

Taxonomy: Make or Buy

Susan Cisco, Ph.D., CRM, FAI
Program Manager, Western Region
Iron Mountain Consulting Services
4121 S. Industrial Drive, Suite A
Austin, Texas 78744

Susan.Cisco@ironmountain.com
www.ironmountain.com

Wanda K. Jackson, Ph.D., PMP
President
WHL Information Solutions, Inc.
11152 Westheimer, #743
Houston, Texas 77042
(281) 752-9643
Wanda.Jackson@whlis-inc.com
<http://www.whlis-inc.com/>

More than 90 percent of new business information is created electronically, and 40 percent is never converted to paper (Ingram, 2003). This onslaught of mostly unstructured digital data raises serious issues related to retention, storage, and accessibility. At the same time, organizations are driven by regulators and other stakeholders to enhance transparency, demonstrate accountability, and implement controls.

The higher the level of scrutiny by regulators and other stakeholders, the greater the need organizations have for applying management controls to records creation and storage. For maximum efficiency in retrieving records, organizations must develop and implement taxonomies with associated metadata to complement text searching, provide multiple access points to information, and incorporate retention requirements.

Until recently, taxonomies have been adopted primarily by highly regulated organizations with large volumes of mission critical information. Pharmaceutical companies, for example, must make massive amounts of digital drug information secure yet accessible for legal, business, and regulatory purposes. Without an appropriate taxonomy and properly classified documents, their search engines would return large numbers of documents for users to wade through to find the relevant information.

Organizations can pay dearly for not having better controls and an appropriate taxonomy when they attempt to conduct searches across multiple record repositories during divestitures, regulatory investigations, and digital discovery in response to courts and regulators. To properly apply management controls to electronic records, organizations need to assess the value of taxonomies for at least their at-risk and mission critical records.

A Little History...

Taxonomy originated in the life sciences and can be traced back to Aristotle's theory of categories. He "espoused the idea that things are placed into the same category on the basis of what they have in

common” (Taylor 1999, p174) and are arranged hierarchically with things either inside or outside the container.

Among the earliest applications of classification of knowledge were ten broad categories used by Callimachus, a Greek poet and scholar, for organizing works in the Library of Alexandria. These ten categories remained fairly stable until the late Middle Ages, then expanded significantly in the 19th century with the rapid growth of libraries and an increased demand to provide users with easier access to books. Two large classification systems were developed and put into widespread use to address this need: the Dewey Decimal Classification (DDC) and the Library of Congress Classification (LCC).

By the end of the 19th century, a movement was underway in Europe to go beyond providing access to books and provide access to parts of books and other documents. The Universal Decimal Classification (UDC) system was developed not “as a library classification, but rather as a means to organize and analyze documents” (Taylor 1999, p50).

Whereas the aforementioned classification systems were expressly developed to classify physical objects that existed in physical locations, technological advancements in the 20th century brought an explosion of information (both digital and physical) that forever changed the notion of classifying an “information collection.” Perhaps the greatest contribution to 20th century classification was S. R. Ranganathan’s ideas regarding facets and facet analysis. Although he subscribed to the notion of “universal principles” inherent in knowledge and originally proposed a specific set of facets (personality, matter, energy, space, time), his goal was to describe books using compound subjects or facets. His theory of facet analysis paved the way for the Classification Research Group, and others, to develop flexible classification schemes that are well suited for describing an object using compound subjects and searching for that object using Boolean logic.

21st Century Information Management

A 2004 Delphi Group research report indicated that constantly changing information was the biggest impediment to relocating and retrieving information (Delphi Group, p19). Frequent updates are critical when time sensitive information is required, but they can be very disconcerting to users who find themselves hunting for moving targets.

Compounding the problem of frequently changing information is the increasing volume of hardcopy and digital records that organizations are required to maintain for legal and business purposes. A Reuters study (2000) indicated that “Every day, approximately 20m [20 million] words of technical information are recorded.” While the largest library collection in the world, the Library of Congress, consists of nearly 128 million items (<http://www.loc.gov/about/>), a large organization can easily maintain tens of millions of physical and digital records.

One tool that is instrumental for managing the increasing volume of records is a taxonomy: a structured, often hierarchical, classification system of concepts or subject categories. Taxonomies speed up the process of retrieving records because end users can select from subject categories or concepts. This enables them to narrow the search field and find relevant information rather than relying solely on a blank text search field and their ability to construct an effective query. Taxonomies also provide “serendipitous guidance” by providing additional information that can be inferred by seeing where the concept is placed within the context of the taxonomy (Bruno and Richmond 2003, p45).

An end user who is not knowledgeable about a particular topic might begin the search process by navigating through the taxonomy. When a particular concept is discovered, a text search could be executed against only the records in the particular concept. Conversely, the user might start with a text search that produces hundreds or thousands of records. With the integration of a taxonomy the results could be displayed as a customized set of folders that organize the content by related concepts. Enterprise Content Management (ECM) applications enable taxonomy integration, allowing users to search across

repositories, present records from multiple repositories in response to user queries, and personalize these responses based on the relationship of the requestor to the enterprise (Delphi Group 2004).

Despite the compelling argument for the use of taxonomies for records and information management, more than 70% of organizations that invest in unstructured information-management initiatives do not achieve their target return on investment due to under-investment in taxonomy building (Knox and Logan 2004, p2). In most organizations there is still no way to search for electronic records across multiple repositories except to search each repository separately.

The following sections describe, at a high level, the processes of buying pre-built taxonomies, building taxonomies, and automatically generating taxonomies.

Buying Pre-built Taxonomies

Pre-built taxonomies covering common business functions are available from suppliers such as Data Harmony, Entrieva, Factiva, and Intellisophic. Some taxonomy providers develop pre-built taxonomies geared toward a particular vertical market such as the pharmaceutical, government, and education sectors. ECM vendors also play a significant role in the taxonomy market because their applications use taxonomies to categorize documents for retention and retrieval purposes. ECM vendors often provide proprietary taxonomies to clients implementing their products.

Industry associations are another source of pre-built taxonomies. In the oil and gas industry, for example, the Petrotechnical Open Software Consortium (POSC) and the PPDM Association, along with work done by Shell Expro and Flare Consultants, have produced a taxonomy catalog (POSC 2004). This catalog was developed to provide a logical, standardized way to index and catalog exploration and production information so that it can be easily identified and retrieved in the right context.

There are also worthwhile pre-built taxonomies in the public domain. The Taxonomy Warehouse (<http://www.taxonomywarehouse.com/>) provides a free directory of more than 530 taxonomies, thesauri, classification schemes and other authority files from around the world, plus information about taxonomy references, resources, and events. The taxonomies are classified by 73 subject domains, such as patents, real estate, and taxation, each with ordering instructions.

Pre-built taxonomies usually include a hierarchical structure of topics/categories and a set of rules for implementation. It is important to make sure that the rules provided are compatible with an organization's content application. For example, if the taxonomy is to be implemented in software that maintains multi-level classification, then the taxonomy must allow for multi-level relationships, not flat lists with limited relationships as certain vendors offer (Lederman, p26). Some custom business rules will still need to be defined by the organization buying the taxonomy.

To research the feasibility of buying pre-built taxonomies, the authors queried suppliers for taxonomies covering two specific subject matter areas (Taxation and Architecture/Design) for use in any content application. Several vendors such as Convera, Documentum, and Verity were eliminated because their pre-built taxonomies can be used only with the vendor's own technology. Nearly 30 taxonomies were identified for the two subject matter areas, ranging in price from free to an annual subscription fee of \$69,930. For further details about pre-built taxonomies, refer to the Appendix. Caution must be exercised when reviewing the stated prices since "even giving a price range often misleads customers because each tends to key in on either the low or high figure and it sets an impression that's either unreasonably favorable or unfavorable" (Klein 2005).

Building Taxonomies

In constructing and implementing a taxonomy, the goal is to develop a conceptual organizational structure that can be used to classify and search for information. The general process is roughly the same whether a manual or an automatic approach is used. Four interrelated phases must be considered:

- Phase 1: Planning and analysis
- Phase 2: Design, development, and testing
- Phase 3: Implementation
- Phase 4: Maintenance and change control

Phase 1: Planning and Analysis

Planning and analysis is the first and most critical phase of the taxonomy development process. It requires gaining a thorough understanding of the total information environment in which the taxonomy will be implemented and developing a realistic strategy for integration. The following tasks will ensure successful planning and analysis:

- Assess resources to be involved in the taxonomy project and determine how users plan to use the taxonomy. If necessary, identify outside consulting resources to assist.
- Collect and analyze terms to be used in the taxonomy and decide on the structure of the taxonomy. A good place to start is with all taxonomies and classification schemes already in place at the company (e.g., database structures, networked servers or file systems, file plans, records retention schedules). It is also wise to consult external sources of terms (e.g., pre-built taxonomies, indices, and thesauri related to your industry).
- Select a development strategy and identify appropriate technology to assist in developing the taxonomy and later categorizing content.
- Budget for both development and ongoing maintenance.

Information discovered in the planning phase is used to firm up the project plan and select key milestones to demonstrate project success.

Phase 2: Design, Development, and Testing

Frequent changes to a taxonomy can be expensive and time-consuming after implementation, so taxonomies should be designed for both short-term and long-term needs, anticipating potential organizational structure changes. “People do not like information architecture to change. Spend the time to get it as right as possible [the] first time” (McGovern 2002). Design and development, therefore, should be an iterative process based on feedback from stakeholders at every major stage of the process. Develop a high level structure and test it with a pilot group of stakeholders. Modify the structure based on their feedback and then test again until the general consensus is that the objectives set for the taxonomy are being met.

Phase 3: Implementation

Good planning and design provide a solid foundation for implementing the taxonomy. However, smooth implementation can only be achieved if people, processes, and technologies have been identified and prepared for this phase. The change management process begins early on in the project through open communication and expectation setting. It is formalized with training for stakeholders on the processes to be used around categorizing new information, search and retrieval of information, and the technologies employed in the effort.

Phase 4: Maintenance and Change Control

Even when taxonomy developers consider short-term and long-term needs in the planning process, change is inevitable. A taxonomy is a strategic part of an organization’s information architecture that will require maintenance for many years. It evolves as business needs change, sophistication and

understanding grow around records management, and technology is advanced. Documentation of decisions made throughout the development and implementation process is instrumental for efficiently assessing requests for change and making changes to the taxonomy as necessary. The change management infrastructure (people, processes, and technology) that was implemented in Phase 3 should be maintained for the life of the taxonomy.

Manual and Automated Taxonomy Development

There are two basic strategies to building taxonomies: top-down or bottom-up. A top-down strategy generally uses a manual approach in development. It offers control over the broad general concepts found at the highest levels of a taxonomy and is useful for aligning the taxonomy with business plans and goals. A bottom-up strategy can use automated technologies to extract basic concepts from the content itself and make generalizations on top of these concepts (Ramos and Rasmus 2003, p5). Both strategies have advantages and disadvantages, and both are important for taxonomy development and implementation.

Manual taxonomy development frequently uses a top-down strategy. It offers significant control over the meaning and arrangement of concepts and can be deliberately shaped to reflect common knowledge and practice in an organization. However, manual categorization of documents to the concepts in the taxonomy is low in accuracy simply because of the human judgment involved. The cost of developing and maintaining a manual taxonomy is high because it is a resource intensive process. At the same time, it is a significant task to “train” automated classification tools to categorize documents to the taxonomy. Depending on the tool, it also may be impossible to train it if there is insufficient distinction in the meaning of concepts, or if there are not enough documents (25 to 100) to create a training set for each concept.

Automatic classification tools can automate the process of categorizing content for an already developed taxonomy and/or generate the taxonomy structure itself. Tools that automatically generate the taxonomy apply various algorithms (statistical analysis, Bayesian probability, and clustering) to a corpus of documents in a bottom-up strategy. An automatically generated taxonomy offers little control over the meaning and arrangement of concepts, particularly at the higher levels and, consequently, requires significant refinement in order to make sense to users. These tools can categorize a larger number of documents more accurately and faster than humans, however, the addition of new concepts requires that each concept be trained for automatic classification.

The Make or Buy Decision

Selection of a taxonomy strategy and associated tools should be based on the goals of the taxonomy development project.

Pre-built Taxonomies - Pros and Cons

Pre-built taxonomies can speed up the taxonomy creation process, enabling organizations to deliver immediate results while still allowing for taxonomies to be fine-tuned to organization-specific requirements. Pre-built taxonomies should have already been checked for consistency so, for example, an accounts payable invoice is not called a “bill” in one subject category and a “posting” in another. Furthermore, they incorporate industry best practices and can introduce a more efficient and effective method for organizing records. Even so, it is still necessary to test the concepts in the taxonomy with a subset of the content in your organization to make sure the model is suited to your content and your users.

A significant disadvantage of pre-built taxonomies is that they are not specific to an organization and its objectives, and therefore may have limited applicability. Each organization has its own culture and its own way of categorizing. Using pre-built taxonomies is likely to introduce unfamiliar terminology, making user training more time-consuming. In addition, it is labor intensive to procure a pre-built

taxonomy. It took several days of research to acquire the information in the Appendix. It takes even more time to evaluate each taxonomy and negotiate the purchase price or licensing fee.

Manually Developed Taxonomies - Pros and Cons

Unlike pre-built taxonomies, a manually developed taxonomy can be very specific to an organization, its objectives, and culture. The developer has control over the selection of terminology to make sure it reflects the understanding and needs of an intended audience, as well as the range of content to which it is applied. In some cases, building a taxonomy is the only solution because there are no other existing taxonomies that cover a particular environment.

The primary disadvantage of building a taxonomy from scratch is the time and effort it takes—some of which is hidden because it is distributed across an organization and may include external experts. It is usually faster to use a pre-built taxonomy or to customize one that is compatible in scope and application; however, trying to customize an incompatible taxonomy could be just as time-consuming as building a new one and even more challenging. IBM estimates that it takes 1.5 to 2.5 hours to create each concept (Pohs, Gates, and Doree, 2005) and locate it in a taxonomy. In addition, it takes about 2 hours to review and approve each concept. Obviously, the cost of manual development will depend on the size of the taxonomy.

Automatically Developed Taxonomies – Pros and Cons

The cost of automatically deriving a taxonomy structure is also high because some time-intensive tasks still require human intervention. One may be able to start with automatically generated "concepts," but a person must still examine each concept to see if it meets its purpose and is named appropriately. Human judgment must determine if some concepts should be deleted or new ones added and if the final taxonomy "matches" human understanding and purpose. The time it takes to generate the concept in the first place will depend on the system in use, but the rest of the steps cannot be omitted.

Bear in mind that acquiring a taxonomy, by whatever means, is only part of the picture. All taxonomy structures must be evaluated for suitability to a particular information environment and tested with end users to determine usability. If automatic classification is implemented, data training sets have to be assembled and used to train the classifier, if not, classifiers will have to be trained by more traditional methods. Once implemented, all taxonomies must be maintained.

Best Practices

Just as an awareness of the importance of taxonomies to business strategies has grown over the years, so has agreement over best practices in taxonomy design and development. Based on the authors' experience, the following practices are absolutely necessary for a successful taxonomy development effort, whether you create your own or purchase a pre-built one:

- Make sure the taxonomy is clearly related to business strategy. This provides one standard against which to measure success and help in controlling the scope of work to be done.
- Incorporate existing taxonomy and metadata resources whenever possible. Some resources may be available internally (e.g., departmental taxonomies, records classification schemes from existing file plans, records retention schedules) while others are available externally (e.g., pre-built taxonomies, country codes published by the United Nations Code for Trade and Transport Locations, API well numbers published by the American Petroleum Institute).
- Even in large and complex taxonomies, make sure concepts are well-defined and distinct. If the meanings of concepts are too similar, it is difficult for both people and machines (automatic tagging and categorization) to make distinctions.

- Develop a high level taxonomy, test it with users, expand it, and test again. This iterative technique increases the probability that the right concepts are identified and encourages buy-in from stakeholders.
- Keep the taxonomy as simple as possible. Expand to a useful level but avoid so much detail that information becomes fragmented.
- Provide for adequate resources to maintain the taxonomy. Taxonomies are not static and will change over time. The appropriate change management infrastructure (people, processes, and technology) must be put into place to support necessary change.

Conclusion

Faced with an ever-growing challenge to provide efficient search and retrieval across growing record repositories, organizations are looking for ways to create order out of chaos, and taxonomies are a primary tool. Taxonomies enable proper classification of records and provide end users with the ability to select from standardized concepts and hierarchical structures of information, enabling them to narrow the search field and find relevant information faster. Good planning and design provide a solid foundation for establishing a taxonomy, whether it is pre-built, custom built, or a combination of the two. It is also crucial that impacted people, processes, and technologies are identified and prepared for implementation. Understand that a taxonomy, like a records retention schedule, is a strategic part of an organization's information architecture, and maintenance will require a long-term investment of organizational resources.

References

- Bruno, Denise and Heather Richmond. March/April 2003. The truth about taxonomies. *The Information Management Journal*: 44-46, 48-50, 52-53.
- Delphi Group. June 2004. Information Intelligence: Content Classification and the Enterprise Taxonomy Practice.
- Foskett, A.C. 1996. *The Subject Approach to Information*. London, UK: Facet Publishing.
- Ingram, Brian. September 29, 2003. Locate Smoking Guns Electronically. Internet. Available from http://www.law.com/special/supplement/e_discovery/smoking_gun.shtml; accessed January 7, 2005.
- Klein, Jon R. July 22, 2005 [e-mail]. LexisNexis Taxonomy Solutions. Available email: jon.klein@lexisnexis.com.
- Knox, Rita E. and Debra Logan, September 10, 2003. What Taxonomies Do for the Enterprise. Gartner Research.
- Lederman, Paula. March/April 2005. Implementing a taxonomy solution. *AIIM E-DOC Magazine*: 25-26.
- McGovern, Gerry. October 2002. A step-by-step approach to web classification design. Internet. Available from <http://www.gerrymcgovern.com/la/wcd.pdf>; accessed January 28, 2005.
- Pohs, Wendi, et al. March 3, 2005. Implementing Enterprise Taxonomies with WebSphere Information Integrator OmniFind Edition. Internet. Available from <http://www-128.ibm.com/developerworks/db2/library/techarticle/dm-0503doerre/>; accessed August 10, 2005.
- POSC. Work. Program Summary for 2003. February 3, 2004. Internet. Available from http://www.posc.org/workprgm/summary_2003.shtml; accessed January 7, 2005.
- Ramos, Laura and Daniel Rasmus. January 8, 2003. Best Practices in Taxonomy Development and Management. Giga Information Group.

Reuters Studies. 2000. The Reuters guide to good information strategy. Dow Jones Reuters Business Interactive Limited.

Taylor, Arlene. 1999. The Organization of Information. Englewood, CO: Libraries Unlimited, Inc.

Appendix - Pre-built Taxonomy Vendors for Use in Any Content Application

Company	Product	Product Description	Test Cases: Taxonomies for Taxation and Architecture
<p>Data Harmony www.dataharmony.com</p>	<p>Knowledge Domains</p>	<p>Knowledge domains (include thesauri with rulebases) on more than 40 subjects such as business and finance, education, technology, and pharmaceutical.</p>	<p>Business; 37,874 terms, created for automatic filtering of office-related information (about 34,000 terms are geographical terms). There are tax-related terms in the Business and Finance domains.</p> <p>Available as XML, MARC, Comma Delimited, and Tab Delimited format. Available for license or purchase. Thesaurus is \$5,000. Rulebase is \$10,000.</p>
<p>Entrieva www.entrieva.com</p>	<p>Semio Taxonomy</p>	<p>27 subject matter areas such as high tech, legal, and petroleum; single fee of approximately \$1.50 per node/concept which includes the hierarchical list and related rules; pharmaceutical is the largest (24,000 nodes).</p>	<p>No taxonomies for taxation or architecture.</p>
<p>Factiva (Dow Jones and Reuters company) www.factiva.com</p>	<p>Factiva Intelligent Indexing Taxonomy</p>	<p>Industry, Region, and Subject Taxonomy, includes hierarchical tree-structure, definitions, and alternative names for each term in the taxonomy.</p>	<p>Includes terms for taxation and architecture/design. The entire taxonomy is licensed for annual fee of \$25,000. Updates are available on a quarterly basis. Data are delivered in XML. Business rules are not delivered with the fee. Factiva has a consulting service that builds rules specific to each client’s taxonomy and/or that customizes our rules for the client’s content set.</p>
<p>The Getty www.getty.edu</p>	<p>Art & Architecture Thesaurus</p>	<p>Hierarchical vocabulary of 131,000 terms for 33,700 concepts.</p>	<p>Considered to be a long established and complete resource for architecture. Fixed term license is granted for 5 years. Data updates are offered periodically for an additional license fee. Raw data files are available in three formats:</p>

Company	Product	Product Description	Test Cases: Taxonomies for Taxation and Architecture
			relational tables, XML, and MARC. To obtain pricing, provide 1) Contact information; 2) Description of commercial business or not-for-profit entity; 3) Description of intended use.
Intellisophic www.intellisophic.com	Multiple taxonomies	More than 100 subject matter areas. Includes table of contents, concept tree, and matching rules.	Architecture and design have six subject areas and a subscription cost of \$69,930 for the entire package. Taxation has five subject areas and a subscription cost of \$22,500 for all five. Taxonomies are licensed on an annual subscription basis. The subscription includes all updates, maintenance and technical support throughout the subscription term. Taxonomies delivered in XTM Topic (open XML based format), relational database, or Web Ontology Language.
Lexis-Nexis www.lexisnexis.com	LexisNexis® has subject-specific taxonomies and full industry and subject taxonomies	In June 2005, LexisNexis® launched a program that allows companies to license their proprietary taxonomies.	Terms in the subject and industry taxonomies are related to architecture and taxation. Pricing depends upon the nature of the implementation and the amount of custom terms. Generally, most taxonomies include a one-time implementation fee for engineering consulting and support of initial integration, and yearly licensing fees to cover updates and maintenance. Taxonomies can be purchased with or without the business rules for applying the indexing.
Synapse, the Knowledge Link Corporation, was acquired by Factiva June 2005 www.taxonomywarehouse.com	Taxonomy Warehouse	Directory of more than 530 taxonomies, thesauri, classification schemes and other authority files from over 300 publishers in 40 languages. About 100 of these taxonomies are directly licensable through Taxonomy Warehouse.	Art and Architecture – 12 options; of which 3 are from The Getty (see above). Taxation – 7 options. The Gale Taxation Thesaurus, for example, contains approximately 179 primary terms plus additional non-preferred and related terms, and is priced at \$1,244.05. There are other Gale thesauri related to taxation, which are subsets to the Gale Business Thesaurus. Prices vary based on size and other factors and range from approximately \$300-\$8,300.

Authors

Susan L. Cisco, Ph. D., CRM, FAI

Susan L. Cisco is a Program Manager with Iron Mountain Consulting Services. Her client engagements include retention schedule and vital records program management, electronic records assessments, document imaging assessments, and records management program development. Dr. Cisco is a published author on records management and has educated graduate students, practitioners, researchers, and staff members on records and information management. She holds an M.L.S. and Ph.D. in Library and Information Science from The University of Texas at Austin. Her dissertation is the seminal study of the petroleum industry's use of document imaging systems. She is a member of ARMA International, and in 2000 was named as one of ARMA's Company of Fellow award winners.

Wanda K. Jackson, Ph. D., PMP

Wanda K. Jackson is an IT professional with a variety of project management and knowledge management experience. Her special expertise in taxonomy development and collaborative efforts enabled her to develop enterprise-wide taxonomies for knowledge management and records management initiatives in several large Exploration and Production companies. One of those projects required collaboration with 100 people in seven countries. Dr. Jackson holds a Ph.D. in Library and Information Science from the University of Texas at Austin and is a certified Project Management Professional. Her dissertation focused on understanding information overload of field managers in the oil and gas industry.