Taxonomy Development Advice

Accessing and using your information is what you need to do. Developing a taxonomy in conjunction with ECM can help you do it.

By Tom Reamy

Taxonomies and enterprise content management (ECM) have a great deal in common. They are both enterprise resources (at least if they are done correctly) and have historically been underachievers—too big and cumbersome to develop, and not enough real business value. So why would you want to put the two together and get two giant projects that fail? Actually, there are three good reasons. First, taxonomies are beginning to deliver great value. Second, it turns out that if you combine taxonomies and content management, you can get greater value out of each and the synergy creates even more value. Third, if you do not combine them, you are much more likely to fail attempting either one. In fact, according to a Gartner research study last year, 70 percent of new content management projects would fail due to an underinvestment in taxonomy.

The fundamental reason for the increased interest in taxonomies is simple—despite years of effort and new technologies, professionals are still spending more time looking for information than actually using the information they find. The basic problem is also simple: search engines search for text strings, not concepts, and that means that they do not understand meaning. The text strings could be chicken scratches or random marks on a rock; all the search engine can do is look for that exact mark somewhere in a few million documents.

Google, of course, helps, but only on the Internet where its link analysis algorithm works well to find the most popular website or document that contains a lot of those chicken scratches. Moreover, search engines are getting better with things like stemming, but even then, there is very little conceptual meaning. Any good search engine can "understand" that law and laws are essentially the same. However, how does that work in legal, jurisprudence, or courts? How are they related to what you are looking for?

Finally, in the enterprise world, search is much more difficult—Google's link analysis algorithm doesn't work, and you are very often not looking for the most popular website or document, but one official document.

Taxonomies to the Rescue?

How do taxonomies help? Taxonomies help by introducing a level of meaning to the chicken scratches and thereby helping people find information. However, there are different types of taxonomies and different types of semantic structures, and how they help is in part dependent on what type they are.

It is beyond the scope of this article to discuss the different types in any detail, but there are essentially three kinds of taxonomies: 1) browse taxonomies, 2) formal taxonomies, and a new form of taxonomy/metadata application, 3) faceted taxonomies. There are different applications that can be built with the three types, but the important thing to remember is that all three types are enterprise resources that support a wide range of information access applications.

Taxonomy and Content Management

While both taxonomies and content management are considered platforms for building or supporting information applications, it is important to remember that neither ECM nor taxonomies are ends in themselves. They both work in connection with search, search applications, and other information and knowledge applications. They also perform that platform function much better when they are combined, but only if they are combined properly.

The basic way that a taxonomy can be used within an ECM application is through tagging documents as they get published with subject matter keywords. This is something that users hate and are not very good at if left to their own.

So how do you get good results that will enable users to add good metadata tags, ones that will enable enhanced searching and finding of information? The first issue is how to avoid user overload, both real and perceived. If you only add taxonomy input into an ECM interface, chances are your users will rebel and creatively figure out ways to avoid using either the taxonomy or the ECM software. There is not a simple, single technique on how to do it, but rather, finding the right blend of four dimensions or enterprise contexts: people, process and policy, technology, and semantics or content and content structure.

The fundamental reason for the increased interest in taxonomies is simple—despite years of effort and new technologies, professionals are still spending more time looking for information than actually using the information they find.

Content Management, Taxonomy, and the Four Contexts – in Practice

Semantics. Let's start with semantics; the basic process of adding metadata to content. Thinking up keywords to describe a document is a difficult skill and is very different from simply being a subject matter expert. In fact, people who know a topic very well often choose very different kinds of keywords than non-experts who are looking for that information.

However, the quality of keywords can be greatly improved by asking people to choose a word from a well-organized set of concepts such as a hierarchically organized taxonomy. Choosing words from a list is cognitively easier than creating it and the structure of the taxonomy gives additional context to the decision, thereby enabling better choices.

In addition, the structure of the taxonomy gives more power to the keyword by enabling the search engine to place the users input in a well-ordered relationship with other terms. For example, by "understanding" that courts can be a law related term and that jurisprudence is related in a number of interesting ways.

However, you still have the problem of asking users to add metadata and this is where people, process, and technology can all help.

Process and Policy. ECM creates a set of procedures and policies around publishing a document that fosters an environment that is already more formal and complex than simply saving to your local or shared drive. These policies and procedures can easily be extended and adapted to incorporate the use of taxonomies. The psychological cost of adding one or two steps to a process that has already added five new steps is relatively minor, as long as the addition is not too onerous and the benefits are well socialized.

Technology. In addition to standard "carrot and stick" ECM policy methods, technology can help a great deal. For a recent project, we recommended a faceted taxonomy approach that works extremely well for finding information, but requires that even more metadata have to be added—at least one subject keyword plus one value for each facet. The client wanted as few facets as possible, but more facets tend to work better for finding. The solution was to use taxonomy/text analytics software to auto-populate as many of the metadata fields as possible.

Let's say we have a keyword metadata field and three facets: organization, client, and instruments. The auto-categorization function could offer suggested subject categories for the keyword field. Plus, an entity extraction feature could find all the client names that the document contained as well as specific instruments. The organization facet could be auto-populated either by entity extraction or more likely, by a simple publishing rule.

The net result is that users are presented with simple selections instead of the daunting task of filling out four metadata fields. Most users will probably just accept the suggested values unless there is something quite strikingly wrong with them.

Social/People. Finally, the social dimension can be used to enhance the process of adding metadata and making it easy, and sometimes even fun. Folksonomy (a really bad name) and other Web 2.0 concepts are getting a lot of press lately. While much of it is more hype than reality, if folksonomy tag clouds are combined with a central team of editors or taxonomists (even Wikipedia has taken the plunge), they can be a valuable tool to both keep a taxonomy up-to-date, and to make the experience of adding metadata easier and more enjoyable.

Content Management, Taxonomy, and the Four Contexts – Strategy

The success of a combined taxonomy/ECM initiative depends on one other essential ingredient; a well developed and articulated strategic vision. This strategic vision can be built by developing a deep understanding of the same four dimensions or contexts discussed above. We have used a technique called a knowledge architecture audit to develop this deep understanding for a variety of projects from developing a taxonomy to developing a strategic plan.

It is beyond the scope of this article to go into this process in any real depth, but it consists of a series of interviews, focus groups,

text analysis, and other research techniques to come up with a knowledge map that characterizes the four dimensions within a particular environment and the relationship among them.

Semantics. A good starting place is to develop a content catalog of all the content within an organization. This includes structured and unstructured, internal and external, documents and digital assets. It should also include a basic categorization of the content and any associated structural elements like vocabularies, glossaries, taxonomies, and metadata. This catalog becomes a resource that can feed into any number of projects such as ranking content to create different levels of metadata efforts or doing a gap analysis to drive the purchase of new external content.

It is essential for a good ECM policy to develop a metadata standard that is truly a standard (i.e., for the whole organization). Too often, IT, training, research, and even a business unit are all independently developing metadata "standards."

Social/People. The second dimension is the social dimension or people and communities. Typically, we try to develop a map of all the formal (organizational) and informal communities or networks within an organization characterized by subject matter and primary and secondary communication channels (i.e., portals, push or pull, iPods or CDs, or person to person). We also try to discover and map any community-specific vocabularies and categorization schemas.

Another important aspect is the various information behaviors and needs of individuals and communities. Groups that require known item look-ups in a few seconds have fundamentally different content and presentation needs than a research group that needs to research a topic extensively for a week or so.

One other important theme in this area is how to create a taxonomy/search/content management team with central editors, distributed authors, subject matter experts, IT support, and high-level executive participants. Neither top-down nor bottom-up approaches work very well, but what is usually best is a distributed and integrated team with specific roles. Also, this is an infrastructure and interdisciplinary team, not a simple IT housed or ECM/library housed taxonomy team.

Technology. The third dimension, technology, is where many people start without a clear understanding of how technology functions in the overall enterprise, or how making good decisions and getting good value is almost pure chance. Some of the important themes in this area are:

- Using technology to develop content structures or taxonomies
- The relationships with taxonomy software, content management, search, and other specialized information applications
- The development of enterprise platform technologies like knowledge management (KM) or learning management system (LMS) software

There have been major advances in taxonomy and text analytics in the last few years with new categorization techniques, better vocabulary extraction and analysis software, and more mature products. Many of the best of these products are stand alone, but both search vendors and content management vendors have been buying or developing these capabilities and integrating them. A good case can be made for any of those options, but the main split is between search and ECM with the latter requiring authors to add metadata with help and search trying to automatically categorize on the fly.

My own experience is that the answer (as is so often the case) is all of the above. Taxonomy development and content tagging need the same functionality and it is simply a matter of how to integrate them. On the other hand, these features at search time can be valuable, but only if you have a taxonomy with highly developed categorization rules— something for which you need categorization and text analytics functionality.

If you only add taxonomy input into an ECM interface, chances are your users will rebel and creatively figure out ways to avoid using either the taxonomy or the ECM software.

Process/Activities. The last dimension that needs to be researched and well understood is the actual business or day-to-day activities that all of this categorizing, searching, and managing of content is designed to support. This is the area where real value is created, and it is the most variable and dynamic of the four dimensions. It can include everything from formal process management maps to very high-level descriptions of generic processes to specific actions.

And finally, if you map the four dimensions against each other you have an integrated foundation on which you can build almost any information or knowledge initiative that you want to create the kind of enterprise changing impact that content management, search, and taxonomies have been promising for so long.

Tom Reamy (tomr@kapsgroup.com or 510-530-8270) is chief knowledge architect at KAPS Group (www.kapsgroup.com).