DOCULABS

Doculabs White Paper:

Ten Questions on Taxonomy

Content management is now becoming an enterprise-wide imperative, as organizations are being driven by compliance requirements and competitive pressures to get their unstructured content under control. Likewise, categorization of that content is increasingly becoming an imperative, to enable users to share information across the organization. The tool that makes this possible is an enterprise-wide classification for the organization's documents and records – an enterprise taxonomy.

A taxonomy allows you to address where and how content is stored – in advance. Unlike a search tool, which attempts to locate information after the fact, a well-thought-out enterprise taxonomy helps your users put things in the right place the first time, greatly easing subsequent retrieval and allowing your organization to make optimal use of the technology it deploys to manage content. For this reason, defining how unstructured content is to be organized within your enterprise content management (ECM) system is as critical as rolling out the technology itself.

What does it take to create an *enterprise* taxonomy – a classification that reflects the business-critical documents of the organization as a whole?

In this white paper, Doculabs answers the top ten questions that our consulting clients are asking us about taxonomy, to help you understand the key concepts and the issues involved in creating and implementing an enterprise-wide classification of your organization's information assets, and to help you plan for a taxonomy initiative for *your* organization.

Overview

Why is an enterprise taxonomy a critical aspect of an ECM deployment?

- To provide a structure that allows users to classify information appropriately at the time it is created
- To enable users to share information on an enterprise basis
- To provide a consistent user experience
- To make searches more efficient (including discovery searches for litigation support)
- To enable automated capture of metadata attributes needed to classify a document (including attributes needed for records management)

Information overload is a common complaint in organizations today, as electronic content (desktop documents, scanned images, e-mail) continues to grow at an ever-increasing pace. It becomes frustrating for executives as well as knowledge workers when valuable content can't be used effectively – either because it can't be found, or because it is organized in ways that are unique to each department, making the discovery and assembly of relevant information assets either cost-prohibitive or impossible altogether.

Many organizations are now looking to address the problem by implementing enterprise content management (ECM) technology across their organizations – taking ECM enterprise-wide. These technologies are mature and provide all the functionality an organization needs to get its content under control. But control of unstructured content is only part of the solution. The other part is making content available and easily accessible to users throughout the organization – ensuring that users can search, find, and reuse the information assets that are stored on the ECM system. Thus a critical component of any strategy for truly "enterprise" ECM is a classification for all of the organization's documents and records – an enterprise taxonomy.

In most organizations, each department has developed its own classification for categorizing the documents its users rely on for their business processes, with categories that meet the business purposes of the department and using terms that make sense to users within that department. But many documents cross multiple departments – and potentially multiple business processes, too. If your goal is to provide cross-organizational access to information, then your company's documents and records need to be stored in the ECM system in categories that cross all of these information domains.

In recent years, an estimated 50 percent of "enterprise" ECM deployments have failed to reach enterprise adoption levels, much less to deliver the expected business benefits. In these instances, user acceptance is less than optimal, and at some point, many of these deployments tend to stall out.

Doculabs has consulted for many organizations that have struggled with less-than-enterprise-wide deployments of ECM technology. What we've found is that most of these systems were implemented with little or no regard for how the content was to be organized at an enterprise level. If an organization is to realize its key ECM objectives, it cannot rely on the technology alone – it must also come to a company-wide consensus on how it will categorize the information assets it plans to manage using ECM technology. That consensus is realized through the development of an enterprise taxonomy.

In this white paper, Doculabs makes the case for why you need an enterprise taxonomy, and explains what an enterprise taxonomy can do for your organization. We also provide some basic guidance to help you get started – and, just as important, guidance on how to maintain your taxonomy and ensure that it continues to reflect the way that documents are used across your organization, even as that organization grows and changes.

Doculabs' Top Ten Questions

Why the new interest in content classification?

The unprecedented growth in unstructured content is behind much of the new interest in content classification, particularly from the content management and storage management standpoints.

But compliance, records management, and legal discovery concerns are also driving organizations to find ways to retrieve information faster and more efficiently.

This white paper compiles the top ten questions our consulting clients are asking us about enterprise taxonomy, together with answers that will help you as you begin to consider your own organization's taxonomy requirements. The questions are as follows:

1. **What is a taxonomy?**
2. **What are the key components of a taxonomy?**
3. **Why does my organization need a taxonomy?**
4. **What is involved in developing a taxonomy, and how long does it take?**
5. **What is the difference between records management and taxonomy programs?**
6. **We already have a records retention schedule. Can't we use that for our enterprise taxonomy?**
7. **Which should come first: implementing a content management system, or developing the taxonomy for organizing content on the system?**
8. **Should we use an independent tool, or can we use the taxonomy that comes with our content management system?**
9. **How many index fields does my organization need?**
10. **What is involved in maintaining a taxonomy, and where should the ownership of the taxonomy reside?**

1. What is a taxonomy?

Discussion

The most common uses for taxonomies:

- To organize content and aid in navigation
- To support compliance
- To retrieve information and clarify results
- To identify patterns and overlaps according to common interests and association
- To facilitate collaboration
- To organize projects, processes, and other abstract items by type, topic, and other metadata

A taxonomy is a high-level classification scheme. It is typically in the form of a hierarchical classification that shows the relationships of the entities within the hierarchy. In the context of ECM, a taxonomy is a structure for classifying documents and records; it focuses on the relationships of how information is used within an organization.

The hierarchy of an enterprise taxonomy is represented by a tree-like diagram that shows the relationships among standardized topics or subjects within a company or organization. It also shows relationships between topics that span the hierarchy and defines a controlled vocabulary of agreed-upon terminology that is to be utilized by all persons who create information and store it in the ECM repository.

The structure of an enterprise taxonomy generally has three levels:

- **Area:** the highest level, usually corresponding to a key business area or organizational subset, such as Human Resources, Operations, or IT.
- **Function:** a group of common processes within an Area, typically processes with considerable overlap that are managed by a common executive or team. Within the Human Resources Area, examples include Recruiting, Employee Management, and Training and Development.
- **Specialization:** a process within a Function that has its own unique set of document types. Within Employee Management, typical document types include documents associated with time reports, periodic performance reviews, and general personnel files.

The figure below shows the three-level hierarchical structure of a taxonomy.

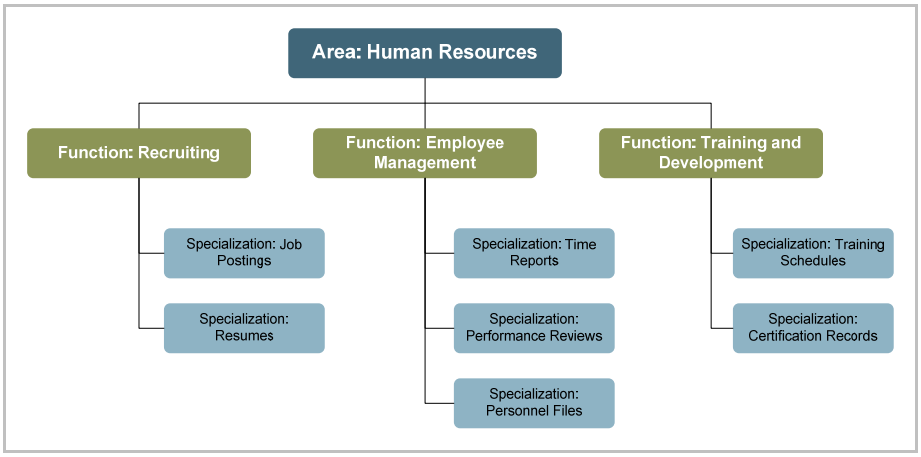

Figure 1: Example of the Taxonomy's Hierarchical Structure

An enterprise taxonomy allows the categorization of documents that cross business processes, or of documents that serve business processes that cross multiple departmental boundaries. As such, a taxonomy allows these documents to be retrieved by users who need access to them, irrespective of the department or business unit in which those users work.

Factors to Consider

A taxonomy has many different representations and can be used for a variety of purposes, depending on the context, application, and audience. A taxonomy can range from the simple hierarchical directory structure, to a navigational taxonomy, with attributes that pull from controlled vocabularies. It can also include the metadata standards that need to be aligned with the taxonomic structure, along with mapping of fields between applications (where a branch of a taxonomy can feed multiple systems).

As the foundation architecture for managing documents within an enterprise; a taxonomy also serves as the foundation architecture for records management (see Questions 5 and 6 of this white paper). It facilitates the management of recorded information throughout its lifecycle – from the creation of a document, through its capture in the ECM system and its management, to its ultimate disposition.

Clearly, taxonomies have many constituencies and stakeholders, which underscores the need for governance processes and change management procedures in order for the taxonomy to retain its value in the long term (see Question 10).

Doculabs' Opinion

A taxonomy must be enterprise-wide if its benefits are to be fully realized. Within your organization, many departments may already have classification schemes for their own documents. An enterprise taxonomy, however, enforces consistency in the way that information is organized across all departments of the organization.

While you may leverage aspects of these existing document classifications toward the development of an enterprise taxonomy, it is unlikely that a departmental taxonomy can be adopted for enterprise use. Consider a specific functional area such as Accounting (or, in a multi-line financial institution, a specific business line such as Mortgages), which may already have built its own complete, well-thought-out taxonomy. A department-level taxonomy such as this becomes inadequate once external factors are brought into play. The isolated model proves too inflexible to be applicable beyond the boundaries of the department; in addition, the departmental taxonomy will likely present redundancies once other departments' documents are included. Finally, departmental taxonomies are likely to present competing governance structures that make them unworkable for enterprise adoption.

2. What are the key components of a taxonomy?

Organizational functions to involve in your taxonomy initiative:

- IT
- Subject Matter Experts (SMEs) from the lines of business
- Records Management
- Legal/Compliance

Discussion

Any enterprise taxonomy development effort should address four components, as shown in the figure below. Together they form the foundation for effective, ongoing content and information management on an enterprise level.

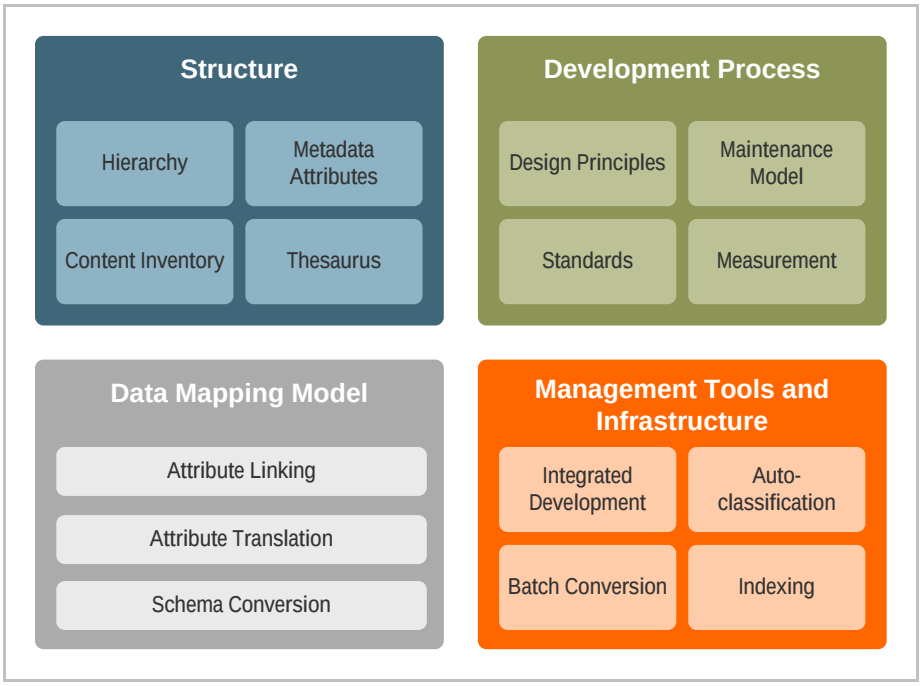

Figure 2: The Components of a Taxonomy

Factors to Consider

Each of the components includes a number of key elements, as listed below:

Structure

- *Hierarchy* – the classification that shows the relationships of how information is used within an organization, represented by a tree-like diagram
- *Metadata attributes* – literally, data that describes data; the "tags" given to documents and records, corresponding to key characteristics (such as Author, Title, Document Type, Subject) that are attached to each document in the ECM repository, typically assigned to the document at the time the item is saved in the repository
- *Content inventory* – the itemized list of an organization's content sources
- *Thesaurus* – inter-departmental cross-reference that lists the "approved" terms, with clear definitions of what the terms mean and when the terms should be used

Development Process

- *Design principles* – a detailed description of the overall taxonomy development approach; for example, should the taxonomy be primarily structured around business processes or by organizational structure or when is it necessary to break a single taxonomy node into multiple nodes, etc.
- *Maintenance model* – a framework of defined policies, procedures, and staffing to manage and maintain the taxonomy over time and to enforce the use of the taxonomy throughout the organization
- *Standards* – the documented specifications to be used in assigning documents and records to the structure of the taxonomy
- *Measurement* – the ability to assess the impact the taxonomy has on business metrics, as well as the ability to measure the value of the taxonomy itself (which can manifest as user satisfaction, fewer searches to find information, etc.)

Data Mapping Model

- *Attribute linking* – connecting the key characteristics of documents (such as Title, Creator, Subject, Description) in order to carry forward relationships that exist in the business use of the document
- *Attribute translation* – ability to map attributes and metadata in one system to attributes and metadata in a different system or repository in order to provide seamless information integration between the two systems
- *Schema conversion* – providing the capability to map one schema to another in order to facilitate data transfer (import or export of data between two systems, usually using XML)

Management Tools and Infrastructure

- *Integrated development* – development tools and/or development environments that allow people to create, maintain, test, update, and deploy taxonomic artifacts and information
- *Autoclassification* – the ability to classify content by analyzing its text and then automatically assigning it to a pre-defined class
- *Batch conversion* – a program that can leverage taxonomic information to mass-load content from one file format or repository into another without user interaction
- *Indexing* – assigning standard, pre-defined metadata to documents prior to storing them in the repository, such that the metadata can then be used as search terms for retrieving documents from the repository

Doculabs' Opinion

All four components are critical to the implementation of an enterprise taxonomy. Furthermore, addressing each of the components requires the participation of both IT and the business units. While the taxonomy bridges all the document-based systems and content repositories of an organization, ultimately, it must serve the needs of the users.

In many case, a taxonomy is a hybrid of how documents are stored from a database perspective, with business hierarchy placed upon it. The risk, however, is being too tied to a structure that a system has forced upon you.

Subject matter experts from the business should be involved in the taxonomy, as they represent the best source of current process information. They have the best understanding of how content is actually used on a day-to-day basis and how document types should be ordered and classified for greatest efficiency. Other groups to involve include Records Management, Legal, and the Compliance function of your organization.

3. Why does my organization need a taxonomy?

Discussion

The implementation of a common taxonomy is particularly critical for organizations that have multiple content management solutions in place. Without an enterprise-wide classification of unstructured content, users that access the various systems are confronted with a confusing and inconsistent user experience that can lead to the loss of information.

Large, global organizations in particular stand to benefit from implementing enterprise taxonomies. A taxonomy provides a controlled vocabulary and consistent reference framework which can facilitate social networking and collaboration, thereby creating organizational synergies. Departments, and even lines of business, can more easily collaborate across divisional and geographic boundaries – which in turn can reduce costs or increase innovation.

In many organizations, as the volume of unstructured content has grown, users have been demanding improvements in how that content is organized – or at least better ways of searching across that content. The ability to search is the major reason for putting an enterprise taxonomy in place. Tens, hundreds, even thousands of people in an organization can be creating content on a daily basis. In this growing volume of unstructured content, the challenge is to find a way to categorize this information at the time of creation so that the knowledge within the content can be made available to the organization as a whole.

Putting in place a common taxonomy that all users follow in creating and storing their documents results in a consistent user experience. This consistency in turn helps to ensure that information is not only captured, but also that the information is organized in a way that makes searches more efficient. A taxonomy thus makes it easier for users to retrieve the information they need, when they need it. This is a key aspect of discovery searches for litigation support, as well as for business intelligence.

Figure 3: Taxonomy as the Bridging Foundation of Information Management

Factors to Consider

While an ECM system provides a central location for storage and retrieval of information, many of these solutions provide limited ability to associate documents or content with a master classification or taxonomy. Instead, the systems rely on search functionality, such as keyword search or full-text search, to enable users to search a particular subject. The limitations are clear: the larger the organization and the greater the volume of content in its repositories, the more unwieldy the search results. And different departments may assign different keywords to their content – in which case a search conducted across departmental boundaries will not turn up all relevant documents.

The application of taxonomies to other content sources

A taxonomy is particularly useful in organizing content for intranets, extranets, and portals – places where consistency is critical, given that their categories and subcategories of topics are likely to be accessed by users representing many constituencies, and must be easily navigable to allow those users to find the information they are looking for.

Implementing a taxonomy in conjunction with ECM provides a predetermined context that supplements the powers of the system's search tools. The result, from the user's perspective, is clearer search results and more effective retrievals – and an improved work environment, particularly for knowledge workers, whose productivity may well increase. We've found that implementing an enterprise taxonomy generally provides very good cost/benefit and efficiency returns across *all* user constituencies, in the form of improved response time and better decision-making – the various quality-of-service benefits that arise when users are able to search enterprise content more effectively.

One of the most significant benefits of an enterprise taxonomy is reduced time (and costs) spent on discovery. A major issue in finding documentation to respond to a lawsuit is the fact that such information typically is located throughout the organization, in multiple repositories – and is probably organized differently within each location.

A taxonomy also makes it easier to manage an organization's content from the standpoint of the information lifecycle. When enterprise content is categorized, it is easier to manage the various categories of content as records, applying records management retention and disposition policies and rules based on taxonomic metadata. It provides a better understanding of the usage and lifecycle of content, allowing the IT function to apply the right storage solution to specific content types to manage the organization's content more cost-effectively from the storage and operational perspectives.

Furthermore, when ECM is deployed with a well-developed taxonomy, the ECM system itself will be able to automatically capture more than 90 percent of metadata attributes needed to classify a document, including attributes needed for records management, through contextual references and pre-defined classification rules. Automating the capture of metadata minimizes the amount of content that users will have to enter when they store a document – which can greatly increase user acceptance of the ECM system itself, leading to higher adoption rates.

Doculabs' Opinion

Today, content management is an enterprise-wide need. To obtain maximal business benefit from enterprise-wide deployment of ECM, information classification must be undertaken on an enterprise scale, if users are to share information on an enterprise basis.

We see many companies where information is fragmented because it is organized by business function. Putting in place an enterprise classification of information not only leads to more effective information retrieval, it also makes it possible to unify information from legacy systems into a single, virtual repository for the enterprise – i.e. a single source for the organization's information. In organizations with many operational systems, a taxonomy also simplifies the integration of existing content applications and any new ECM standard components that you may add to your technology environment, because the taxonomy development process forces identification of precise vocabulary for organization of documents and records.

4. What is involved in developing a taxonomy, and how long does it take?

Questions to ask at the outset of a taxonomy initiative:

- How large and complex is your organization?
- How complex are your business processes?
- What are the common goals of authors, content creators, and others who participate in building your organization's institutional memory?
- What types of content are being created, and where and how is it stored and retrieved?
- What are the common vocabulary terms used within your industry, and how does your organization's vocabulary differ from the "standard" terminology?
- From a cultural perspective, can your users understand the benefits of classifying information, and do they have the discipline to enforce its classification?

Discussion

The first step in the process of developing an enterprise taxonomy is to assemble a team of business users and domain experts, together with IT and records management personnel to oversee the ongoing taxonomy development process.

This team will then compile and evaluate your organization's current document inventories, file plans, and retention schedules. Surveys issued to the business units, followed by interviews with subject matter experts within each business unit, are a good way to proceed. The initial discovery process should also include a "knowledge audit" of source repositories, with an analysis of the type of content and metadata that currently exists.

Next, review the information gathered in this initial discovery process and create an inventory representing each business unit's documents, file plans, and retention schedules. Based on these inventories, a draft version of the structure of the classification can then be created, using terms that describe universally the concepts and categories into which your organization's documents can be classified. Note that these terms should be descriptive enough to be both meaningful and unique.

The draft version should then be submitted to the business units for review, and any discrepancies should be resolved. Users need to be convinced that each type of content they work with has a place in the enterprise taxonomy. Note that multiple review cycles will be required to "rationalize" the taxonomy so that its structure meets the needs of all business units. Part of this process will involve reconciling language issues and terminology. Where consensus cannot be reached, a thesaurus entry can be created as a cross-reference between the terms.

The figure below shows the timeline for a typical taxonomy project. Note that the IT implementation of the taxonomy will extend beyond the timeline shown; also, that the timeline for your own taxonomy initiative may vary, based on the size of your organization and the resources you assign.

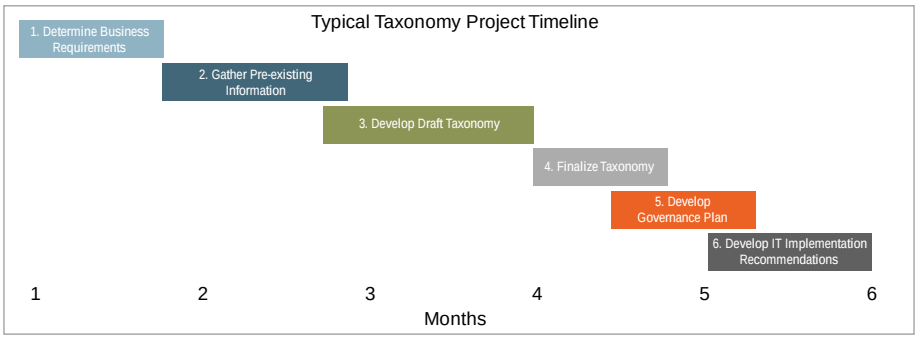

Figure 4: Timeline for an Enterprise Taxonomy Initiative

Factors to Consider

A number of factors can affect the time that will be required to develop an enterprise taxonomy for your organization. Following are the ones that can have an impact on the progress, as well as the success, of a taxonomy initiative:

- **High complexity:** Large, complex organizations with extremely complex business processes should anticipate and plan for slower progress.
- **Level of management support:** Key influencers should be on board and drive the initiative in their own organizations to ensure timely completion.
- **User buy-in:** If representatives of the business units regard the taxonomy effort as unimportant, information quality will be poor. It's critical to have buy-in from all levels that participate in the process.
- **Availability of knowledgeable resources:** Often one or two key people know quite a bit about a department or function, and as a result their time is at a premium. Use a standard set of data collection tools and methods to keep the process efficient.

In general, however, with a core dedicated team supported by interdisciplinary subject matter experts on a part time basis, you should expect the taxonomy creation process to take approximately 3 to 4 months, depending on scope and complexity.

The value of the document inventory

One of the first steps in developing an enterprise taxonomy is to conduct an inventory and assessment of your organization's content assets and repositories (see Question X). Many organizations have never performed such an inventory, and thus have an incomplete understanding of their information assets. The discovery activity that begins the taxonomy development process is a highly effective mechanism for understanding where your organization's information assets lie.

Doculabs' Opinion

In your initial information-gathering, leverage existing taxonomic sources (web page categories, existing metadata, categories in shared drives and Microsoft Exchange) as indications of how the users in different departments and business units now organize their content. Also gather any available taxonomies from external sources (societies and associations, technical indexes, reference libraries), and investigate whether taxonomies offered by your ECM solution provider might meet the needs of your organization. (EMC Documentum and Open Text, for example, provide extensive taxonomies for the pharmaceutical industry, and IBM/FileNet has many vertical taxonomies available that can serve as starting points.) Groups such as the Dublin Core initiative conduct research and recommend best practices for metadata, thesauri, and other techniques for good information management.

Keep in mind, however, that each organization requires its own unique taxonomy. Each organization it has its own history, its own processes, its own organizational configurations, and its own unique core competencies – all of which play a role in how that organization should optimally classify its information assets. The taxonomy discussion can turn into the typical "build-versus-buy" decision, with many organizations choosing to build their own general structure, using components of a packaged taxonomy for specific subject areas only.

5. What is the difference between records management and taxonomy programs?

Discussion

A records management program is concerned with managing the required retention periods of only those documents that the organization has defined as corporate records. In contrast, a taxonomy program is exemplified by a hierarchical classification of all information within the organization, capturing the relationships of the information within the hierarchy. This is particularly true for reference content – i.e. document that are not necessarily business records, but are critical to the day-to-day activities of a business.

A records classification and retention schedule applies specific classifications to business records, defining how long such records are to be maintained. A taxonomy is a structure for organizing documents *and* records; it focuses on the relationships of how information is used within an organization.

The following figure lists out the major differences between a taxonomy and a records plan.

Characteristics	Taxonomy	Records Plan
Scope	Broad – business records, reference materials, correspondence	Narrow – business records strictly defined
Declaration	As deemed by policy or user discretion	As deemed by corporate records policy
Disposition	By policy or user / administrator discretion	As deemed by policy or as required via legal hold
Primary Access Purpose	Search	Discovery
Applicability to Paper or Electronic Materials	Both	Both

Table 1: Taxonomy vs. Records Plan

Factors to Consider

Many people tend to use the terms "file plan," "taxonomy," "records inventory," and "records schedule" interchangeably. But they are different. Briefly, a records inventory is a particular *type* of file plan or taxonomy – specifically, a records inventory is a file plan whose primary purpose is to help create a records schedule (rather than to facilitate search, organize company information to improve efficiency, etc. – the purposes to which an enterprise taxonomy would be put).

If you were to combine your records inventory with a set of defined retention rules, the result would be a records schedule – a schedule that could then be used to manage the retention and disposition of your organization's corporate records.

Leveraging a taxonomy initiative toward establishment of a records management program

If your organization has not yet developed a comprehensive records classification and retention schedule (particularly for your electronic records), gathering the information for an enterprise taxonomy through surveys and interviews at the same time you gather information for your records classification and retention schedule is a very effective approach for completing both tasks in a thorough and comprehensive manner.

However, a taxonomy can serve as the foundation architecture for records management. Having a taxonomy in place allows you to define effective records retention and disposition policies and rules, based on the taxonomic metadata. It also allows you to manage content more effectively from a storage and operational perspective (also known as Information Lifecycle Management).

But a taxonomy that is developed to support a records management program must often go to the document level, which can be three or four levels deep, depending on the strategy (for example, function, process, record series, document). The levels are necessary to capture the information needed to support records retention policies, as the taxonomy will be responsible for including retention information as part of the metadata.

Doculabs' Opinion

One thing to consider is that not all taxonomies support records management – i.e. not all items in a taxonomy are necessarily corporate records. An organization may have an excellent taxonomy for classifying information, but that taxonomy has no relation to the document-level retention plan; there is a disconnect or redundancy that makes it difficult to manage.

Additionally, a taxonomy may have been developed for search and retrieval as well; this may be suitable for navigation, but does not meet the organization's needs from a storage management perspective. As part of the scoping effort for your taxonomy, it will be important to set the goals for the taxonomy up front; keep in mind that creating a good information organization which can be used for navigation, security, records management, and storage management will require additional planning and structure to make sure it meets the needs of all areas. Focusing purely on records management will accomplish your goals from a retention and compliance perspective, but may prove inflexible in other areas.

The most effective approach to implementing an enterprise taxonomy is to use the existing records classification and retention schedule as input for building the taxonomy and as a key element in the content metadata. As ECM solutions are deployed, the indexing and repository structure can be effectively developed using this unification of the taxonomy classification data and records classification and retention data.

6. We already have a records retention schedule. Can't we use that for our enterprise taxonomy?

Discussion

The objectives of a records retention schedule and a taxonomy are different.

The objective of a records retention schedule is to define the lifecycle of a company's corporate records – i.e. how long they should be retained.

The objective of an enterprise taxonomy is to provide a way of classifying all unstructured content so that users can find it and share it, on an enterprise basis.

The key things to remember about a records retention schedule are 1) that it applies only to those documents that are defined as corporate records, and 2) that it is organized by document, within each department or business unit in your organization.

Recall that the objective of an enterprise taxonomy is to provide a classification for *all* documents (not just those that are records), across the entire organization (not just department by department). A records retention schedule cannot show you the cross-connections between documents that the hierarchical structure of a taxonomy is designed to capture. Furthermore, the taxonomy development process will identify duplicate information, as well as document types that should be treated in the same way in the hierarchical structure. The goal of an enterprise taxonomy is consistency: those documents that have the same relationships in the hierarchy occupy the same place in the hierarchy, whereas in a retention schedule, that relationship is generally not made explicit.

Factors to Consider

Recall that one of the first steps in developing a taxonomy is to compile and evaluate the existing document inventories, file plans, and retention schedules (see Question 4). Use your organization's records retention schedule as a part of this discovery process, as it will provide valuable information about the business-critical documents in individual departments, as well as the terminology that each department uses for its documents. Keep in mind, however, that these terms and categories may well be subsumed within the taxonomic structure or in subsequent rationalization process.

Doculabs' Opinion

If your organization has a records retention schedule in place, as well as effective records management policies and procedures, it's likely that your organizational readiness for an enterprise taxonomy will be fairly high. Your users will already be familiar with the concept of enterprise-wide document policies, and the communications and training requirements will probably be lower – first, as you undertake the taxonomy development process; and later, as you implement your enterprise taxonomy across the organization.

Make no mistake, however: effective information management requires both: the records retention schedule, for managing the disposition of corporate records throughout their lifecycles; and the enterprise taxonomy, to manage information that users access in order to do their jobs.

7. Which should come first: implementing a content management system, or developing the taxonomy for organizing content on the system?

Taxonomy and multiple content systems

Note that if your organization has ECM systems in place, the indexing structures of these systems' repositories will need to be accommodated and absorbed into the new enterprise taxonomy.

Discussion

Ideally, you should develop the taxonomy first, as it will be used in a key part of the installation and configuration of the ECM solution: setting up the file and folder structure for the system's repository. This is done manually, folder by subfolder by sub-subfolder, based on the taxonomy you have created for your organization. But in many instances this is impractical, so a certain amount of "retrofitting" will likely be necessary.

Factors to Consider

Today, very few organizations have the luxury of starting an ECM implementation from scratch with a new taxonomy in place and an empty repository awaiting its first document. The reality is that most organizations will have multiple content systems in place already, each with its own embedded indexing schemes, as discussed previously.

If possible, it is advantageous to allocate a new repository and use this as a proof-of-concept area for new taxonomy rollout. This allows you to not only create the new structure, metadata, and other components, but also to validate that the solution can be implemented in the technology solution you have at hand.

To this point, each content management application has its own idiosyncrasies related to the implementation of taxonomy and records management structures. Some, like EMC Documentum, use an inheritance model, while others are more of a flat structure. Considerations such as the treatment of metadata and whether this is carried down from parent to child objects in the object model, and whether fields may be multi-purposed for holding different key values, will determine the exact implementation strategy.

Doculabs' Opinion

Given the realities of technology solutions in most organizations today, it is not reasonable to expect a "green-field" approach to taxonomy development. However, with good management of scope and focus, a contained instance of a new taxonomy can be created as a starting point and then built upon.

Additionally, knowing that the taxonomy is intended to be implemented in an ECM system and not just as a file plan on paper, it is critical to have subject matter experts from IT involved throughout the taxonomy development process. As noted above, there are system restrictions on how you implement taxonomies; you may have to compromise the perfect "academic" structure in order to implement something that will work not just in the repository but also from an interface perspective. Be prepared to discuss and make decisions about these trade-offs as the project progresses. If you find yourself having to compromise too much, it may be time to investigate a new technology solution.

8. Should we use an independent tool, or can we use the taxonomy that comes with our content management system?

Taxonomy specialists provide applications that generate taxonomies, as well as:

- Tools to manage changes in the taxonomy over its lifecycle
- Tools that auto-classify content, store, manage, and share metadata
- Recommend "expert contributors" based on the relationships between people and organizations
- Entity extraction – tracking the number of times a piece of information is successfully found (and accepted) by a person searching for it

Common capabilities of these systems include the following:

- Ability to automatically scan, or "crawl" an existing file system and evaluate its contents, ultimately creating a suggested classification or taxonomical scheme
- If a pre-built taxonomy is used, these systems can import it and then use the results of a crawl to assign content to the taxonomy's nodes
- Benefit of concept extraction – the ability to determine a piece of content's theme, and its relevance to a particular taxonomy

Discussion

Creating an information taxonomy is a large and complex undertaking. Many organizations choose not to create their own taxonomies, but to leverage the work of others in similar organizations or industries. For those that follow this path, there are two options: buying a technology solution that will generate a taxonomy automatically; or using a pre-written, publicly available or for-purchase taxonomy.

As mentioned previously (in the answer to Question 4), your ECM system will include a basic taxonomy. As the demand increases, more and more ECM vendors are starting to offer basic industry-specific taxonomies as part of the package. Additionally, there are sources of industry-standard taxonomies, such as the Federal Enterprise Architecture standards (for entities that work with the U.S. federal government), as well as web sites such as TaxonomyWarehouse that collect sources of information that you can use as the basis for building your organization's taxonomy.

Within the last 3 years, however, there has been an emergence of information taxonomy and classification software – a new breed of software applications designed to help create information taxonomies and address the problem of information organization. These commercial taxonomy and classification software solutions help reduce the labor by automatically examining the existing information and documents in an organization and extracting key concepts from which it generates a taxonomy – which, in theory, is an exact representation of how an organization already "thinks." Vendors such as Convera and Autonomy offer basic taxonomies in a number of industries such as financial services and insurance, which provide low-level details for specific areas.

Factors to Consider

Taxonomy tools fall into three broad categories, each of which presents strengths and weaknesses from a solution-comparison standpoint:

- **Independent taxonomy creation and management tools:** tools that create, manage, and export within an application; used to export to a common XML format for use in other compatible systems
- **Independent taxonomy creation and management tools with connectors to ECM platforms:** tools that create, store, and manage taxonomies externally, but synchronize with the ECM system when necessary (for example, in minor updates, only the changed part of the taxonomy is applied)
- **Taxonomy tools within an ECM system:** tools that are integrated with the administrative interface of an ECM platform; used to manage the hierarchy and organization of content types within the repository

Doculabs' Opinion

Using the tool that comes with your ECM system has the definite advantage of ease: when you apply the taxonomy structure, all of the relevant folders are automatically created for the repository. (The level of automation provided by the solution varies from vendor to vendor, but the process is basically the same.) At implementation time, applying a taxonomy structure is certainly faster than the process of creating a custom repository structure by hand, when that structure is based on your own inhouse-developed taxonomy.

Using a publicly available and free taxonomy for your industry, if such a taxonomy is available, will require extensive efforts (by qualified personnel, such as library scientists) to customize it for your organization.

As stated above, a commercial taxonomy and classification software solution can help reduce the labor of developing your organization's enterprise taxonomy. Keep in mind, however, that these solutions come at a high cost, with best-of-breed taxonomy generation applications costing $250,000 or more.

Ultimately, however, the trade-off between these approaches and creating your own enterprise taxonomy is in how well the taxonomy (and the organizational structure of the repository) corresponds to your organization's unique requirements, as discussed previously (in the answer to Question 4).

9. How many index fields does my organization need?

Discussion

As you begin to define a taxonomy for your organization, look at standards such as Dublin Core for the basic metadata elements for describing cross-domain information resources. The core defines fifteen elements (such as "Creator," "Contributor," "Rights," "Format," etc.), that form the basis of any metadata repository. The key is to target the appropriate level of detail for your organization and for your taxonomy's particular purposes. The question, then, is how far to go.

Involve your IT function in the taxonomy development effort

A taxonomy is a hybrid of how things are stored from a database perspective, with business hierarchy placed upon it.

The risk, however, is being too tied to a structure that a system has forced upon you. Keeping IT involved will help the taxonomy team understand the tradeoffs.

Factors to Consider

The academic perspective advocates defining a full set of metadata and keywords, with all possible unique values, to ensure that there is no ambiguity among document types and classes. Carried to this level, the result can be a seven-level taxonomy – a daunting, multi-month effort, to begin with – but the taxonomy that results will present you with two major challenges: first, your ECM solution will grind to a halt trying to support it; and second, the taxonomy will be nearly impossible to maintain.

The academically "perfect" taxonomy is infeasible to implement and support. Indexing issues are considerable, as is the assignment of metadata for storage purposes. ECM systems are implemented on relational databases, with the unstructured content stored in flat structures, linked from the database which acts as the directory. While many ECM vendors claim their products can support a document hierarchy of "unlimited" levels, the reality is that performance begins to suffer beyond three or possibly four levels.

Metadata and keywords also pose critical trade-offs to content management. More metadata leads to better classification, and thus easier retrieval, but more metadata is harder on the system from an indexing perspective, and also places a heavier burden on the users who will need to assign terms – or requires that you have a smart auto-classifier.

Doculabs' Opinion

Typically, beyond the Dublin Core fifteen elements, anything beyond ten additional terms raises the level of difficulty of management. The difficulty stems not only from the need to enter (or derive) that information for every piece of content, but also to maintain the controlled vocabulary for those items. (Note that some of the security and access components in the Dublin Core may already be "baked" into your ECM tool, and don't need to be redefined.)

The table on the following page shows the typical types of index fields, with examples, and provides ranges of the typical numbers of fields to capture.

Types of Index Fields	Examples	Typical Number of Fields
System-generated	Date, Format, User Logon	2-4
Enterprise-required	Title, Creator, Type	2-4*
Department-required	Subject, Description, Publisher, Rights	3-5*
User-optional	Contributor	2
Total		9-15

Table 2: Index Field Recommendations

The challenge is to combine the goals of the thoroughness of an academic taxonomy against the limitations of a manageable technical solution. Know the limits upfront, and understand the tradeoffs. If four levels is the maximum your ECM system can support, plan ahead and know where to draw the line on your taxonomy effort. Prioritize the value of specific metadata elements, so you can decide which ones to keep and which to handle in a different way. Keeping these things in mind allows you to create a good, implementable taxonomy that your ECM tool will be able to support.

10. What is involved in maintaining a taxonomy, and where should the ownership of the taxonomy reside?

Discussion

Be advised that maintaining a taxonomy can be a challenge, particularly at a large organization.

Like any effort that forces individuals to come to consensus on a common understanding, taxonomies confront different interpretations of reality and make them explicit and tangible. Understand that resolving explicit differences will be an ongoing challenge.

Any single taxonomy will not serve all users equally well, nor will all users agree on a common classification scheme that gets applied consistently over time. Part of the purpose of the governance function will be to develop methods of enforcing the use of the taxonomy – methods that are appropriate to the culture of the organization.

No taxonomy is ever final. Your organization's taxonomy is a living entity, subject to revision and modification as the organization itself grows and undergoes change – for example, as the result of a merger or acquisition, regulatory changes, or changes in the nature of the organization's work. Other events, such as reorganization or departmental consolidation, can also present challenges for an existing taxonomy.

Once your taxonomy is implemented, you should develop policies, procedures, and guidelines to cover the ongoing maintenance, governance, and ownership of the taxonomy, to ensure that it continues to meet the needs of the enterprise.

Factors to Consider

In a large organization, there are typically four roles/bodies that influence a living taxonomy:

- **Taxonomy Governance Committee:** responsible for issuing the final approval on taxonomy changes; the team includes an Enterprise Taxonomy Manager (below), and meets periodically to discuss additions to the taxonomy and to approve, deny, or modify requests as necessary
- **Enterprise Taxonomy Manager:** overall owner of the enterprise taxonomy; responsible for maintaining the integrity of the taxonomy data and structure and for enforcing standards set by the Taxonomy Governance Committee
- **Departmental Subject Matter Experts (SMEs):** experts in a business or functional area who can be consulted for initial taxonomy development and who will receive requests for changes and additions to the taxonomy
- **IT SME/System Administrator:** responsible for consultation concerning technical feasibility of taxonomy governance issues

Once a base taxonomy has been established, the Enterprise Taxonomy Manager must define a specific process for the future evolution of the taxonomy. Doculabs recommends an approach that incorporates the following task areas:

- **Schedules:** The Enterprise Taxonomy Manager publishes a schedule, identifying the schedule for review meetings (monthly, quarterly) and the timelines and procedures for submitting taxonomy change requests.
- **Taxonomy change requests:** Departmental SMEs complete taxonomy change requests as needs dictate and post these to the Enterprise Taxonomy Manager according to the schedule.
- **Review meetings:** The Governance Committee reviews submitted requests and determines the appropriate action.
- **Implementation:** The Enterprise Taxonomy Manager and the IT SME/System Administrator work together to schedule and then implement taxonomy schema changes.

Doculabs' Opinion

Communication is critical for the ongoing evolution and use of the taxonomy. One best practice that Doculabs recommends is setting up a taxonomy portal or intranet web presence. Typical contents of this site would be:

- Links to read-only views of the current taxonomy
- The current meeting schedule and agendas of the Taxonomy Governance Committee
- The Change Request Implementation Schedule
- The Change Request Form and instructions for completing the form

Having this information in an active intranet site is typically more effective than issuing a newsletter or other paper communication, as the information must be current to be valuable. The Taxonomy Manager and Departmental SMEs would be the primary users of this web site.

Additional communication tasks include general education, not only of policies and procedures, but on the subject of how best to develop a taxonomy within the rules and standards set up specifically for your organization. Many firms develop a taxonomy education and style guide to provide a starting point for employees who are new to the process.

Final Word

The average organization is now seeing its volumes of documents, images, reports, web pages, e-mail, and other forms of unstructured content growing at an unprecedented rate. While this electronic information has been with us for two decades, the volume of information is now pushing some organizations to a state of gridlock, from the content- and storage-management perspectives.

Factor in the need for faster and more efficient information retrieval (driven by compliance and records management initiatives as well as by legal discovery concerns), and you begin to see the dimensions of the problem: users are simply unable to find the information they need – and that is costing organizations both money and resources.

What if all of your organization's information assets were stored in a logical structure, within a common hierarchy, with a consistent set of keywords and metadata applied every time, to every piece of content? Retrieving this information would be more efficient, and search results would be more accurate and consistent. In a perfect world, this is exactly what you would do if you had the luxury of building your corporate information model from scratch.

But in the real world, a taxonomy is the next-best thing: it can bring structure to the content your organization has accumulated over the years, and allow your knowledge workers and other users to access the information assets of that content. Implementing a taxonomy can also help you roll out ECM to the enterprise, ensuring that your organization achieves the full business benefits of truly "enterprise" ECM.

Granted, the upfront effort of a taxonomy effort is considerable, requiring commitment and buy-in at all levels of the organization to put in place these enterprise standards and terms. But the benefits of providing consistent access to corporate information are both measurable and significant. As content continues to proliferate, few organizations can afford *not* to have an enterprise taxonomy in place to provide structure to their unstructured content.

Doculabs' Services in Taxonomy Development

Doculabs offers recognized expertise in the creation of enterprise taxonomies, including proven methodologies for taxonomy development and governance.

To learn more about Doculabs' consulting services, please contact us at (312) 433-7793 or at info@doculabs.com.

About Doculabs

DOCULABS

200 West Monroe Street
Suite 2050
Chicago, IL 60606
(312) 433-7793
www.doculabs.com

E-mail Doculabs at:
info@doculabs.com

Doculabs is a consulting firm that helps organizations develop sound technology strategies for content- and process-related applications. Our engagements focus on helping clients leverage their existing enterprise content management (ECM) investments on a broader enterprise basis through objective analysis and in-depth market knowledge. This approach is based on our fundamental belief that in order to protect a client's long-term interest, technology advisors should not be implementers.

Doculabs helps clients deliver on their technology objectives through consulting engagements that address ECM opportunities such as strategic planning, center of excellence creation, taxonomy development, and maturity assessments. Through more than a thousand engagements for organizations facing technology-, compliance-, and process-related challenges, our proven approach has provided our clients the information and advice they need to make confident and well-informed decisions.

Hundreds of leading organizations in the Global 2000 and in state and local government have turned to Doculabs for assistance with their technology strategies.

For more information about Doculabs, visit our web site at www.doculabs.com or call (312) 433-7793.