



Demystifying Taxonomies

Understanding taxonomies and their
use in content management

Author: Richard W. Erskine

Table of Contents

What's the problem?	3
What is a taxonomy?	3
Taxonomies in other fields.....	3
Business taxonomies.....	4
Taxonomies in content management.....	6
Problems with taxonomies.....	7
Alternative structure.....	8
Pragmatics	9
Conclusions.....	10
Appendix—glossary	11

What's the problem?

The problem is that there is a lot of confusion regarding the use of terms such as taxonomy, thesaurus, ontology, metadata, semantic web, tag clouds, and others. All of these terms are used in varying contexts and are often misused.

The aim of this EMC Perspective is to clarify what a taxonomy is, what it is not, and to provide an explanation that will de-mystify this important subject area, clarifying the terms used.

What is a taxonomy?

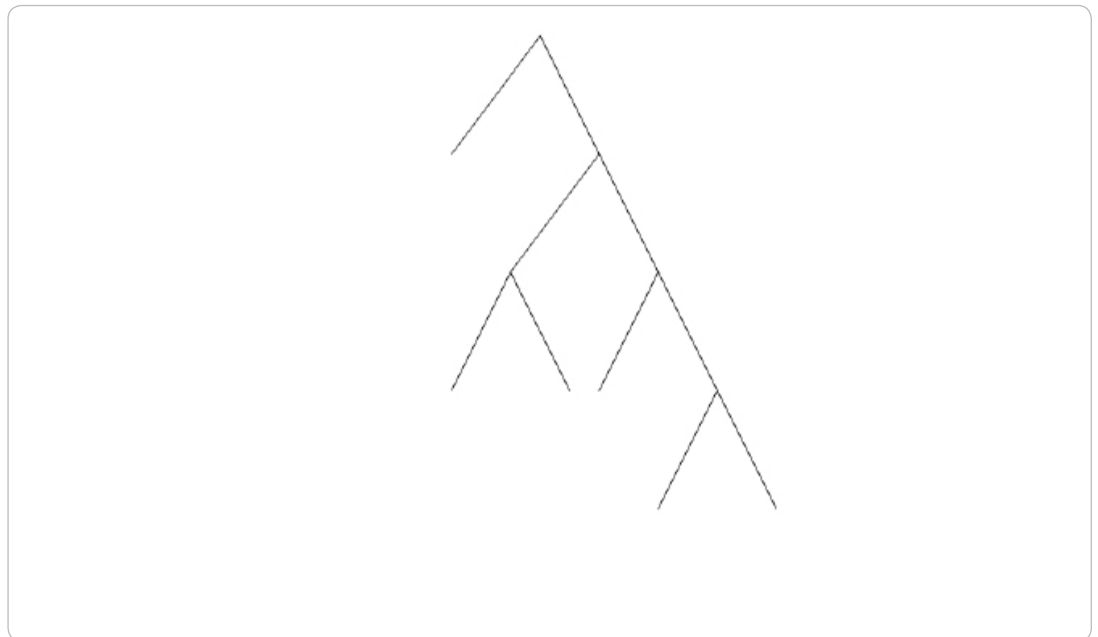
The idea of systematic categorization of concepts in nature and elsewhere is not new. The ancient Greeks, as we might expect, provided some original thinking. See, for example, the *Tree of Porphyry* based on Aristotle's *Categories*. The archetypal use of the term taxonomy is in the field of biology.

In terms of a modern, scientifically based form of biological classification, the term taxonomy is most strongly associated with Carolus Linnaeus (1707-1778) who created a system for uniquely classifying all living things by providing a hierarchical naming scheme. This is the "Linnaean Taxonomy," which usually presents classification levels as *kingdom*, *phylum (or division)*, *class*, *order*, *family*, *genus*, and *species*.

This taxonomy says that a given species belongs to one (and only one) genus; a given genus belongs to one (and only one) family; and so on up the tree, so that each species is uniquely named and classified.

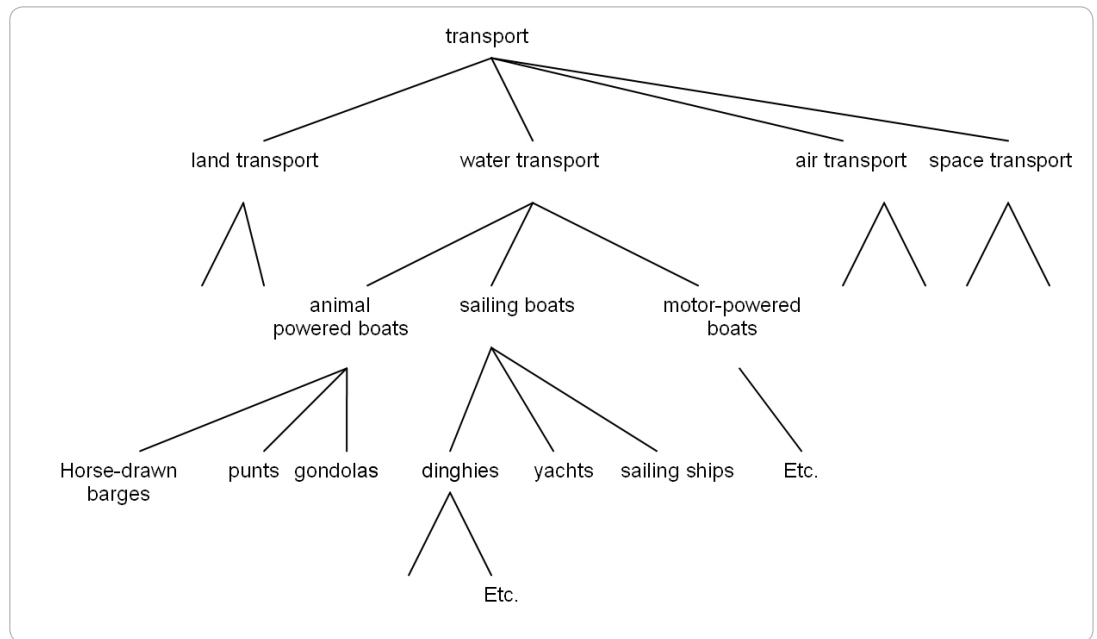
Taxonomies in other fields

To use the term 'taxonomy' for domains other than biology (for example, for transport), one would need a similarly structured way of thinking about classification that ensures that anything within that domain is uniquely and unambiguously positioned in a simple hierarchy as follows:



In a simple hierarchy, each node in the hierarchy can have only one parent. In the transport example, the scheme could emulate the biological one, so that the top level (the equivalent of *kingdom*) would be represented by the main evolutionary groupings: *land transport*, *ships*, *air transport*, and *space transport*. The argument could be made that submarines are quite distinct from *ships* and therefore warrant a separate top level grouping, or that spacecraft are just a specialization of *ships*. In this case, the decision can be made that water transport is a key characteristic of *ships*. Then a decision could be made that the next level of subdivision is based on the nature of the power system (natural systems

such as wind versus man-made systems such as combustion engines). In this way, we might end up with a taxonomy such as this:



As with the biological taxonomy, this taxonomy shows how the classification uses ever more specialized criteria for naming and differentiating things. So, for example, a dinghy does not have a keel; a sailing ship is distinguished more by its sheer size, traditional design, and its function of war or commerce; and the largest yacht can carry tens of people and its function is pleasure or sport. Also, the criteria become less abstract as the hierarchy descends.

Subject matter experts will argue about the correct concepts and leveling to use. There is therefore no single method. Even the taxonomy for classifying species has several variations that biologists have argued about over the years. Nevertheless, the point is that once a taxonomy has been chosen, it must be stuck to and applied consistently—otherwise it could be very confusing.

Any additional indexing can still be applied independently of a taxonomy to reflect other information (such as the genetic information, geographical population, and spread). Therefore a taxonomy may be a primary element in the way we represent the information about the class of things in question, but is normally supplemented by a wide range of additional information.

Business taxonomies

The use of the term 'taxonomy' is increasing in business and by those implementing IT solutions. It is not always obvious how to define or apply taxonomies within a specific business. Often the issue of how to classify information arises within the IT organization; this is where a taxonomy is often first discussed. The analogy with the biological taxonomy is that *species* is replaced by the products that represent the business (for example, drugs in pharmaceuticals) or more abstract entities (for example, business processes). The taxonomy then becomes a means for defining the main characteristics of the subject matter of the business.

However, businesses are often complex and multidisciplinary—one view of the business may be quite different from another or there may be greater or lesser emphasis placed on areas which are familiar or unfamiliar to those in question. For example,

- The drugs created by a pharmaceutical company could be classified according to disease groups, medical indications, treatment regimes, or chemical or biochemical classification.
- A pharmaceutical company can be viewed according to its main business functions: discovery, pre-clinical, clinical, regulatory, manufacturing, sales, and marketing.
- Other viewpoints may reflect the geographical markets; the regulatory environment; the partners and channels to market; and much more.

The right classification scheme is often only definable by the individual, within a particular business context and information need, asking a specific question. For example:

“What is the regulatory status of treatments for Swine Flu in India, and what trials are currently active that will have an impact on that regulatory status?”

Businesses create products or deliver services, and the information needed to answer such questions is embodied in a web of documents and data that describe the physical artifacts of the business (the Drug Dossier; the Clinical Report Form; the Marketing Plan for India; and others). The documents and data are therefore often reflections of the real-world business entities, and can be organized or indexed in a way that reflects the conceptual model of the real-world artifacts.

Sometimes, the documents and data become real-world artifacts in their own right. For example, when a financial company produces financial instruments, such as derivatives that can be invested in, the derivatives only exist as documented agreements and data in systems. In our knowledge economy, there are many such forms of information that have value in their own right, distinct from physical artifacts or assets. In this example of a financial company, a taxonomy can help classify financial products, and include *derivatives* as one important branch of the hierarchy.

In a complex business, there is often a temptation to include all of the criteria needed to classify information within one hierarchy. This is almost always a mistake because it will often lead to a very deep taxonomy that tries to do too much. For example, if in a pharmaceutical company there is one taxonomy dealing with regulations, indications, and markets, the taxonomy would likely be:

- enormous and difficult to use
- possibly biased towards a point of view that is not natural to other disciplines
- error prone because people are unsure how to select an appropriate term, either for indexing information, or for subsequent search and retrieval

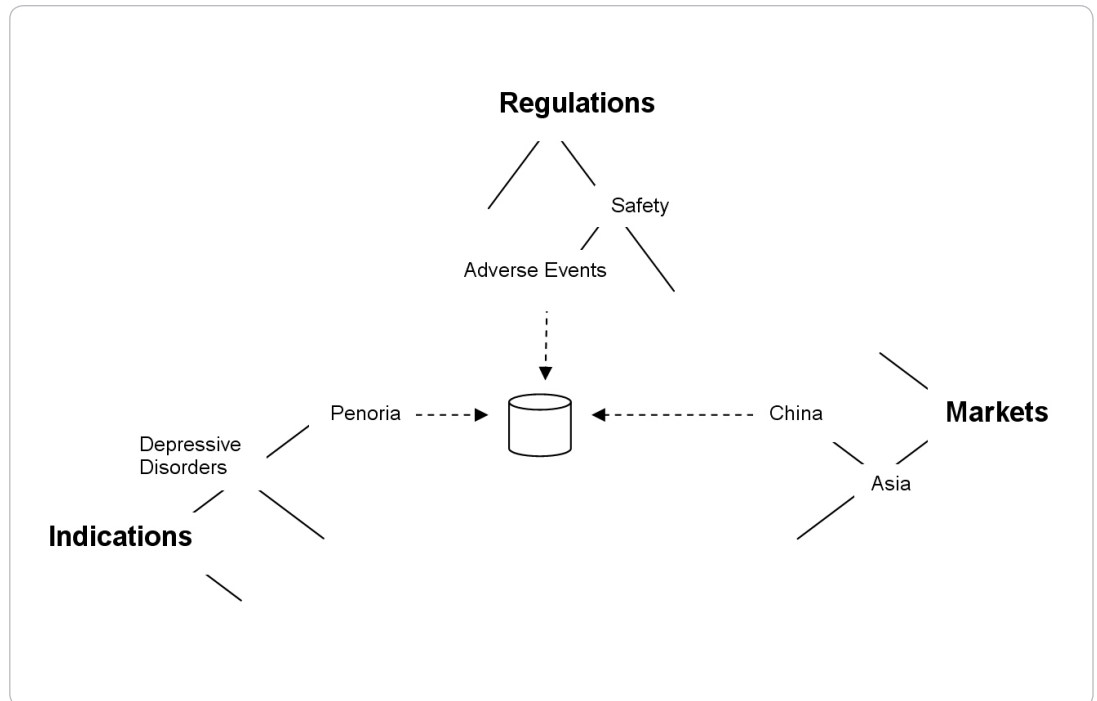
A better practice would be to have separate taxonomies for the distinct subject areas (for example, for regulations, indications, and markets) so that one could navigate to a piece of information via one of more of these taxonomies as appropriate.

This practice is illustrated in the following figure, where a document is classified using a combination of the three taxonomies (*regulations, indications, and markets*), navigated as follows:

Regulations -> *Safety* -> *Adverse Events*

Indications -> *Depressive Disorders* -> *Penoria*

Markets -> *Asia* -> *China*



Taxonomies in content management

Business taxonomies are often used to provide a method of navigating a website, for example. Equally, a taxonomy could be envisaged as the primary folder structure (or file plan) used for organizing documents in a document management system. The word ‘primary’ is crucial here because it is rarely the case that content or documents are classified using one type of attribute (and therefore one hierarchy is rarely sufficient). The primary taxonomy would provide just one dimension for the classification, and generally more are needed in order to efficiently find a document. This is why file systems or file shares are so poor at providing a basis for information management—they often lead to deep, complex, and unusable folder structures.

Consider the transport example described earlier. Additional information, or attributes, can be used to annotate and enrich¹ a taxonomy. These attributes could include *displacement, seating capacity, sail area*, and others. So while a taxonomy may provide the primary means for navigating to documents and content, attributes can help refine the overall classification scheme.

In business there is no single taxonomy that can sensibly be used to organize content hierarchically. In an oil and gas business, for example, there are geologists, economists, lawyers, well engineers, and others, each with their own subject matter expertise (and hence each with their own view of how to classify information). A classification scheme or information architecture should address all needs without being too complex to manage or use.

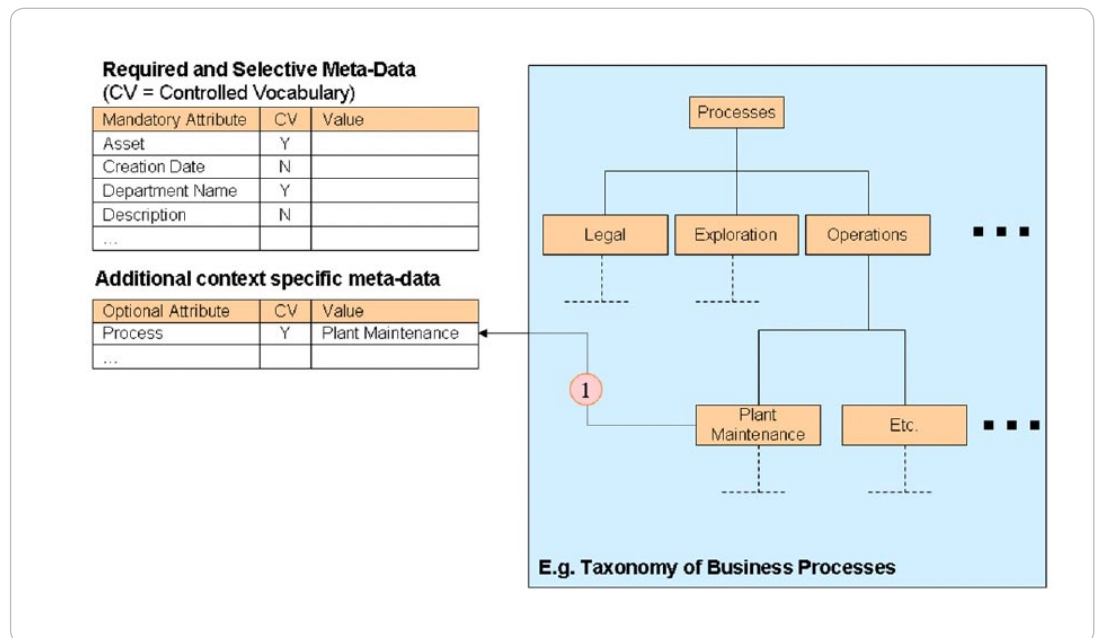
¹ In an enterprise content management system such as EMC Documentum, there are a number of mechanisms for providing enrichment. The folders themselves may have attributes attached to them (and the values may be inherited down through the hierarchy to reduce the manual effort needed); also links between objects can be used to allow multiple inter-locking hierarchies to be created.

Another use of a taxonomy is as a source of terms that can then be used for indexing. In this case, the earlier pharmaceutical example can be seen in a different way. Instead of *regulations*, *indications*, and *markets* being viewed as ways to navigate to a document, they can be seen as three index fields, or attributes, as follows:

Attribute	Value
Regulation	Adverse Events
Indication	Penoria
Market	China

However, when populating the values, the same taxonomies would need to be used as before, but in this case, as navigable vocabularies.

The oil and gas example can illustrate this idea. Consider an index used in oil and gas that contains a number of attributes such as *asset*, *department name*, *process*, and others. Suppose that the *process* value is populated using a process taxonomy that is navigated by a user when indexing a document. This is illustrated in the following diagram:



The taxonomy performs the same role as a simple pick list—to provide a controlled set of values (a controlled vocabulary) when populating attributes. The company could create multiple subject matter taxonomies for other attributes in the same way, each maintained by the subject matter experts. For example, *exploration keyword* could be an attribute, with allowed values such as *field*, *basin*, *granite*, *fault*, and others which have been organized in some structured hierarchy of terms meaningful to geologists and geophysicists.

Problems with taxonomies

There are two problems with taxonomies as illustrated so far in this paper:

- **Confusion between the use of a taxonomy as an organizing principle versus its use as an aid to indexing:** This problem is manageable and often can be handled by simply having a clear understanding of the user interaction and the technical options for implementation.

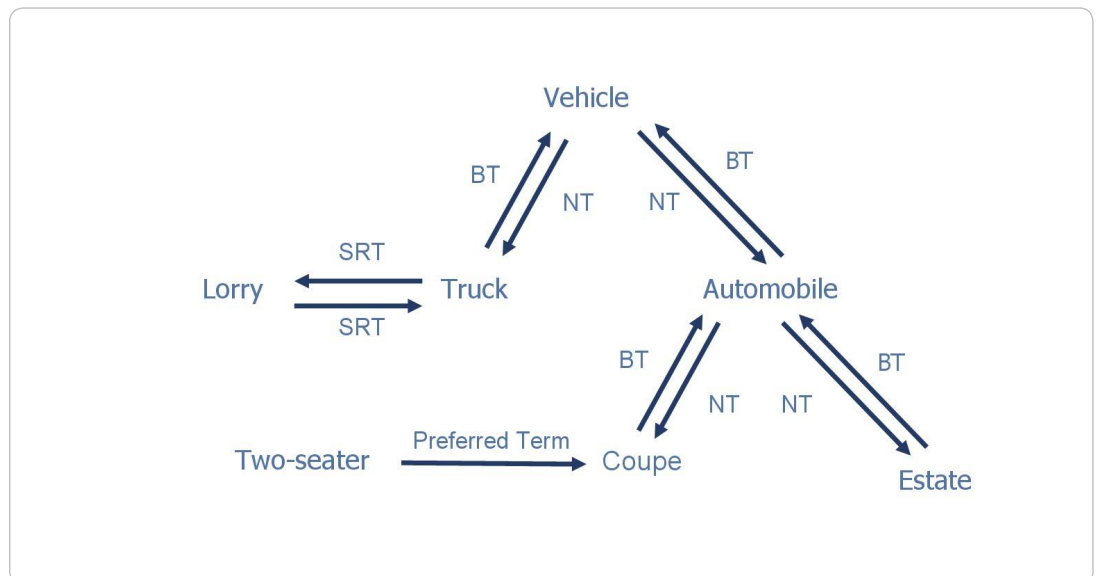
- **When using taxonomies for subject-based key-wording, the terms do not always fit easily within a simple hierarchy (even allowing for the separation into subject domains we have illustrated earlier).** There are two manifestations of this:
 - The terms are often related in different ways—for example, synonyms—and so the simple hierarchy is not the right way to organize the terms. For example, the terms ‘truck’ and ‘lorry’ describe exactly the same concept (they are synonyms). But users need the ability to index using a preferred term while searchers of information have the ability to use either term. These kinds of problems require something more sophisticated than a simple hierarchy of terms.
 - The taxonomy is too deep—the deeper it gets the greater the confusion that arises in where to locate a new term. This is often the result of a poorly structured taxonomy or one that is being overloaded with multiple domains. The solution may be to have a number of distinct orthogonal taxonomies to represent these different subject domains.

Alternative structures

For subject-based terms, used for describing content, it is often better to use an alternative structure to taxonomy.

Thesauri

A thesaurus is focused on structuring descriptive terms using a range of relationships between terms, not just the parent-child one found in taxonomies. We would use ‘Broader Term’ and ‘Narrower Term’ as reciprocal relationships between terms that would have been used in a taxonomy, but then add additional relationships such as ‘Related Term’ (RT), ‘Synonym Related Term’ (SRT), ‘Preferred/ Lead Term’ (LT), ‘Language Equivalent’ (LE), and so forth. Following is a simple example of a thesaurus, using ISO 2788 (a well established standard for describing a Monolingual Thesaurus), describing terms related to ‘vehicle’:



ISO 2788 includes 13 relationships in all and these have been used for many years as a way of structuring terms. Multi-lingual Thesauri are addressed in ISO 5964.

More recently ANSI has published a “Guideline for the Construction, Format, and Management of Monolingual Controlled Vocabularies” ANSI/NISO Z39.19-2005, which enriches the description and use of Thesauri and provides a lot of practical guidance on how to use them. However, even with the addition of such things as qualifiers to help in the management of thesauri, thesauri have limitations when it comes to providing a general model for representing knowledge.

Ontology

Ontology, in the sense used in information management, attempts to go beyond taxonomies in order to represent knowledge in a chosen domain. For advanced information providers there is a need to retrieve relevant information with a very high level of fidelity, and the terms used need to be used in a very precise way. To achieve this high level of fidelity, ontology allows the designer to create customized inter-term relationships.

For example, a thesaurus might identify Formic Acid as a 'Related Term' to Vinegar:

Formic Acid RT Vinegar

Ontology, by contrast, provides a much more precise relationship, very specific to the chemical meaning:

Formic Acid is-an-active-ingredient-of Vinegar

In a food science ontology there would be a number of objects and relationships such as this to help represent our knowledge of food science.

Ontology, therefore, provides a formal way of conceptualizing a domain of knowledge. This is much more than simply a controlled vocabulary to support some indexing system. An ontology attempts to model the semantics used to represent objects and their relationships.

A number of initiatives exist that are supporting the development of ontology for the Web (for example the Web Ontology Language (OWL) is a standard being developed under the auspices for W3C). This kind of thinking has been used in the development of the 'semantic web' concept. Many public domain initiatives exist to develop the use of ontology to connect information sources, and thereby create innovative new services.

Pragmatics

Organizations need a classification scheme (or information architecture) to help organize, index, and then subsequently find documents or other records. Doing this is a basic, but critical, need and often organizations struggle to achieve it.

Returning to the oil and gas exploration example, the need to classify documents is crucial to enable multiple functions across the business to gather and share information to support activities such as discovery, production, and safety management.

One can imagine the following attributes being applied (optionally) to objects such as documents:

- Country
- License
- Block
- Field
- Well
- Discipline
- Document type

But for each of these attributes, a controlled vocabulary can be used, relevant to the attribute in question:

- A simple list may be used (for example, for discipline: "geophysics," "drilling," and others).
- Values can be selected from a hierarchy when, for example, attribute values cannot be represented as a simple list (there may be 10,000 wells to select from). This might be a taxonomy that uses a geographic/geophysical scheme (Country -> Basin -> Field -> Well) for navigation.
- More complex selection mechanisms may be needed (for example, there are conditional relationships between License and Block).

So, for each attribute, an appropriate mechanism is used to enable accurate indexing of information.

This gives rise to the ability to subsequently find documents using a combination of these attributes. For example, searching on Block “XYZ/123,” discipline “Drilling” and document type “Budget,” will meet a clear business need.

In addition to this indexing scheme, there may be a range of file plans appropriate to each main function of the business (geophysics, drilling, HR, and others). Often these plans may reflect organizational concepts or information governance requirements. Despite these being hierarchical file plans, they may not always be best referred to as taxonomies (we will often see the use of the term ‘file plan’ used for such folder structures).

The above mixture of elements² together make up what might be called an ‘information architecture,’ and we should avoid using the term ‘taxonomy’ in this context. In practice, when we implement an information architecture within an electronic system, the following mixture of mechanisms is available: folders, indices, controlled vocabularies, object linking, facets, and others. The end user can then apply a variety of techniques for accessing the information they need: navigating file plans, navigating thesauri or ontology, searching on combinations of attributes, and searching on the content of documents (known as ‘full text retrieval’).

Often there is a desire to use the more sophisticated techniques outlined in the previous sections in situations where a more pragmatic approach will yield results much more quickly. Advanced techniques such as use of ‘thesauri’ and others become more relevant for information mining and knowledge management where there is a greater need to locate documents based on the meaning of their contents. This is much harder to achieve and requires additional effort in indexing documents as well as additional software to support auto indexing and other techniques. It also implies an investment in creating and maintaining these more advanced assets.

For most businesses, it makes sense to start simple before adding the more advanced techniques and to do so when the business case is clear. There are enhanced benefits from using the more advanced capabilities, but also a need for an organization to put in place the right governance, people, and processes to manage, for example, the thesauri it may wish to use.

If the information strategy can identify a roadmap that allows an organization to progressively increase its maturity in relation to the management of controlled vocabularies, then the organization can start simple but also accrue increasing benefits from information and knowledge management over time.

IT will often focus on the technical elements of a plan; but if the effort put into the design of classification schemes (information architectures) is not equivalent to that expended on software architectures and technical architectures, the overall information management objectives are unlikely to be met fully over time. Computer Science must work with Information Science to be truly effective.

Conclusions

Taxonomies are very useful. There are two main applications for them in document/content management:

- As a means for defining the primary hierarchy used for organizing documents/content, which is then used for navigation of folders and/or web sites.
- As a form of controlled vocabulary for indexing documents/content (which can be for one or more attributes).

Understanding the distinction between these two kinds of usage will overcome the more common confusion surrounding taxonomies. The less common confusion arises when the term ‘taxonomy’ is used too loosely to describe any kind of controlled vocabulary or indeed any kind of information architecture.

For example, when a business is using terminology for descriptive subject-matter specific terms (for example, law, geology, transport, and others), simple hierarchies are often not sufficient. Taxonomies may need to be replaced by more advanced structures like thesauri.

² Attributes model for objects; controlled vocabularies for attributes; taxonomies as one form of controlled vocabulary; file plans for organization and governance of documents and content.

Of course, the effort put in must be balanced against the benefits. For an information provider, advanced and efficient information retrieval is essential; but for many organizations this may be too much effort and sophistication³. Often, a combination of using simple indexing (using a number of attributes) and full-text searching on content will provide pragmatic information retrieval without the need to develop complex terminology structures such as thesauri and ontology. The decision on how much effort is put in to design and develop controlled vocabularies must be based on a cost/benefit analysis.

Having said this, the benefits of taking a strategic approach to information management and key assets such as controlled vocabularies can be extremely high and are key to the success of any IT/information management initiative. It is therefore crucial that businesses understand and consistently apply the language of information science—terms such as ‘taxonomy’—when discussing their information management strategies and plans, and not allow themselves to be blinded by jargon.

APPENDIX: Glossary

The following glossary provides a summary of the key terms used in this EMC Perspective.

Term	Description
Attribute	A named field that takes a value and is used to index content. See Classification Scheme. ‘Attribute’ is also referred to by synonyms such as ‘property’ or ‘metadata.’
Classification Scheme	A combination of set of properties used to index content (for example, for a book this might be ‘Title,’ ‘Author,’ ‘Subject,’ and ‘ISBN’). The scheme will include information about each property such as (a) how it is created and maintained, (b) what Controlled Vocabulary is used to assist in populating values, and (c) whether it is mandatory. A useful guide to defining a classification scheme can be found at http://www.getty.edu/research/conducting_research/standards/intrometadata/setting.html
Controlled Vocabulary	Some structure for organizing reference terms in a controlled way, such as a pick list, taxonomy, thesaurus, or ontology.
Ontology	A conceptual model for knowledge, using objects and relationships to model the meaning of terms. An example use of these ideas in the Web is the ‘semantic web’ (http://www.w3.org/2001/sw/).
Metadata	Nicholas Negroponte defines metadata as “data about data.” In respect to content management, see Classification Scheme.
OWL	Web Ontology Language (see http://www.w3.org/TR/owl-features/).
Property	See ‘Attribute.’
Semantic Web	The Semantic Web is an extension of the Web, aimed at enabling innovative applications to be created that connect people and content in new and enriching ways. See http://semanticweb.org . The engineering of the Semantic Web applies and contributes to techniques related to ontology design, such as OWL.
Taxonomy	Originally used for classifying species in biology, but now used more generally. Taxonomy is a simple hierarchy of terms used to classify things (in content management the ‘things’ are documents or other content).
Thesaurus	A set of terms and their relationships, using a fixed set of relationships such as ‘synonym’. Popular form for English terms is Roget’s Thesaurus. Technically defined in terms of various standards, for example, for a monolingual thesaurus there is ISO 2788 or ANSI/NISO Z39-19-2005.

³ However, given the extent to which product providers (for example, in pharmaceuticals, aerospace, and finance), are also information providers, it should not be assumed that their information needs are any less sophisticated than those in the on-line media business.



EMC Corporation
 Hopkinton
 Massachusetts
 01748-9103
 1-508-435-1000
 In North America 1-866-464-7381
www.EMC.com

Take the next step

To learn more, visit www.EMC.com or call 800.607.9546 (outside the U.S.: +1.925.600.5802).