Notes from experience in producing a thesaurus taxonomy

William N. Watson, Physicist
<watsonw@osti.gov>
DoE Office of Scientific and Technical Information

Purpose of the thesaurus taxonomy

To help people find standard descriptors by zeroing in on those that relate to the concepts they're interested in, without having to wade through all the rest.

Procedure

Start with the relatively few top-level descriptors (those with no broader terms).

List the top-level terms with plenty of space between them and note what subject(s) or field(s) each one relates to. This process turns out to require more than one pass since, as you get to later terms in the list, you will think of new subjects that turn out to apply to some of the first terms in the list. Iterate until consistent.

Afterward make new separate lists for each subject, then see what *types* of descriptors there are (processes, characteristics and features of things, natural objects, principles, &c.) and note them on the new lists. Some types will be common to many subjects; some types pertain to very few subjects. Name the different types consistently across subjects. It's easier to become familiar with fewer descriptor types named consistently across different subjects than to learn many different descriptor types that have similar, but not identical, names under different subjects. Some types of descriptors may be numerous enough to subcategorize.

Sometimes narrower terms fit under different categories than their top-level terms do. For the ETDE thesaurus this didn't seem common, but it is a possibility to be aware of and take into account when categorizing similar thesauri. With the ETDE thesaurus categorization, narrower terms were often examined to help determine how to categorize the top-level terms, so this was easy to stay aware of.

Some things to aim at, and other things to not be concerned about

The elements of any category are either subcategories alone, or descriptors alone, but never both subcategories and descriptors. This is expected to make things clearer for the user and thus save his time.

Keep the number of choices under any category as few as practical; subcategorize as necessary. Finding one item in a set of 50 isn't very hard, but having to find one item in a set of 500 isn't much of an improvement over having to find one in a set of 5,000. Few people will bother with either of the latter.

Choose category names that make it as easy as possible for users to know what sorts of descriptors will *and won't* be under them just from the name, without their having to dig into the category and see the descriptors themselves first. In keeping with this, *do not use names like "Miscellaneous" or "Other"*.

Changing a category name implies at least a potential change in a category's content. Renaming categories, even slightly, requires reviewing all of at least the top-level descriptors to be sure that everything in the thesaurus that belongs under the new category is included.

The categories are meant to filter out irrelevant descriptors, not divide the descriptors up into mutually exclusive sets. So the categories overlap as needed. Descriptors tend to belong in multiple categories.

We have many ways to categorize our concepts, none yet clearly the best. Pick any way that works well. But after that, expect additions and improvements over time, as you do with other designs.

Pacing

Categorization takes real man-months. And it can, after a while, get mind-numbing. Plan work accordingly.