An AIIM Briefing

Helping you manage and use information assets.

How to Develop Taxonomies to Support Navigation, Information Discovery, and Findability

Produced by AIIM Training

By Carl Weise, CRM Industry Advisor

aiim®
The Global Community of Information Professionals
aiim.org | 301.587.8202

Table of Contents

Introduction

Taxonomies are all about organization! With the business trend referred to as "Big Data" reminding us of the terabytes and petabytes of content that organizations have, the idea to organize this content and records for navigation, information discovery, and findability is critical for the success of businesses and government entities. This AIIM Training Brief provides the direction for you to develop the different taxonomies that you need.

The first function of a taxonomy is to help you understand the structure of your knowledge domain at one easy glance. You should be able to look at the structure and be able to predict where you can find out about the parts of your domain in more detail. Predictability turns out to be the most important feature of good taxonomy design. This means it is necessary to understand the natural categorization patterns of your different user communities, and to balance out the ways that they compete or conflict with each other.

Go to www.aiim.org/training to learn of public courses being held in your area and their dates, and please contact training@aiim.org if you have any questions. Course structure, objectives and topics are subject to change without notification.

The Different Forms of Taxonomies

List

This, the simplest form of taxonomy structure, is a collection of items that have some basic relationship to each other, e.g. similarity of attributes or purpose (shopping list, list of files on my shelf), steps in a process, a sequence of actions, activities that are frequently done together, or project roles that you manage. Usually when you look at a list, you should be able to understand the principle of similarity that brings them together.

The main drawback with a list is that it becomes difficult to scan quickly and make sense of with a list over about 12-15 items. The exception is for lists that are very familiar and have entered our long term memory – e.g., lists of countries.

Making lists predictable in structure is sometimes a challenge – for example, a list of job roles might traditionally be sorted by hierarchy, activities might be sorted by their sequence. However, if you are working with a lot of lists, having multiple sorting principles adds an extra burden to your users – they have to figure out what the sorting sequence is before they can predict where to look for what they want. In this case, there is an argument for sorting everything alphabetically, even if this meets some resistance – it makes navigation absolutely consistent and predictable, even if it contradicts the preferred ways of organizing some lists.

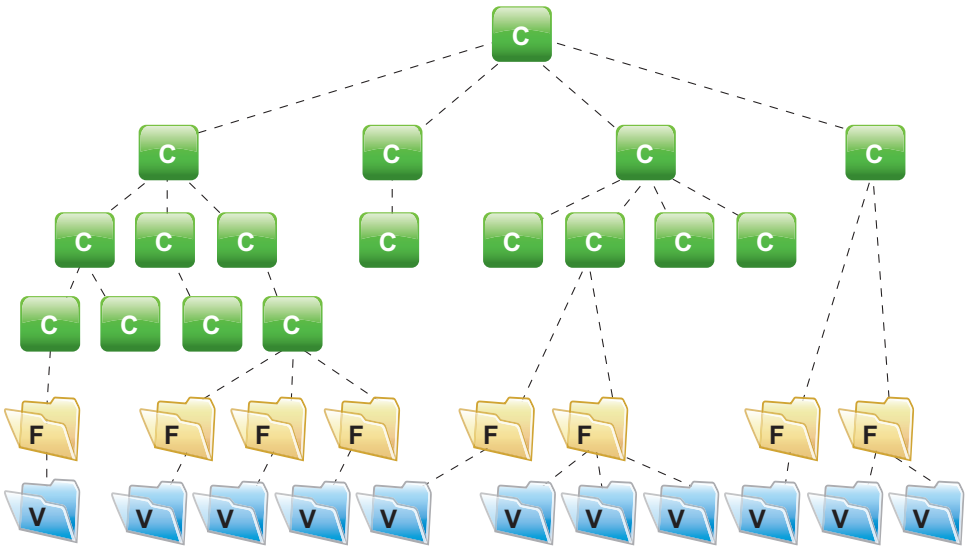
Tree

Once a list gets too long, you usually start breaking it up into clusters and create parent categories for it - creating a tree structure. The tree structure is the structure most traditionally associated with taxonomies, subdivided at the top level and sub-categories underneath – e.g., like in a folder structure on your network share drives.

Most individually-created tree structures are not very easy and predictable to navigate for other people. The reasons are that we are often inconsistent in how we apply principles of subdivision with and between levels. You might sort things into some folders by types of documents but, in others, you might use people or activities, topics, or dates. This lack of consistency is what makes tree structures unpredictable and hard for other people to use.

Hierarchy

A hierarchy is a tree structure that follows very strict rules about how it is subdivided. The same principle of subdivision must be consistently applied at every level. This makes the hierarchy very easy to predict. A true hierarchy should be exhaustive (it contains all possible topics) and categories must be mutually exclusive (there cannot be any ambiguity or overlaps between them. There can be no ambiguity in a true hierarchy. Each topic can only exist in one place in the structure



Records Management Hierarchy

However, in real life, you often want to be able to organize your information according to different principles, and different people have different organization needs. After all, this is why tree structures become so inconsistent. Strict hierarchies are not very practical, but you can learn from how they create predictability. If you reduce the number of ways you organize your content and if you follow consistent rules about how you subdivide the tree, then it becomes easier for different users to find their way around.

Polyhierarchy

A polyhierarchy tries to solve the problem of naturally occurring ambiguities in our information world, by selectively breaking the strict rules of the hierarchy. In a polyhierarchy, a topic can have more than one parent, if there is more than one way that people want to look for it. For example, "pneumonia" might be a topic that could have a parent concept "lungs" for people who are interested in the parts of the body affected and a parent concept "viral illnesses" for those who are interested in causal factors. A "report" could have a parent concept "document types" as well as "activities".

The first function of a taxonomy is to help you understand the structure of your knowledge domain at one easy glance.

The problem with creating too many cross linkages like this is that it starts to break down the consistency and predictability of the hierarchy. The more you break the rules, the less predictable the hierarchy becomes. So, people who work with polyhierarchies use them very selectively with very clear principles for use.

Facets

A facet is a taxonomy structure that takes into account only one attribute of a piece of information. Normally, you would use a system of facets, where each facet describes a different attribute of the information. For example, one facet could list of the document types, another could list all the business activities, and another could list all the job roles, and so on. Each facet can be a simple list, or a tree or a hierarchy.

Too often in taxonomy work, you sort things logically into warehouse type arrangements, instead of understanding what "relatedness" means to our typical users in the course of their everyday work.

Used in combination, facets give a rich description of the content, but each facet also provides a distinct way of organizing and finding the same content. This can overcome the problem of having different competing ways of organizing the same information among different user groups. So people who like to organize by document types can find it that way, others who organize by job roles can find it that way, and so on.

Facets are a clever way of making sure that groups with different organizing principles can each be satisfied. The only downside is that the content has to be tagged multiple times for each relevant facet, and users do not always decompose their searches in terms of different facets.

To be successful, each facet must be completely orthogonal to the others, which means there can be no chance of confusing them. A facet for documents types is completely distinct from a facet for business activities, but a facet for steps in a process might easily be confused with a facet for activities, and so they should not be used in combination. Facets are very popular for providing easy ways to navigate large collections of content.

Matrix

A matrix structure is usually two (sometimes three) facets presented in a table format, so that you can explore the combinations of facet attributes where the facets intersect on the vertical and horizontal axes.

Matrix structures work well when you have a limited number of entities to work with, and they are exhaustively described by the facets when used in combination. One of their benefits is that they can help you spot gaps in your inventory – e.g., when a cell in the matrix is empty.

Grouping "related things" together might seem like a simple thing to do. Good taxonomies make it appear simple.

The limitations of a matrix are that people find them difficult to use when there are large numbers of entities to be described, or when the facets do not consistently combine to describe entities. For example, a facet for customer types would not work well in a matrix with document types because many documents may be internal administrative documents bearing no relation to customers.

System Map

System maps are visual representations of a knowledge domain, where proximity and connections between entities are used to express the relationships between them.

A true hierarchy should be exhaustive (it contains all possible topics) and categories must be mutually exclusive (there cannot be any ambiguity or overlaps between them.

System maps can be descriptive or conceptual. Examples of descriptive maps might be a map of a transport system or a map of the arteries in the human body. Mind maps or concept maps are examples of the more conceptual system maps, as are process maps. All of these maps help to organize concepts and entities, and they are often used to communicate the key nomenclature or vocabulary of a domain.

In a Web-enabled format you can also hyperlink further information resources to the elements in the map.

Choosing which structure or combination of structures you want to use will depend on:

- The number of entities you want to cover in your taxonomy – lists, matrices, and system maps tend to work better at smaller number of entities; trees, hierarchies, and polyhierarchies work in the mid-range, and facets work well for very large numbers.
- The extent to which you want to give a visual representation of a domain that is easy to comprehend and navigate – system maps and matrices do this best, facets don't support this function quite so well, because users have to decompose their queries into the various facet elements.

The feature that is common to all these structures is their ability to give a predictable structure to users to help them navigate and find the information they want.

Classification to Group Related Things

The second function of a taxonomy is to group related things together. This is the classification function of a taxonomy and this is why in Library and Information Science taxonomies are described as classification schemes.

The purpose of a classification scheme is to support resource (information) discovery. If I go looking for applies in a supermarket, I know that I am likely to find other fruit nearby. Similarly, if I go looking for press releases in my knowledge repository, I may be interested in finding speeches and other corporate communications items related to my topic of interest. Grouping "related things" together might seem like a simple thing to do. Good taxonomies make it appear simple.

But, in practice, what seems like a natural relationship to you might not seem so to others. If one person is responsible for safety, they might want the incident reports close to the relevant regulations. If another person is responsible for plant maintenance, they might want the same incident reports close to the relevant documentation and manuals for the machinery affected.

The more you break the rules, the less predictable the hierarchy becomes.

Anthropologists who study the workplace talk about the concept of a "taskonomy" – by this, they mean the arrangement of tools and resources around the most frequent and important tasks they serve. For, example, when a blacksmith is tidying up her workshop at the end of the day, she doesn't usually sort all her tools by type; she lays out the different combinations of tools she's going to use the next day, so they are close at hand when she needs them. Stores often do the same thing. In many furniture stores, you will see the furniture laid out in typical living combinations as they will be used in practice, and it's only in the warehouse that you'll find all the chairs together in one place, all the tables in another section, and so on.

Too often in taxonomy work, you sort things logically into warehouse type arrangements, instead of understanding what "relatedness" means to our typical users in the course of their everyday work. This means you actually force people to go to lots of different places to gather the information they need, instead of finding them all close to each other ready to hand for the work they serve.

Understanding our users and their patterns of information use are the only ways you can overcome this.

Capture the Vocabulary of a Domain

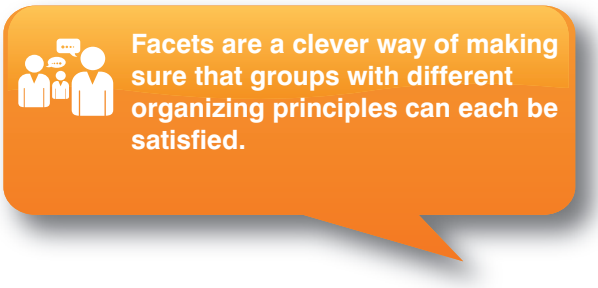
The third major function of a taxonomy is a semantic function, which means that it captures the key vocabulary of your domain. This is why you spend so much time in taxonomy design making sure that you have the terms admitted to the taxonomy correctly so that they can be understood without ambiguity by the different user groups. In fact, you can think of a taxonomy as a kind of dictionary of terms quite apart from its structure. The technical term for this is a "controlled vocabulary".

You define a controlled vocabulary as a collection of terms for which there are job roles, processes, and rules controlling whether or not terms can be admitted to the vocabulary. This means it is somebody's job to either accept or reject terms for inclusion, there are defined processes for doing so, and there are rules or principles governing what should be admitted, and what should be excluded.

The point of controlling a vocabulary is to ensure that it meets the objectives set for it. With no controls, it is very easy for a vocabulary to acquire terms for which there are ambiguities, overlaps in meaning, or duplication of meaning. This clearly has implications for the governance of a taxonomy.

A controlled vocabulary does not need to be a taxonomy. It can be a simple controlled list of terms which is used in your databases or metadata structures. For example, the list of employee names and email addresses in your LDAP server is typically a controlled vocabulary. In many CRM databases, the names of customers are strictly controlled to avoid variant names entering the system and scattering information across several entries. Any field which can have values selected from a predefined set is using a controlled vocabulary.

A taxonomy is just one instance of a controlled vocabulary, and a faceted taxonomy will have a number of controlled vocabularies, one for each facet.



A thesaurus is something more than a simple dictionary or controlled vocabulary. It also contains information about the relationships between terms. When coded into an XML format, a thesaurus can be used by an information system to generate a structural view of the taxonomy, or an alphabetical dictionary view, and it can be used by a search engine to ensure that when related or non-approved alternate terms are used by searchers, they are linked to entries associated with the approved, controlled taxonomy term.

An ontology in information science refers to a data model that has some similarities to a thesaurus. Like a thesaurus, an ontology maps relationships between entities. A basic ontology works on the structure of triples – concept – relationship – concept. However, an ontology is much richer that a thesaurus, which only has three kinds of relationship – broader term, narrower term, and related term. In an ontology you can have as many kinds of relationship as you like: is a part of, is a kind of, is a member of, has the same audience as, has value of, is made of, and so on.

For example, an entry in a list of companies might have a relationship "is supplier to" an entry in a list of projects, and it might also have a relationship "is customer of" an entry in a list of products and services. This gives you great flexibility because you don't have to worry about creating a predictable taxonomy structure that can record only on kind of relationship at a time – you break down the entities into lists and simply track the relationships between them.

But because of this dense interweaving of relationships, the visual structure of an ontology is virtually impossible for humans to navigate. The navigation and overview function of a taxonomy is not served at all by an ontology. This flexibility and richness makes an ontology much more complex than a thesaurus or a taxonomy. It is typically used where you are working with a number of taxonomies and controlled vocabularies across multiple databases and information collections, when you want to be able to get more value out of your data by connecting these vocabularies.

Because of this complexity, ontologies are difficult and expensive to design, negotiate, and maintain. They are virtually impossible to navigate on a human scale, and they typically work best in data processing applications, with designed human presentation interfaces like dashboards or predesigned queries. The effort really only pays off in very large-scale information systems with multiple taxonomies and vocabularies. Here, you want to track meanings and relationships consistently across multiple collections.

Summary

These, then, are the three functions of a taxonomy, i.e., mapping, classification, and description. You can see that a taxonomy should support the three basic needs in your enterprise information system.

- Navigation
- Information Discovery
- Findability

If you are supporting browse and navigate, the structural form you choose is very important.

If you want to support resource (information) discovery, you have to understand and balance the natural perceptions of relatedness among your user communities.

If you want to support findability, you need to collect, understand, and standardize the language in use among your user communities, and assist them in connection related concepts to each other.

You often have to make trade-offs between these functions, so you can't just say you'll do everything. And there is effort involved in building and maintaining the different structures and type of vocabulary.

You now have a good understanding of the different forms of taxonomy, classification to group related things, and the need to capture the vocabulary of your domain. You should have a high-level understanding of how to develop taxonomies to support navigation, information discovery, and findability.

aiim
Taxonomy and Metadata
Training Program

About AIIM's Taxonomy and Metadata Practitioner Program

Develop Your Taxonomy.

Start today!

AIIM offers online and classroom instruction covering the many in-depth topics of Taxonomy. The successful students achieve the Taxonomy Practitioner designation.

Taxonomy^p
aiim practitioner

The Taxonomy Practitioner program covers the strategies, standards, methods, and best practices of developing those taxonomies you absolutely need. The 5 kinds of taxonomies will be thoroughly discussed along with how they should be applied within your organization. An evidence-based form of developing your taxonomies will be presented in detail and the different testing approaches for your taxonomy will be thoroughly examined. An in-depth examination of metadata is provided, including the four key purposes of metadata, and the multiple means of collecting metadata will be extensively discussed. Students will work through four scenarios throughout the course to add practical insights and will be challenged to address their own current circumstances in their own organizations.

This course is designed for:

- Information architects
- Taxonomy and metadata professionals
- Information and records management professionals
- Business analysts
- Managers, project managers and technical staff
- Business unit staff (line management and staff)
- Implementation team - IT and business
- Solution integrators and providers, vendors and their sales staff
- Change agents
- Users

Develop Your Taxonomy and Metadata.
Start Today.

This training supports the AIIM Certified Information Professional (CIP) Certification covering the area of Risk/Liability Focus, in particular, but also the areas of knowledge of Value Focus, Governance Focus and Social Focus.

For more information on the Certified Information Professional Certification, click on:

http://www.aiim.org/certification

For more information on the Taxonomy training, click on:

http://www.aiim.org/Training/Taxonomy-course

Or, email AIIM at: training@aiim.org

About AIIM

AIIM (www.aiim.org) has been an advocate and supporter of information professionals for nearly 70 years. The association's mission is to ensure that information professionals understand the current and future challenges of managing information assets in an era of social, mobile, cloud, and big data. Founded in 1943, AIIM builds on a strong heritage of research and member service. Today, AIIM is a global, non-profit organization that provides independent research, education, and certification programs to information professionals. AIIM represents the entire information management community, with programs and content for practitioners, technology suppliers, integrators, and consultants.

About the Author

Carl Weise joined AIIM in 2006 and is a Program Manager/Industry Advisor. He has contributed to the development of many of the AIIM courses and is a global instructor of these courses.

Carl has over thirty years of senior level records management and project management experience in the financial, IT, manufacturing, electric power, legal, and government environments in both Canada and the United States. He has worked for a records management software provider and worked as a Principal Consultant in Enterprise Content Management (ECM). He is currently providing enterprise content management (ECM), electronic records management (ERM), taxonomy and social media governance (SMG) courses throughout North America and other countries. He has reached over 1,200 students. He is aware of what is happening with records and information management in organizations across North America.

Carl is a Certified Records Manager (CRM) and has given presentations at AIIM and ARMA conferences and chapter meetings. Carl has served on the ARMA Conference Program Committee, including Program Chairperson. He served as a Board of Regent for the ICRM (Vice-President of Exam Administration). Carl developed and taught community college level records management courses and has given many seminars on records management, electronic records management, e-discovery, compliance, risk management, and enterprise content management in cities across the United States and Canada. He has written articles on records management which have been published in North America and Japan.

About this Training Brief

As the non-profit association dedicated to nurturing, growing, and supporting the community of information professionals, AIIM is proud to provide this white paper at no charge. In this way, the entire community can leverage the education, thought leadership, and direction provided by our work. We would like this research to be as widely distributed as possible. Feel free to use this research in presentations and publications with the attribution – "© AIIM 2012, www.aiim.org".

Rather than redistribute a copy of this report to your colleagues, we would prefer that you direct them to www.aiim.org/research for a free download of their own.

aiim®
The Global Community of Information Professionals

AIIM
1100 Wayne Avenue, Suite 1100
Silver Spring, MD 20910
301.587.8202
www.aiim.org