May/June 2008 ▪ Volume 11 ▪ Number 3

Intranets
ENTERPRISE STRATEGIES AND SOLUTIONS

Making Decisions in Taxonomy Creation

HEATHER HEDDEN

Taxonomies, or controlled vocabularies, are structured sets of terms used for indexing or categorizing content. They facilitate search and retrieval in a wide variety of applications ranging from periodical indexing, image archiving, enterprise content management, commercial product categorization, online news service interfaces, and website information architecture. Taxonomies are especially useful for categorizing content and facilitating search and navigation on intranets.

Developing a hierarchical taxonomy (or set of taxonomies) involves a great deal of decision making regarding the kind of categories, the number of categories, the number of levels, the wording of the terms, and so on.

While taxonomy owners or stakeholders can provide some of the answers, they often leave a number of decisions up to the individual developing the taxonomy (the taxonomist), who "knows best." The taxonomist therefore needs to anticipate issues that will arise and know what questions to ask in order to make the best decisions for specific applications.

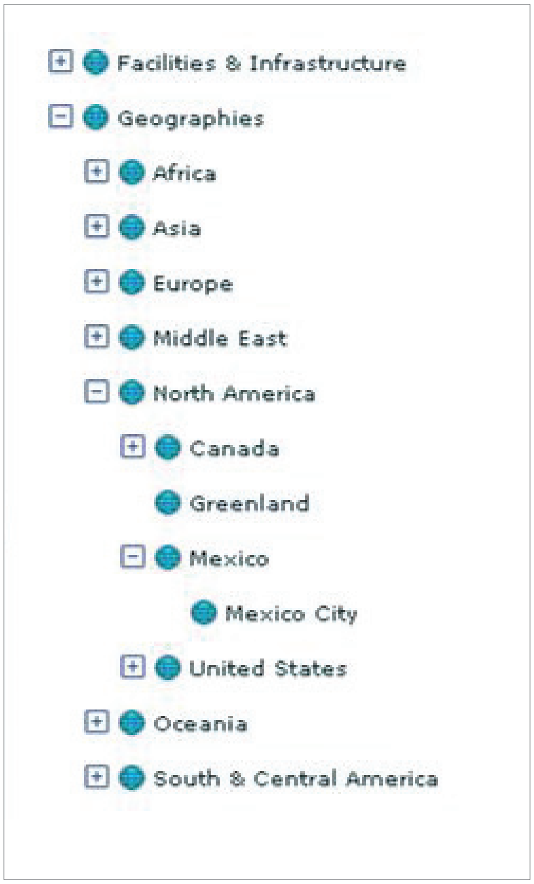
Decisions that are typically made by the taxonomy owner include the basic dominating structure of the taxonomy, the design of the interface, and the means of indexing. The structure is typically either a hierarchical tree or a set of category facets. The interface determines the layout and display of terms (also called nodes). Finally, the indexing of content can either be organized by means of an automated search/categorization system or by human indexers/taggers.

Even with these parameters determined by the taxonomy owner or manager, there are still a number of decisions that are typically left up to the taxonomist. These include, but are not limited to, the following:

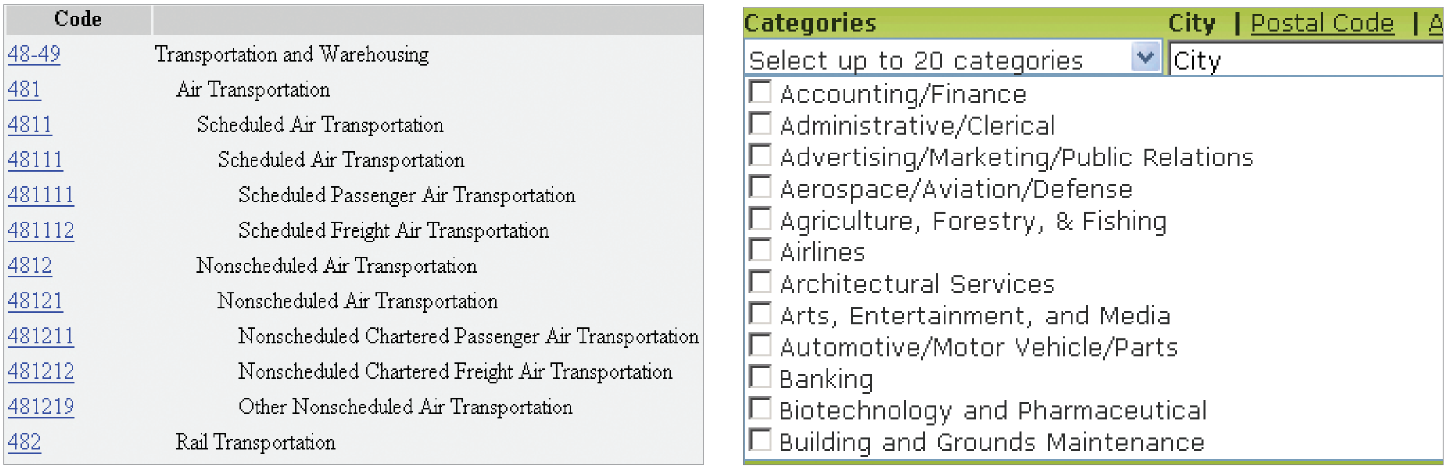
- *The final number of levels and the number of nodes per level*
- *The arrangement of the node hierarchy and the specific placement of nodes*
- *The extent of term precoordination or postcoordination*

NUMBER OF LEVELS, TERMS PER LEVEL

For an HTML-based taxonomy, a popular rule of thumb is to go only three



An example of an expandable tree hierarchy user interface for a taxonomy from the Viziant enterprise search system.



Code	
48-49	Transportation and Warehousing
481	Air Transportation
4811	Scheduled Air Transportation
48111	Scheduled Air Transportation
481111	Scheduled Passenger Air Transportation
481112	Scheduled Freight Air Transportation
4812	Nonscheduled Air Transportation
48121	Nonscheduled Air Transportation
481211	Nonscheduled Chartered Passenger Air Transportation
481212	Nonscheduled Chartered Freight Air Transportation
481219	Other Nonscheduled Air Transportation
482	Rail Transportation

Categories | City | Postal Code | A

Select up to 20 categories | City

□ Accounting/Finance
□ Administrative/Clerical
□ Advertising/Marketing/Public Relations
□ Aerospace/Aviation/Defense
□ Agriculture, Forestry, & Fishing
□ Airlines
□ Architectural Services
□ Arts, Entertainment, and Media
□ Automotive/Motor Vehicle/Parts
□ Banking
□ Biotechnology and Pharmaceutical
□ Building and Grounds Maintenance

Excerpts of two different taxonomy methods used to classify industries: The NAICS (North American Industry Classification System) taxonomy is up to five levels deep, but it does not have many nodes per level. The industries on Monster.com are in a flat list of 55 nodes. NAICS also separates out manufacturers, wholesalers, and retailers, whereas Monster.com's industries are more vertical market-oriented.

levels deep with only six to eight concepts per level. These numbers are based on user experience tests that have shown that users only have the patience to click down to a third level and the ability to scan only six to eight term entries at once. This makes sense for a small taxonomy that is integrated into an intranet's navigation menu, where the space is confined to a strip along the top or a margin along the side.

In reality, however, a taxonomy that fully covers all the content of an intranet may have hundreds or thousands of nodes. Therefore, in a hierarchical taxonomy the taxonomist needs to consider how best to balance the number of levels (depth) and the number of nodes per level (breadth).

The taxonomist must first consider the user interface display, including any vertical or horizontal space limitations. If each level will display in a separate webpage, then as many terms as can fit onto one page in two or three columns without scrolling will work nicely.

If the taxonomy will display as an expandable tree in which sublevels appear (such as by clicking on a plus sign), then more than three levels is fine, but the number of terms per level ought not extend beyond a single page length in a single column, because the position in the hierarchy cannot be seen.

The taxonomist also needs to consider the nature of the content and the users' needs and expectations. If the content is detailed and job-specific, such as information on components or manufactured parts, users will expect many levels. General areas of an intranet accessed by all employees, such as human resources, will more likely have fewer levels but a greater number of top- level nodes and nodes at each level.

Another issue to keep in mind is that the greater the number of levels, the less consistency there inevitably will be across levels. This may or may not be important. If users are experts in the subject area or are using the taxonomy frequently, such as for their own functional area, providing accurate terms is more important than providing a uniform-looking hierarchy.

On the other hand, if the userbase doesn't access large parts of the intranet frequently or includes all of the company's employees, then a logically organized taxonomy with a consistent number of levels and a consistent degree of specificity at each level is desired.

NODE HIERARCHY ARRANGEMENT

Related to the issue of deciding the depth and breadth of each hierarchical level is the decision of how to design the structure when different alternatives for classification exist. Departments within an organization, for example, can be categorized by function, by market served, or by geographical location. Industries can be organized by a standard classification system such as SIC or NAICS codes, or by vertical market sector.

A similar decision would be whether to organize products by function served, manufacturer, or customer type (consumers, business, or government). In other areas, people's names can be listed by employment status, function, or location. Finally, buildings and facilities can be grouped by geospatial location or by type of facility. Even if a taxonomy supports "polyhierarchies" (the presence of the same term under more than one broader term), there still needs to be an overarching general structure to the taxonomy so that it appears logical and is easy to use.

In determining the arrangement of the node hierarchy, the taxonomist needs to consider users' needs and expectations. Based on their background and perspective, how would the majority of users most likely classify the subject matter? For products, classification by end use makes most sense for users who are consumers, whereas classification by material type is more appropriate for wholesalers. For organizing national and international government agencies, a U.S.-centric taxonomy structure would be appropriate for U.S. federal employees but not for international users.

It is important to know whether retrieval of documents will be recursive, that is, whether a node retrieves not only the documents indexed with that node but also the documents indexed with all of its narrower nodes. Recursive retrieval might be suitable for product categories, industries, or geographic places. If recursive search and retrieval is implemented, then hierarchies should be carefully created to retrieve expected results.

PRECOORDINATION OR POSTCOORDINATION OF TERMS

Taxonomies vary to the degree in which their terms are precoordinated or postcoordinated for achieving search results. “Pre” or “post” refers to before or after search execution. An example of a precoordinated term is “discontinued product customer support,” to which documents have already been indexed, whereas a search on the term “discontinued products” and the term “customer support” in a Boolean “and” operation is an example of postcoordination of terms. Additional examples of precoordinated terms are “pre-employment drug testing,” “customer information files,” and “management training workshops.”

Precoordinated terms, if used correctly, provide more precise retrieval results, are better suited for specific custom taxonomies, and are more likely to match multiword search strings entered by users. A drawback is that the existence of such specific nodes might be overlooked by the user, and it is more complex to correctly index with a taxonomy comprising many precoordinated terms. Precoordinated terms are typically found in a hierarchical tree-type of taxonomy, in which the user browses down the levels in order to find the most specific terms.

A faceted taxonomy, on the other hand, serves postcoordination, whereby the user chooses a combination of terms from multiple facets. But the distinction is not always clear cut. Hierarchical topic trees may be used in a user interface that supports postcoordinated searching, and hierarchies may exist within facets. Thus, the taxonomist needs to determine in a hierarchical thesaurus whether to expect postcoordination on the part of the user, and in a faceted taxonomy whether (and to what extent) to include precoordinated terms.

In making the decision between precoordination and postcoordination, factors to be considered include the search interface, the scope and volume of the content indexed, and the knowledge of the end user. If the primary method of searching will be by entering terms in a search box rather than browsing a hierarchy, then more precoordinated terms are likely needed to increase the likelihood of matching a user-entered term. Content comprising a wide range of article types and subject areas is also better served by precoordinated terms, which are less ambiguous across subject disciplines. Finally, users who are subject area experts are more likely to search for precoordinated concepts. Flexibility in the degree of precoordination versus postcoordination is acceptable, but consistent application will make the taxonomy easier to use.

Each taxonomy project is unique. Creators of taxonomies need to understand who the users are, what their needs are, what the nature of the content is, what the user interface looks like, what the system supports, and how the content will be indexed. Creating a good intranet taxonomy requires the combined skills of information science and usability experience. I

HEATHER HEDDEN (heather@hedden.net) is an information taxonomist with Viziant Corp. and a continuing education instructor of taxonomy creation and web indexing at Simmons College Graduate School of Library and Information Science.