Everything You Ever Wanted to Know about Taxonomies … but were afraid to ask

Susan Hanley

The term **taxonomy** comes up a lot in discussions about both portals and content management solutions. Most of us have a vague idea about what a taxonomy is, but because it's an official sounding word, we assume that there is some magic or hidden secret in taxonomies. Taxonomies *are* absolutely critical to successful portals and content management solutions. However, they aren't as scary as they seem. This brief paper answers a few of the basic about taxonomies, explaining why they are important, what makes them so tricky, and what you really need to know.

What is a Taxonomy?

A taxonomy is a collection of relevant topics and subtopics arranged in a hierarchical or networked structure. A library card catalog is a classic example of a taxonomy. The hierarchical structure on Yahoo is another example. In Microsoft SharePoint Portal Server, the taxonomy for a collection of documents is represented in the Topic hierarchy as well as in the overall site architecture.

Why is a Taxonomy important?

In a portal or content management system, an effective taxonomy helps users to navigate to documents in which they are interested without having to do a search (although, in practice, most studies seem to show that users use a combination of taxonomy navigation and search when both are available). Taxonomies also allow users to see documents in a context, which helps the user assess whether a document is relevant for what they are trying to accomplish.

How do I get started building a Taxonomy?

Three key skills are required to effectively build taxonomies. The first, and probably most important, is content organizational skills – a combination of data modeling and library science. The second is some knowledge of the domain to be modeled. The third is knowledge of the end user of the applications that will leverage the taxonomy. In general, when there are trade-offs to be made in taxonomy design, design for the end user of the content, not the contributor.

There are several ways to get started generating a taxonomy. Manually developing taxonomies requires significant effort and cost, but some manual effort is almost always required. A good taxonomy designer starts with existing structures – typically found in folder hierarchies, existing Intranets, industry sources, or organizational charts (though the most useful business taxonomies will be organized functionally, not organizationally, since organizations are rarely static) – and uses these existing structures as a starting point to review proposed taxonomies with domain experts. There are also automated classification tools that attempt to suggest a taxonomy by analyzing the content from a collection of documents. Some of the automatic classification engines include machine-learning algorithms that help the engines train themselves from example data. At best, today's automated classification systems can help get started with taxonomy building. For the most part, building effective taxonomies requires at least some manual effort.

What makes a good Taxonomy?

A good taxonomy has to be comprehensible to users (so they can use it for navigation with little or no training) and has to cover the domain of interest in enough detail to be useful.

What are some of the challenges associated with Taxonomies?

When you first deploy a content or knowledge management solution, the taxonomy is well structured and, usually, content is appropriately catalogued because designers and application sponsors have taken a lot of time to ensure that the initial implementation is successful. Over time, new content enters the system, along with new knowledge areas, and before you know it, the well structured taxonomy is devolving into chaos.

There are many reasons that a taxonomy can degrade over time, including:

- End users may incorrectly associate content with a topic
- When the system allows end users to add topics of their own, they may create a redundant topic
- If the search engine doesn't support a thesaurus, new terms may get added that are merely synonyms for existing terms
- When users can't find a "bucket" in which to place their new content, they may put it in a "miscellaneous" topic, which makes searches and queries far more difficult
- The organization can change direction so that the taxonomy becomes less relevant to the business

When the taxonomy becomes less relevant, so do the applications that depend on it. When that happens, users become frustrated and management wonders why they continue to make investments in IT.

The key to overcoming the challenges associated with maintaining taxonomies is to recognize up front that maintaining a taxonomy requires a continual investment. Building a successful taxonomy is not a "build it once and walk away" process. There are several business process recommendations that we've learned that can help our clients manage and maintain their taxonomies:

- Assign **content managers** to ensure that new content is assigned correctly. Content managers can be domain experts who allocate a portion of their time to review new contributions to the portal or content management system. Content managers can also be **librarians**, specialists who help design meaningful taxonomies, tag content as it appears, and maintain the taxonomy over time. Yahoo has more than 200 librarians on staff who assure accuracy and quality for what may be the largest taxonomy on the planet. Microsoft employs 7 librarians just to manage and maintain the taxonomy for their internal intranets. The new graduates from schools of library science are often highly trained information architects who have a significant role to play in the information economy.

- Establish **governance policies** for managing the taxonomy's structure and adding new documents into its directory. Governance policies should define who does which tasks, procedures for performing tasks, and feedback mechanisms for suggesting changes and improvements.
- Leverage **automated classification software** if the volume of content is too large for librarians to study and classify manually.
- **Revise** the taxonomy on a regular basis. At a minimum, conduct a taxonomy review once a year (or more frequently if content is being added continuously).
- **Maintain the content itself** by archiving old documents and monitoring document usage so that content that is not current or is no longer relevant does not appear in search results.

<u>What is the Dublin Core and why do I hear about it associated with taxonomies?</u>

The **Dublin Core**[1] is a set of 15 elements of meta-data intended to facilitate discovery of electronic resources. The Dublin Core has been in development since 1995 through a series of focused invitational workshops that gather experts from the library world, the networking and digital library research communities, and a variety of content specialties. The idea behind the Dublin Core is a set of standard metadata[2] elements that should be included on documents in addition to their domain-specific taxonomy categories.

The following table summarizes the 15 metadata elements in the Dublin Core:

Content	**Intellectual Property**	**Instantiation**
Coverage	Contributor	Date
Description	Creator	Format
Type	Publisher	Identifier
Relation	Rights	Language
Source		
Subject		
Title		

[1] For more information about the Dublin Core initiative, see www.dublincore.org.

[2] Metadata is data about other data. It is the Internet-age term for information that librarians have traditionally put into catalogues and it most commonly refers to descriptive information about online resources.