MONTAGUE INSTITUTE REVIEW

ISSN 1554-303X

June 2003

Ten Questions for Taxonomy Teams

Today, creating a corporate taxonomy is a team activity that brings together experts from multiple disciplines. It's important for team members to participate in learning activities as a group, preferably in a hands-on setting using their own data. This article discusses ten questions that taxonomy teams need to answer.

Who's on the team?

Taxonomy development teams typically involve staff from publishing, library science, and information technology. They can also include representatives from records management, training and development, electronic commerce, finance, and subject matter specialties. Team members are usually selected for one of the following reasons:

- they are experienced in organizing information (librarians, records managers);
- they need better taxonomy tools (publishers, researchers, course developers);
- they can install, customize, and develop hardware and software (IT).

Every team starts out with a vision, mandate, or objective. In our experience, the most common mandate is "create a corporate taxonomy to make it easier to find internal information." That's a good starting point, but it's too limited. For example, it doesn't address such issues as dealing with multiple organizational schemes, integrating print and Web formats, setting and enforcing standards, or improving content quality.

To put it another way, a "taxonomy" is a necessary part of the company's intellectual infrastructure.

Identify multiple perspectives

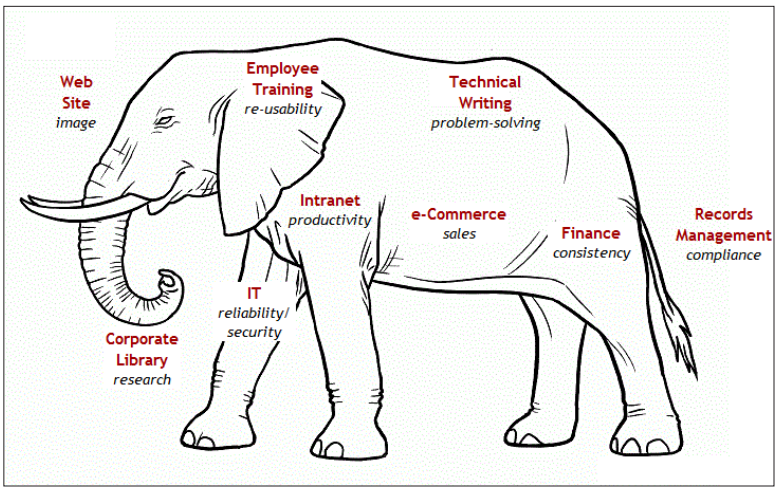
On multi-disciplinary teams, each person views the objective from a different perspective, defines terms differently, and sometimes has a different take on the agenda. Recognizing these differences at the outset sets the stage for innovative solutions. Ignoring them is a recipe for unproductive bickering and stalemate.

Each person on the team sees the taxonomy from a different perspective, defines terms differently, and sometimes has a different take on the agenda (see the drawing below). So the first question involves identifying and understanding the different perspectives — what they can contribute, where they overlap, and where the gaps are.

> 1. How do we develop a common conceptual framework?

Establish a common vocabulary

A related issue is vocabulary — starting with a definition of "taxonomy." For some, a taxonomy is a Yahoo-type directory. For others, it's a whole system of interacting metadata elements



than can be "read" by both computers and humans. Another problem word is "thesaurus," which can have three meanings:

Writer's perspective

To the writer or layman, it's an alphabetical list of words, with synonyms and related words, as in Roget's Thesaurus. Thesauri of this type illustrate a common problem in taxonomies — the same word can have multiple meanings depending on the discipline (e.g. language vs. law).

Indexer's perspective

To the indexer or editor, the thesaurus is a reference tool with specific functions and structures, such as the ACM Computing Classification System. It's used primarily by the association's editors and indexers to prepare navigation tools journal articles.

Some of our members have licensed the ACM thesaurus to create specialized organization schemes for use on an intranet.

Computer scientist's perspective

To the computer scientist, it's a data structure that allows applications such as search engines to use a table of word equivalents. For example, Wordtracker is a data structure that allows applications such as search engines to use a table of word equivalents. Unlike the other two thesauri above, Wordtracker is designed to be accessed by computer programs instead of humans.

A common vocabulary

The process of arriving at a common vocabulary is a good way to introduce and describe the basic taxonomy building blocks — metadata, authorities, thesauri, ontologies — as well as stress the value of accommodating multiple organization schemes.

> 2. How do we develop a common vocabulary?

Translators

Although it's possible to introduce basic taxonomy concepts in a seminar, it usually takes a lot of discussion and examples to get all team members on the same wave length. In our experience, it's very useful to have a translator on hand as a facilitator. The "translator" should be fluent in the following "languages:"

Information technology (relational databases, data modeling, metadata repositories, etc.)

• Library science (subject matter thesauri, authority files, bibliographic metadata standards, etc.)

• Journalism (style guides, structured composition, usability, etc.)

• Project management (objectives, tasks, schedules, deliverables, etc.)

• General management (planning, budgeting, evaluating, etc.)

If you can't find one person that can speak all five languages, try to include a few "boundary spanners" on the team — employees that regularly interact with two or more departments or functions.

Common agenda

The process of exploring all the perspectives and issues can be stimulating. The next task is to refine the objective so that it can be budgeted, implemented, and evaluated — a process we call domain definition. In defining the domain, the team identifies the business context and describes the characteristics of the users, content, and work flow involved. Domain definition comes naturally to the problem solvers and "end users" on the team because they're on the front line. They can pull the answers out of their experience — they don't have to rely on surveys or focus groups.

Defining one domain or application isn't too difficult, but most taxonomy projects span multiple applications. Therefore, it's useful to define at least two domains with a common area of interest. For example, marketing and product development have different vocabularies and use different business processes, but they share a common interest — the product or service to be sold.

> 3. How do we define at least one domain, preferably two?

Key issues

Domain analysis brings to light several key issues:

One taxonomy isn't enough. Even when two business processes use information about the same entity (e.g. a customer), it's likely that the type of data collected, the data source, the frequency of updates, the methods of accessing the data, and the level of detail needed for each process are different. It may be possible to have a single taxonomy architecture, but it's not possible to have a single organization scheme.

> 4. How do we distinguish between a taxonomy (organization scheme) and a taxonomy architecture (the structure and relationships of components in a system of overlapping organization schemes)?

Specialized taxonomies should be linked and shared. Why shouldn't everyone use the same geographic or product codes? Why should the user have to know which collection to search or which department created the data? Why shouldn't the search engine have access to a list of synonyms and definitions, so that when the user searches for COTS (meaning Commercial Off-the-Shelf), he gets information about software instead of beds (cots)?

There are a variety of tools and techniques available to provide access to standard terms and create cross references among similar terms. The team should understand the options, as well as their pros and cons.

> 5. What tools and techniques are available for linking specialized taxonomies?

Departments are both taxonomy users and creators. Content creators are key to keeping taxonomies up to date. If a taxonomy system saves time and yields better information, content creators will not only use it but will

invest in helping to maintain it. The goal is to make it easier to use existing taxonomies than it is to create new ones.

6. What role does content creation play in taxonomy maintenance, and how can the system encourage authors to use taxonomy data?

Proprietary data formats make it harder to link and share. The ease-of-use associated with packaged software sometimes has a hidden cost — the difficulty of exchanging data with software from other vendors. Taxonomy values locked in one proprietary format often can't be imported or accessed by other applications. The trick is to develop a taxonomy system that leverages investments in existing software and allows data interchange among programs (i.e. a "vendor neutral" architecture).

7. What methods are available to extract, exchange, and use taxonomy data stored in proprietary software formats?

Metadata repositories and solutions

A growing number of companies are getting interested in metadata repositories as an answer to these issues. Metadata — information about information — are the building blocks of taxonomies. Examples include:

• content attributes, such as author, title, subject, publication date;
• user attributes, such as topics of interest, security level;
• system attributes, such as database name, software version, date updated.

A metadata repository is a data structure or "virtual holding area" used to store metadata and/or the pathways to external metadata. Some metadata is entered manually through forms during the content creation or classification process. Other metadata is entered automatically by harvesting it through a computer program (see "The economics and ABC's of indexes"). These two methods are illustrated below.

A metadata "solution" consists of one or more repositories along with:

• methods of automatically extracting (harvesting) metadata from documents and databases (see "Quiver's QKS Classifier: a hybrid categorization tool");
• an architecture that unites all the elements (see "What is architecture?").
• methods of accessing and displaying metadata (by both humans and computer programs).

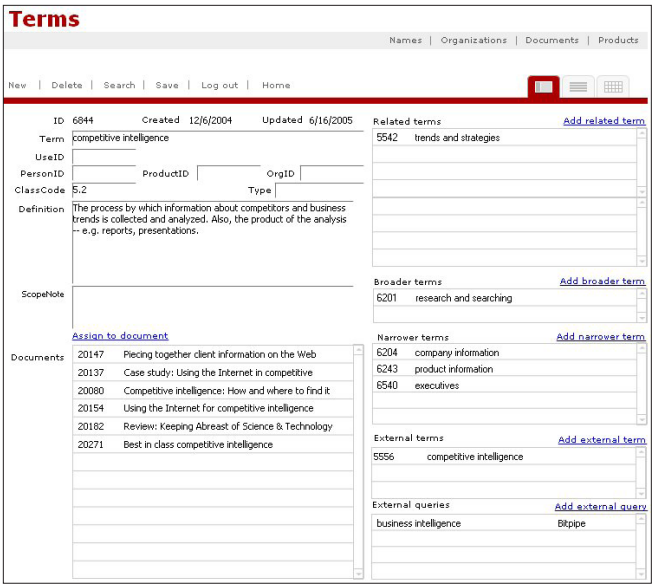
There are two aspects to metadata solutions — conceptual and technical. That's why it's important to have a multi-disciplinary team working from a common set of assumptions, definitions, and objectives. The problem is that both the conceptual aspects (represented by editors and librarians) and the technical aspects (represented by IT staff) are detailed and complex.

8. What are the pros and cons of metadata repositories and solutions?

Theory vs. reality

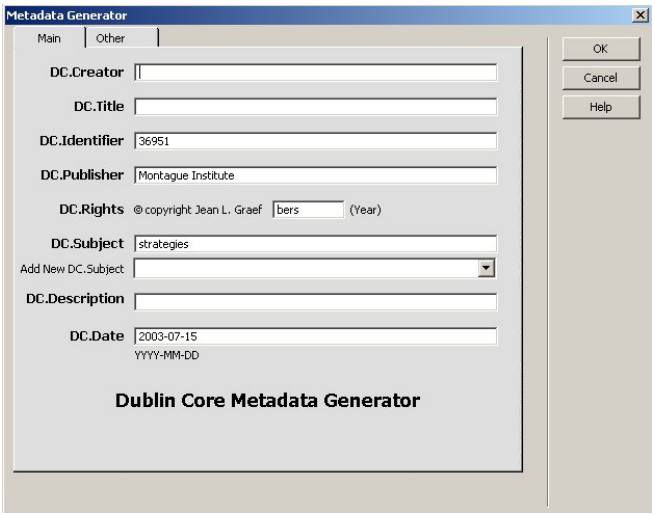
Taxonomies are by definition abstractions. A good way of coping with the abstraction is to conduct a pilot with real users, real content, and a real business application. In this case, the pilot objective is not to test a specific software program or even a specific metadata repository design. Instead, it's designed to test all aspects of the issue — content quality, data interchange, maintainability, external standards, training and support services, ease of use. Ideally, the pilot should involve at least two domains.

9. How should we design, implement, and analyze a taxonomy pilot?



*Above: **Metadata in a database**. Terms and cross references are normally entered manually into this metadata repository. However, because it's a relational database, multiple terms can also be imported from a file or inserted via data interchanges technologies such as Web Services.*

*Below: **Metadata in a document.** Authors can enter metadata for a Web page using this Dreamweaver form. Microsoft Office documents have similar metadata summary forms. Dublin Core is a popular metadata standard for Web documents.*



Organizational "readiness"

The pilot helps the team assess the organization's readiness to develop a taxonomy architecture and metadata solution. Among the questions to be answered are:

• Do we have the metadata we need?
• Do we have a process for keeping the metadata up to date?
• Should we use a phased or an corporate-wide development approach?

It's better to find out up front that some department managers fear losing control over their data or that senior executives believe the proposed solution is too expensive. When some parts of the organization are not ready to move forward with a proposed taxonomy solution, the team can:

• wait a year or two for business conditions to change priorities;
• market the pilot results internally as a demo;
• proceed with a phased implementation for one or more departmental solutions, using a vendor-neutral architecture when possible and international metadata standards when available.

10. How do we assess organizational readiness and what are the options for a phased implementation?

Striking a balance

Increasingly, taxonomy development is being conducted by inter-departmental teams. It's necessary for the team as a whole to learn how to tap the existing skills of individual members, bridge conceptual gaps, define domain requirements, and develop an implementation strategy consistent with the organization's readiness. In answering the questions posed here, teams should be better equipped to strike a balance between centralized and decentralized as well as short term and long term strategies (see "Upstream knowledge management").

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