MONTAGUE INSTITUTE REVIEW

ISSN 1554-303X

November 2002

Ten taxonomy myths

Taxonomies have recently emerged from the quiet backwaters of biology, book indexing, and library science into the corporate limelight. They are supposed to be the silver bullets that will help users find the needle in the intranet haystack, reduce "friction" in electronic commerce, facilitate scientific research, and promote global collaboration. But before this can happen, practitioners need to dispel the myths and confusion, created in part by the multi-disciplinary nature of the task and the hype surrounding content management technologies.

What is a "taxonomy?"

The confusion begins with definitions. Ours is broad enough to accommodate all applications:

> *"A taxonomy is a system for naming and organizing things into groups that share similar characteristics."*

In our view, the "things" (objects) to be organized can be biological organisms, abstract concepts, products and services, geographic regions, and even people. The "groups" (categories) can be expressed as A - Z indexes, thesauri, topic hierarchies, tables of contents, advanced search forms, and other navigation tools.

> *Myth #1*: A taxonomy can only be expressed as a hierarchical list of topics.

The implication of our definition is that every company will use multiple, interacting organization schemes (taxonomies). Some will be very concrete and may even be "invisible" except to computer programs (e.g. product codes). Others will be abstract, designed primarily for use by human beings (e.g. a list of topics on a departmental Web site).

> *Myth #2*: There is only one "right" taxonomy for each organization.

Origins of business taxonomies

In the biological and library sciences, taxonomy development is a long-term, collaborative effort involving classification specialists (see International Association for Plant Taxonomy and Library of Congress). Taxonomies evolve slowly through a consensus process that involves representatives from multiple public and private sector organizations. In business, taxonomies must respond to rapid change in three areas:

1. *Business processes*. Geographic taxonomies often conform to sales territories. Product taxonomies originate in manufacturing processes.

2. *Budgeting and managing.* Budget categories reflect how the company intends to invest its resources. Organization categories reflect deployment of human and physical resources.

3. *Strategic planning*. Categories for concepts relating to future challenges and opportunities reflect the company's world view -- what business are we in, who are our current and future competitors, what technologies hold the most opportunity, who are our most profitable customers?

The implication is that business taxonomies are often parochial (designed for a single task or process) and overlapping. A taxonomy for the sales function in one company is unlikely to work in another company even in the same industry.

> *Myth #3*: You can shortcut the taxonomy development process by wholesale adoption of someone else's taxonomy.

Taxonomy structures vs. taxonomy applications

The structure (architecture) is the taxonomy as the programmer or taxonomist sees it. The application is the taxonomy as the user sees it. Because computers require data to be both predict-

able and comprehensive, a taxonomy structure often requires that each term appear in only one place in the hierarchy and that all terms be included.

These constraints are neither necessary nor desirable in a taxonomy application, where it is often necessary to accommodate the needs of multiple user groups or, at minimum, the different information-seeking behaviors of people in a single user group.

> *Myth #4*: Taxonomy applications (what the user sees) must conform to the same rules as the underlying taxonomy structure (how the data is stored in the computer).

Taxonomies in the information life cycle

Business taxonomies can be stored in several ways:

1. As fields and values in a general purpose relational database.
2. As parameters in a proprietary application program.
3. As metadata in published reports, manuals, or presentations.

In all cases, but especially in published documents, important taxonomic data can be missing or incorrect because the data entry clerk or the author was sloppy, poorly trained, or both.

Unfortunately, most corporate taxonomy development projects begin at the wrong end of the information life-cycle. Instead of tackling the problem at its source -- content creation -- the effort is invested in classifying documents in an existing repository with all its warts. The result is classified mush -- search results with no titles, erroneous publication dates, gibberish descriptions, and too many matching items.

> *Myth #5*: You can create cost-effective taxonomies by investing in the end of the information life cycle (post-publication) and ignoring the beginning (content creation).

User-oriented vs. content-oriented taxonomies

An unfortunate consequence of focusing on the wrong end of the information life cycle is an over-emphasis on content at the expense of user needs. To see the difference, consider the following two ways of organizing information about computers.

Library catalogers and indexers tend to focus on content when developing taxonomies because that's all they have to work with. Database designers tend to focus on making a single business process more efficient. Journalists tend to focus on user needs and interests. All are valid and necessary points of departure, but because journalists are under-represented on corporate taxonomy projects, the user's needs often get short shrift.

> *Myth #6*: A corporate taxonomy should be derived solely from the content in a repository.

Document-centric vs. people-centric taxonomies

What do you classify -- documents or people or both? What is the purpose of your taxonomy -- to find published material or get help? In an academic environment, the focus is on published material for research, writing, and discussion. In a business context, though, you want to solve a problem, get advice, or recruit people to help with a task. Except perhaps in the legal and documentation arenas, documents are a means to an end, not an end in themselves.

But how can taxonomies be used to help find experts? Three common approaches are:

1. *Categorize e-mail*. Use an auto-categorization program to scan e-mail messages and discussion list postings.
2. *Expertise database*. Develop an expertise database where employees enter a profile of their skills and experience.
3. *Documents as "information artifacts."* Publish and categorize key documents prepared by departmental experts. Attach relevant metadata to each document -- author contact information, content owner (departmental publisher), publication date, topics.

The third strategy is probably the most cost-effective as long as you're willing to invest in creating quality information in the first place. Not only does the document help identify the

Content-oriented classification	User-oriented classification
Hardware	User group A (Microsoft Office users)
Large, centralized systems (mainframes)	Pre-sale questions (price, compatibility, features, etc.)
Client/server systems	Installation questions
Portable digital assistants	How-to questions (e.g. can this be done, how do I do it?)
Peer-to-peer networks	Problems and errors
Software	User group B (Content managers)
Operating systems	Planning & budgeting issues
Office productivity software	Technology selection questions
Drawing and painting software	Industry-specific and function-specific issues
Security software	How-to questions
User-focused taxonomy	Problems and errors

expert, but it provides key details that help other employees evaluate his (her) suitability for a task.

> *Myth #7*: It's OK to create separate taxonomies for people and documents.

Integrating taxonomies

Business taxonomies reflect a unique environment that consists of specific content, processes, and users. Yet a single company can contain many such environments representing individual departments, business functions, and even individual "knowledge stewards." Moreover, the firm participates in processes that involve other organizations. Inevitably, methods must be found to integrate multiple taxonomies. Integration is necessary to:

• *Ensure accurate reporting*. If one department calls its supplier a "publisher" and another calls its supplier a "manufacturer," it will be hard to get a total number of suppliers for both departments.

• *Enable data exchange across applications*. The International Standard Book Number (ISBN) is the standard inventory code for the book trade. Amazon.com's system can accommodate the ISBN, but it uses the Amazon Standard Identification Number (ASIN) as its internal standard because it sells other kinds of products as well as books.

• *Facilitate retrieval and discovery*. If you want to find information about technologies that carry people from one floor to another, you would need to search for "elevators" in the U. S. and "lifts" in the U. K. If you want to find ads for a product sold in Canada, a bi-lingual country, you would need to search for both "Annonce publicitaire" and "advertisement."

> *Myth #8*: Personal and departmental taxonomies do not need to be integrated with other corporate taxonomies.

Loose vs. tight integration

Integration can be "tight" or "loose." If computer programs are the primary taxonomy users, the integration must be "tight" (technically compatible, no ambiguities). An example is the U. S. Environmental Protection Agency's Environmental Data Registry, which integrates data from multiple EPA sources. Tightly integrated taxonomies can reduce transaction and reporting costs but can be expensive to maintain as business conditions and technical platforms change.

If human beings are the primary taxonomy users, integration can be "loose," consisting of web-like structures where links point to a variety of resource types (including people). Technical compatibility (i.e. same hardware/software platform) is not necessary, and ambiguity is tolerated to promote discovery. Examples are online magazines and journals, which include links to authors, related articles, topical collections, cited sources, and often an annual A - Z index.

> *Myth #9*: Taxonomies should always be tightly integrated and computerized to achieve maximum efficiency.

Investing in taxonomies

Justifying expenditures for taxonomy projects is a common problem for our Society members and seminar participants. Typically, funds are needed to acquire expensive software and to staff positions for taxonomy maintenance. Unless the firm can show cost savings in existing indexing operation (e.g. an information provider like Reuters) or a direct connection to revenue (e.g. a "dot com" business like Bitpipe), it's a tough sell.

A better approach involves the following alternatives:

1. Use a hybrid budgeting approach that re-allocates resources to the department or division level. Invest centrally in standards, infrastructure, thesauri, and training materials. Invest locally in content creation and selection, specialized taxonomies, training, and application development.

2. Focus on improving the quality of content -- more meaningful titles, better structured documents, accurate metadata, and links to contact information.

3. Use editors and subject matter experts to select the highest quality and most relevant articles for external audiences (i.e. readers outside the department) to minimize the total number of documents available.

4. Include "informal" communication formats such as e-mail, interviews, and discussion groups in the content corpus.

5. To minimize costs associated with changing vendors, use a general purpose relational database such as Oracle, SQL Server, or Filemaker to store the taxonomy structure.

> *Myth #10*: Taxonomies should be funded and managed by a centralized IT function.

Conclusion

Myths are an outgrowth of the multi-disciplinary nature of corporate taxonomy development. A principle that works in one situation can become a myth when generalized to the system as a whole.

Cost effective taxonomy development requires the active participation of many specialties, including IT staff, corporate librarians, departmental publishers, commercial information providers, and international standards bodies. The myths represent conceptual and communication gaps that can impede effective collaboration.

The *Montague Institute Review* is published by the Montague Institute and edited by Jean Graef.