available for the content to which the user will navigate. There are many guidelines you can use, but basically you should use labels that say exactly what you want in as few words as possible. The labels that you determine will then need to be applied consistently across pages in the site.

If you are developing a set of equivalent terms for searching, you are often creating what is also called synonym rings. This involves going through your word stock and deciding what terms should be considered interchangeable when searching. For example, the terms “zucchini” and “courgettes” can be used interchangeably in a database for searching recipes. If you have the software capability, you can store these two terms as a unit. A search using either term will then retrieve all documents tagged as designated.

Synonym rings are often used to create some degree of control over a set of content items tagged using their “natural language.” By natural language, also called “keyword searching,” we refer to the situation where you use the uncontrolled language of documents for indexing, and then search using your own unrestricted vocabulary. Control is achieved by deciding which terms in the text of the content should be made equivalent and achieving this through the search interface, rather than by tagging content with a search vocabulary.

Synonym control is also the first step in creating a controlled vocabulary for tagging content items. In this case, the entry in the vocabulary would be:

Courgettes Use Zucchini

This is the first step in building a thesaurus, the most complete and complex of the vocabularies discussed here. However, the important point is that you can stop here if your goal is just to achieve some control over synonymous terms. Controlling synonyms can be very effective in minimizing the variants introduced by natural language.

The next level of organization of your terms is to arrange them in some way. Usually, a hierarchy is used. This level of organization is generally what people are referring to when they talk use the term taxonomy.

Arranging labels in a navigation scheme results in a navigation scheme that is hierarchical, although other ways of navigating across hierarchies can and should be introduced.

When you are building a search thesaurus, you need to create a list of search terms that are now arranged in what is called a BT-NT, or “broader term/narrower term” relationship. The advantage here is that when users can see this arrangement, they can select terms that are as specific as they want.

For example, to continue with the example above:

```
VEGETABLES
     SQUASHES
          ZUCCHINI
```

This path could be followed either as the navigation hierarchy from top to bottom, or stored as a “tree” that users can view and choose from while searching.

Hierarchies are probably the most ubiquitous organization structures on the Web, and users find them very intuitive. They also serve the important retrieval function of showing the relationships among content items according to their specificity or generality.

What you are really doing here is classifying your terms. Of course, this will probably not be as rigorous as a true classification scheme or taxonomy; it doesn’t need to be. But arranging terms according to specificity from top to bottom is quite useful in many contexts

Finally, it is also possible to relate terms across hierarchies. This is called the associative or related term relationship; it represents the most complex level of control. For example, VEGETABLES can be considered the top term in one hierarchy, while another hierarchy in our database of recipes might be MAIN COURSES. In this hierarchy, you might have the following:

```
MAIN COURSES
     VEGETARIAN DISHES
             ZUCCHINI PARMESAN
```

In a navigation scheme, it is often a good idea to relate these kinds of labels by placing links on the pages to guide the user to the related content.

In a thesaurus, this information is stored as another explicit relationship, the RT, or “related term” relationship. When a searcher inputs a particular term, the display would show all the terms related hierarchically and associatively to that term, as in the following:

```
       Search Term: Zucchini parmesan
```

The content displayed would be that tagged with the search term. The display would also show the UF, or “used for” term, and the BT (broader term) for the user’s consideration in modifying the search:

```
ZUCCHINI PARMESAN
     UF Parmesan of courgettes
     BT VEGETARIAN DISHES
     RT ZUCCHINI
```

Summary & General Advice

One of my main goals in this discussion has been to take the emphasis off the actual terms used and to discuss as clearly as possible the options for various levels of control. Basically, the level of control you choose depends both on what you want to do and what resources you have.

In general, I give the following advice:

- If you are developing a navigation scheme, you will probably want to make it hierarchical if your content lends itself to this arrangement. Most people will call this a taxonomy.
- If you are creating a search vocabulary, you will at least control synonyms, and will often want to arrange the terms hierarchically. The incorporation of related terms makes this a true thesaurus.
- Some control is advisable in a system with a collection of thousands of documents, and can also be useful for collections numbering in the hundreds. The important point here is that natural language search and retrieval does not scale well to larger collections, which is why people are now considering controlled vocabularies in increasing numbers.
- Make sure that the results you get from developing, using, and maintaining a controlled vocabulary are worth the investment. If you don't have a good cost-effective and usable system for tagging your content, it may not be worth the effort to create a full-blown thesaurus. This may also apply if your search engine can't be configured to search using the controlled vocabulary through the interface.

Thesaurus Bibliographies, Lists of Online Thesauri, and Directions to Thesaurus Management Software

Brown University. “Thesauri/Glossaries.”
Koch, Traugott. “Controlled Vocabularies, Thesauri and Classification Systems Available on the WWW.”
“Lexical and Classification Resources.”
Queensland University of Technology. “Controlled Vocabularies.”
Queensland University of Technology. “Taxonomy, Classification and Metadata Resources.”
University of Massachusetts — Amherst Libraries. “Classification, Indexing, Metadata and Thesauri.” “Web Thesaurus Compendium.”
Willpower. “Publications on Thesaurus Construction and Use.”
Willpower. “Software for Building & Editing Thesauri.”

fonte: *A Taxonomy Primer - 2002*