Original Article

Taxonomies and controlled vocabularies best practices for metadata

Heather Hedden
is the taxonomy manager at First Wind Energy LLC. Previously, she was a taxonomy consultant with Earley & Associates and had offered taxonomy development, training and indexing services through Hedden Information Management. She teaches online workshops in taxonomy creation through the continuing education program of Simmons College Graduate School of Library and Information Science. She is the author of *The Accidental Taxonomist* (Information Today Inc., 2010).

ABSTRACT Taxonomies or controlled vocabularies are used in descriptive metadata fields to support consistent, accurate, and quick indexing and retrieval of digital asset content. Designing metadata and controlled vocabularies is an integrated process that takes into consideration which and how many metadata fields will make use of controlled vocabularies. Synonyms (non-preferred terms) and hierarchies are methods to help users find the right term within a large controlled vocabulary. Best practices for creating non-preferred terms and hierarchies are explained.
Journal of Digital Asset Management (2010) **6,** 279–284. doi:10.1057/dam.2010.29

Keywords: taxonomy; controlled vocabulary; metadata; indexing; classification

INTRODUCTION

Designing the metadata for digital assets inevitably brings up the question of controlled vocabularies, taxonomies, keywords or tags. While text documents can be automatically indexed or auto-classified based on search queries matching words within the texts, non-text digital files usually require some kind of descriptive tagging in order to be retrieved in subject searches. Uncontrolled keyword tagging tends to be inconsistent, inadequate, too general and biased, leading to inaccurate retrieval results. The solution for indexing is to implement controlled vocabularies in descriptive metadata fields.

TYPES OF CONTROLLED VOCABULARIES

A controlled vocabulary is a restricted list of words or terms typically used for descriptive cataloging, tagging or indexing. It is controlled because users (catalogers, taggers, indexers) may only apply terms from the list for its scoped area (its metadata value or field). It is also controlled, because only under certain specific conditions and review processes may the terms within a controlled vocabulary change or grow, and this is the responsibility of a controlled vocabulary editor or taxonomist, not the users. The term 'controlled vocabulary' is broad and covers the full range of different kinds of structures for term management. The following are defined types of controlled vocabularies.

Term list

The simplest kind of controlled vocabulary is a flat term list, sometimes called a 'pick list'.

Term lists are often utilized for administrative and structural metadata elements, such as a list of possible file formats, rights status or retention status. Term lists are also used in descriptive metadata elements, such as content type, language, department/source and so on. Controlled vocabularies of subject terms, however, may be too large and complex for simple term lists. Term lists are often displayed within drop-down boxes for a field, but could display as button or check-box items.

Correspondence:
Heather Hedden
First Wind, 179 Lincoln Street, Suite 500, Boston, MA 02111, USA
E-mail: heather@hedden.net

Authority file

An authority file is a controlled vocabulary which includes synonyms or variants for each term which function as cross-references to guide the user from an 'non-preferred term' variant to the equivalent 'preferred term'. In addition, authority files may provide a note for each term as to the authoritative source for the preferred term as it is worded. The designation 'authority file' is used more often with named entities (proper nouns) only, and often authority files are simply called 'controlled vocabularies'.

Taxonomy

The word 'taxonomy' means the science of classifying things, and traditionally the classification of plants and animals, as in the Linnaean classification system. It has become a popular term now for any hierarchical classification or categorization system. Thus, a taxonomy is a controlled vocabulary in which all the terms belong to a single hierarchical structure and have parent/child or broader/narrower relationships to other terms. The structure is sometimes referred to as a 'tree'. The addition of non-preferred terms/synonyms may or may not be part of a taxonomy.

Recently the term taxonomy has also become popular as the term for *any* kind of controlled vocabulary, whether a term list, authority file, thesaurus, or some hybrid combination. This is especially the case in the corporate world, where one might speak of 'enterprise taxonomies'. It's simpler to have a one-word term for the concept of controlled vocabularies, especially when speaking of the people involved, such as 'taxonomists' instead of 'controlled vocabulary creators/editors'.

Thesaurus

The classic meaning of a thesaurus is a kind of dictionary, such as Roget's thesaurus, which contains synonyms or alternate expressions for each term and possibly even antonyms. An information/content retrieval thesaurus shares this characteristic of listing similar terms at each controlled vocabulary term entry. The difference is that in a dictionary-thesaurus all the associated terms *might* be used in place of the term entry depending upon the specific context, which the user needs to consider in each case. The content retrieval thesaurus, on the other hand, is designed for *all* contexts, regardless of a specific term usage or document. The synonyms or near-synonyms must therefore be suitably equivalent in *all* circumstances.

A content retrieval thesaurus is also more structured than either a dictionary thesaurus or other types of controlled vocabularies, because it provides information about each term and its relationships to other terms within the same thesaurus. In addition to specifying which terms can be used as synonyms (labeled as 'used from'), a thesaurus also indicates which terms are more specific (narrower terms), which are broader, and which are non-hierarchically related terms. In addition, some terms have scope note explanations, as needed. With this much information, the user can typically choose among multiple display options for a thesaurus.

METADATA FIELDS AND CONTROLLED VOCABULARIES

Designing the metadata elements for a digital asset collection and designing the controlled vocabulary are integrated processes. Each controlled vocabulary, hierarchical taxonomy or authority file will correspond to a different metadata field. The initial decisions in developing metadata and controlled vocabularies center on the following questions:

- Which metadata fields should have controlled vocabularies?
- How many and what metadata fields should there be?
- How many controlled vocabularies should there be?

Not every metadata field needs to have a controlled vocabulary, so it must be decided which will have controlled vocabularies and which will not. Fields such as title or filename should allow free text, and numeric fields such as size and date, also do not use controlled vocabularies, even though there may be policies pertaining to the entry format. The field for creator, for example, may or may not have a controlled vocabulary, depending on the circumstances. If all content creators are restricted to employees of an organization, then a controlled vocabulary is easy to implement and would support more efficient, accurate

indexing. If, however, creators, could be any possible outside person, then the names should not be limited to a controlled vocabulary.

While the decision regarding how many metadata fields to create is usually made independent of controlled vocabulary development, the way in which controlled vocabularies are managed could impact the number of metadata fields. Returning to the example of creator names, if names are sometimes internal (can be in a controlled vocabulary) and sometimes external (cannot be controlled), the there are two options: (1) create two separate metadata fields, one for internal creators (with a controlled vocabulary) and one for external creators (free text), or (2) create a single metadata field that uses a controlled vocabulary but also allows the option of 'overriding' the controlled vocabulary and entering an unapproved free-text name. The latter is ultimately simpler to use, yet more technically challenging to implement.

Determining the number of standard metadata fields depends on administrative needs and policies. Making distinctions among certain kinds of descriptive metadata, however, may not be obvious and thus can be more challenging. For example, person names and corporate names could be used in controlled vocabularies for a creator or publisher metadata field, but a digital asset may also be 'about' a person or corporate body. Then the question arises: shall names (proper nouns) be part of a single subject taxonomy for the subject metadata field, or shall there be more than one type of subject field? A single descriptive controlled vocabulary would thus include topical subjects (generic/common nouns), person subjects (proper nouns), organization subjects (proper nouns) and so on. The same question may be asked about place names for digital assets that are about a place and not merely the place of asset creation.

In making the decision, there are various factors to consider:

- How large the subject-descriptive controlled vocabulary is – if large, it may be better to break it up into separate vocabularies and metadata fields, but if small it can be kept as one.
- What the ratio of names to topical subjects is – if names are few, then they could more easily be integrated within the topical subjects, but if there are many names, their own metadata field may be justified.
- Whether advanced search permits users to select more than one term at once from within a single metadata field – if so, then a combination of term types within a metadata fields is more acceptable than otherwise.
- How users are most likely to look up names and topics.

While having a greater number of descriptive metadata fields can support more sophisticated searching, too many fields can be confusing to the untrained user. For example, the distinction between names as content creators and names as subjects, or between places of creation and places as a subject, may not be obvious to some users.

In specialized applications, additional types of topical subjects may be broken out into their own controlled vocabularies and consequently their own metadata fields. Examples include product types, industries, facility types, markets/customer types, job titles and so on. Making the determination as to whether a subject category should stand on its own as a separate controlled vocabulary and metadata field, depends on both the nature of the content and the needs of the users searching for the content. The two key questions to ask are:

1. Will users want to search and limit by this particular type of subject?
2. Can the majority of digital assets in the collection be described by this type of subject?

The number and what kind of controlled vocabularies to create should be tailored to the particular digital asset collection and users.

USE OF SYNONYMS (NON-PREFERRED TERMS)

As explained previously in the definitions, controlled vocabularies may or may not contain 'synonyms', which are more correctly called non-preferred terms, since many are not true synonyms. These serve as additional entry points or cross-references to corresponding preferred terms within the controlled vocabulary. They are also known as variant terms, use references,

see references, entry terms, variants and equivalencies.

In addition to synonyms, non-preferred terms may be near-synonyms, alternate spellings, grammatical/lexical variants, slang or technical versions, phrase inversions, acronyms and so on. Since terms in a controlled vocabulary are usually not single words but often phrases of two or three words, there can be many possible non-preferred terms for each term. What is important to keep in mind when creating non-preferred terms is that they should be sufficiently equivalent to the preferred terms in the context of the content repository to serve for retrieving the same content as the preferred terms.

When a controlled vocabulary has non-preferred terms, usually all topical terms have at least one non-preferred term, and many have more than one. Proper noun terms do not necessarily need non-preferred terms to the same extent, so some proper nouns many not have any. While a preferred term may have multiple non-preferred terms, each non-preferred term should point to only one preferred term.

Non-preferred terms serve two kinds of users, those who are doing the tagging or indexing of content and those who are searching for content. They serve the taggers/indexers by helping them quickly find the ideal term to index similar content efficiently and consistently. Speed is always an issue, and consistency is particularly an issue if there are multiple people performing indexing. Non-preferred terms serve varied users who will look for the same concept by different term names. It is possible to have non-preferred terms just for the indexers, or just for the end-users (as part of the search system), but usually non-preferred terms are for both.

The decision to add non-preferred terms depends on the size of the controlled vocabulary. If all the terms in a controlled vocabulary can be seen and skimmed through in a single screen view, such as via a drop-down scrollbox, then non-preferred terms are generally not needed. Thus, controlled vocabularies of 20–30 terms or less probably won't have non-preferred terms. Longer lists may also not need non-preferred terms if it is obvious to the user by scrolling whether a term is present or not. For example, a browsable alphabetical list names of employees or of countries, states, or provinces does not need non-preferred terms. The ability to search for terms by both beginning words and by words within the term can also preclude the need for non-preferred terms in controlled vocabularies of proper nouns. In sum, alphabetically arranged controlled vocabularies of generic topics of over 40 terms or so (a greater number than can be viewed at once) or of proper nouns that number into the hundreds, ideally should have non-preferred terms.

When developing non-preferred terms for a controlled vocabulary that can be browsed alphabetically, the first word of the non-preferred term needs to be a word that users will likely look up. This could include phrase inversions, such as 'Bridge, pedestrian'. If a controlled vocabulary will be browsed only and not also searched, then non-preferred terms beginning the same word as the preferred term are not needed. For example, 'Administrative staff' is not needed as a non-preferred term for 'Administrative assistants'. If the controlled vocabulary can be searched, however, non-preferred terms that begin with the same word, as in the preceding example, should be included.

Non-preferred terms make a controlled vocabulary much more effective, but they also make the controlled vocabulary more complex to develop, implement and maintain. The task of creating and maintaining non-preferred terms usually requires the resource of a taxonomist, at least as a partial responsibility. Implementation also depends on the support of the search system.

USE OF HIERARCHIES

A taxonomy usually means a controlled vocabulary with a hierarchical structure. A hierarchy is an alternative method to an alphabetical list for the user to browse to the desired term within a displayed controlled vocabulary. Users, again, may be either the indexers or the end-user searchers of content. In a hierarchical arrangement, terms are subordinate to others in a parent/child or broader/narrower relationship. A hierarchy could comprise as few as two levels, but three or four levels deep is also quite common.

Although hierarchies are increasingly popular in the online medium (in contrast to the former, print medium), they are not suitable for all controlled vocabularies. Alphabetical lists are better for names and other proper nouns. Subjects that are difficult to classify – such as generic miscellaneous topics, or methods, processes, activities or events – may be more appropriate to organize in alphabetical lists only that include non-preferred terms. Other subjects lend themselves more naturally to hierarchical classification, such as industries, product types, facility types, organizational units and so on. Place names could be arranged either alphabetically or hierarchically, so the exact list of names needs to be considered when choosing the display arrangement places.

It is important to put a pair of terms into hierarchical relationships only if the concepts truly have a broader/narrower relationship and not to do so merely for convenience of grouping. The narrower term must be one of the following:

- a specific type of a generic broader term;
- a proper noun instance of a broader term; and
- a part of an integrated system whole-type broader term.

Thus, for example, narrower terms for a specific company name could include its subsidiaries, then divisions, then other operating units and finally departments, but not any term that has anything to do with the company. Employees, products, markets and locations are *not* narrower terms for a company name. More information and examples of each type of permitted hierarchical relationship can be found in national and international controlled vocabulary standards, such as BS 8723, ANSI/NISO Z.39.19-2005 and ISO 2788 (1986) (soon to be replaced in 2011 by ISO 25964).

The total number of terms is a factor considered differently for a hierarchical taxonomy than it is for an alphabetical controlled vocabulary. A controlled vocabulary too small to be worth the trouble of including non-preferred terms could still benefit from a hierarchical structure. A single displayable list of 20–50 terms, whether fixed on a page or in a drop-down box, could be made easier to browse if organized into a simple, two-level hierarchy. The second-level terms merely need to be indented to effectively represent the hierarchy, and not every top-level term needs to have narrower terms. On the other extreme, large taxonomies of over 500–1000 terms in a hierarchy may become to unwieldy with too many levels deep, and are no longer easy to browse. The largest controlled vocabularies, while possibly having some hierarchy, can be browsed or searched more effectively if relying on non-preferred terms rather than on a hierarchical structure.

Display options for a hierarchical taxonomy vary based on taxonomy size and on the supporting technology, and they may also differ between that of the indexer's view and that of the end-user's view. A small taxonomy has the space to include all narrower terms in a static list, typically indented, as described previously. This is the simplest to implement in a user interface for metadata selection. A large taxonomy would be too long to scroll through if all levels of the hierarchy are always fully displayed. Expandable topics, often displayed as folders, with plus signs to indicate the presence of narrower term levels is a popular method of indicating hierarchy for end-users, but requires greater technology resources to implement, and may not be a metadata field display option available to the indexers. A compromise display that is easier to implement than expandable folders, but less elegant or user-friendly is to use a second metadata field for subtopic, where the lists of subtopics varies and is dynamically driven by the choice of main topic. Subtopics are narrower terms for main topics. Only after the user selects a main topic, does a list of corresponding subtopics appear in the associated subtopic field. This method is feasible if a taxonomy is no more than two levels deep, and more suitable if all main topics have subtopics. Figure 1 shows an example of this method of having a second metadata field for a subtopic.

A thesaurus contains hierarchical relationships, but is not necessarily constructed as an overall hierarchy, and it contains other relationships as well. As a thesaurus is large and has non-preferred terms as additional entry points, an alphabetical display, in addition to a hierarchical display, is often useful. It is indeed more complex to create and technically support a thesaurus, so this kind of controlled vocabulary makes more sense to implement when there is

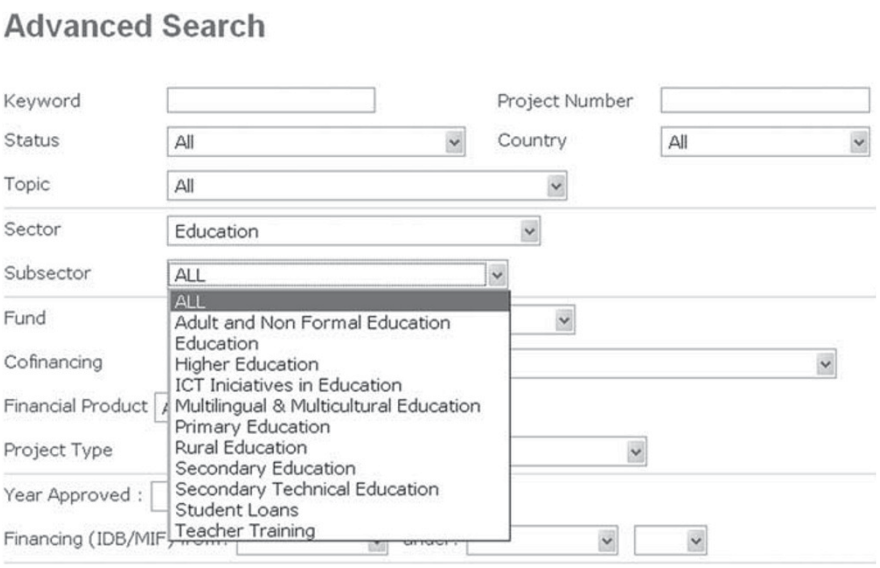

Figure 1: A hierarchical taxonomy of sectors supported by two metadata fields in advanced search for projects of the Inter-American Development Bank.
Source: http://www.iadb.org/projects/search.cfm.

more than one indexer (a person for whom indexing with the thesaurus is their primary job).

CONCLUSIONS

Taxonomies or controlled vocabularies enable consistent, accurate, and rapid indexing and retrieval of content. The fact that both indexers and end-users benefit from them, make controlled vocabularies very desirable. They are especially important for digital assets that do not have text that can be analyzed by a traditional search engine. Controlled vocabularies support various metadata fields and thus vocabulary design needs to be integrated with the metadata strategy. This may require not just one person, but rather a multidisciplinary team of experts to implement.