

The State and Value of Taxonomy Standards

Trudy Levy

Image Integration //www.DIG-mar.com

Taxonomy is one of the tools available for managing the information overload that people now face. While all tools may help to retrieve information and provide information about that information, each tool and its standards provide different benefits. To compare taxonomy to another tool with which many may be more familiar, consider metadata structure. There are standards for metadata structure, such as Dublin Core and MARC that specify the most useful data fields with accepted field titles for capturing information about an item. They set up the slots, in which to insert the data. On the other hand, there are no taxonomy standards, but rather standard taxonomies, such as Universal Standard Products and Services Classification, which specify the terms and the hierarchical order to describe an item. Taxonomy provides the data for those slots. Some taxonomy, which is less formal and close to a hierarchical classification system, enables a user to browse: see the big picture and how a piece of information fits within that picture. The formal taxonomies provide an index category system with which authors and editors can avoid duplicating existing documents and recognize areas that need additional content.

What makes taxonomy, how they are implemented and how to benefit is what we will discuss here.

What defines taxonomy?

Taxonomy is a classification system, a concept borrowed from biology. In that model, to classify something is to find the perfect and only place for it in the order of things. Everything has a place. As people further develop taxonomy, some allows information about documents to be stored in multiple places called facets. Facets are alternate aspects of an item, so a book could have a subject heading in a hierarchy but also an author, date, price, and so on, reflecting the human tendency to cross categories. Relative order and the identifying terms define the object in terms of the whole.

Many are already using several taxonomies without being aware of them, such as the Internet category system that Jerry Yang and David Filo developed and is now the standard for Yahoo and many other Internet search directories. This is a classic browse taxonomy form. Its relationships are based on similar items. Those who use certain DAMs (Digital Asset Management) may have created a hierarchical master keyword (or category) list to describe and locate digital assets. This would be another form of taxonomy used for browsing. An example of a more formal taxonomy is a controlled vocabulary or Thesaurus, such as the Getty Research Institute's "Architecture and Art Thesaurus." In this form, the relationships are based on a subset being part or kind of the "parent" set. These classification systems when they are rendered in a controlled manner are taxonomies. In contrast, the hierarchy of folders and alphabetical file names in most of our computers or filing cabinets may be an uncontrolled system.

Taxonomy expert Tom Reamy of the KAPS Group [tomr@kapsgroup.com, //www.kapsgroup.com] further clarifies the two taxonomy forms: "Without getting into arcane definitional disputes, there are two very fundamentally different uses of taxonomy and unfortunately, the most common use is not really a taxonomy at all.

The State and Value of Taxonomy Standards

Trudy Levy

Image Integration //www.DIG-mar.com

The first definition is of a formal taxonomy like that developed by Linnaeus. It is formal in that the relationship between levels is limited to "is a kind of". Zebra's are a kind of mammal, which is a kind of animal, etc. The second sense is the familiar Yahoo style browse taxonomy, which while, it is hierarchical, is based on all sorts of parent-child relationships. A software is not a kind of Computers & Internet to take a Yahoo example

The differences go beyond the structure. To take one example, formal taxonomies are more often used for indexing documents using concepts while browse taxonomies, or classification schemas as they are often called, are used for browsing to sets of documents or web sites based on a huge variety of user based interests and language.

It is possible to talk about standard formal taxonomies like NAICS but there is no hope, nor should there be, of standard browse taxonomies -They depend too much on context and function which is their strength.”]

Taxonomy standards

While there is no such thing as a taxonomy standard, some guidelines do exist that define what constitutes a taxonomy.

Rule #1

Everything has its place.

Taxonomy is a term that was first used by the biologist Linnaeus to organize the animal species in the 18th century. In other words, it is pre-digital technology. Classic taxonomy assumes there is one object and it must have one and only one place. The benefit of this is a unique identifier to every object as Melvyl Dewey discovered with his Dewey Decimal system for cataloging books. As mentioned above, current taxonomies may have additional facets for an object, but the approach of classifying items is still the same.

Rule #2

There must be a hierarchical order of some kind.

The hierarchical order describes the relationship the item has to this subject matter. Taxonomies are usually based on a subject; it may be as general as products, such as UNSPSC (Universal (formerly United Nations) Standard Products and Services Classification) or specific as Medical terms, such as MeSH (Medical Subject Headings), the National Library of Medicine's controlled vocabulary. The relationship may be specific such as in formal taxonomy structure where something is a part of some group or more general, but there is a defined relationship. The hierarchy may be quite flat with only a few “descendents” of the parent group, or it may have several levels of detail. This is where the art or science of taxonomy is seen and where the big picture is created for data users to see. Imagine what big picture would florist or a garden supplier find most meaningful for roses? Two common taxonomies for product classification perceive a rose in the following way:

Taxonomy Example

The State and Value of Taxonomy Standards

Trudy Levy

Image Integration //www.DIG-mar.com

The UNSPSC (Universal (formerly United Nations) Standard Products and Services Classification, [<http://unspsc.org>])

1000000 Live Plant and Animal Material and Accessories and Supplies

1016000 Floriculture and silviculture products

10161600 Floral plants

10161601 Rose plants

10161602 Poinsettias plants

10161603 Orchids plants

10161604 Azaleas plants

10161605 Cactus plants

10161700 Cut flowers

10161701 Cut gladiolus

10161702 Cut lilies

10161703 Cut carnations

10161704 Cut tulips

10161705 Cut roses

10161707 Cut flower arrangement

People using the UNSPC taxonomy would perceive a rose as a Floriculture product. By the way, floriculture is “the cultivation and management of ornamental and flowering plants” [Webster’s seventh New Collegiate Dictionary], so they would learn quite a bit about a rose, but not as much as they could learn in the NAICS taxonomy, which has a deeper hierarchy.

The NAICS (North American Industry Classification System, developed by the US Census Bureau, [<http://www.census.gov/epcd/naics02/>])

11 Agriculture, Forestry, Fishing and Hunting

111 Crop Production

1114 Greenhouse, Nursery, and Floriculture Production □

111421 Nursery and Tree Production

111422 Floriculture Production

 Bedding plant growing (except vegetable and melon bedding plants)

 Cultivated florist greens growing

 Cut flower growing

 Cut rose growing

The State and Value of Taxonomy Standards

Trudy Levy

Image Integration //www.DIG-mar.com

Now those using NAICS taxonomy would also learn that Floriculture is considered part of Nursery and Tree production, which is a subset of Greenhouse production, etc. With the deeper hierarchy, they have provided a much clearer picture. However, the UNSPC has roses as both cut and bush, but not so the NAICS. A hierarchy should be deep and thorough which brings us to rule three.

However, before we leave rule two, please note: some systems only provide searchers with the unique identifier from the taxonomy, thus robbing them of this Big Picture. While using the unique identifiers that taxonomy generates works very well in computer encoding, it is a waste to only reveal the information available in its structure to the machine.

Also, be aware that the classification order represents someone or some group's perception of the object. While these two agencies have tried to make a taxonomy that is generic and fits all purposes, the mere choice of a word to name an object influences one's perception. A rose is not a rose by any other name. Culture and common practices will influence the structure and term choices of a classification system, whether it is a world, country, industry or a business. This can be the power and the weakness of taxonomy. We gravitate to the government and professional organizations' taxonomies, because they were developed to fit many people's needs. More specialized taxonomy will only work if it accurately reflects the culture of all its users.

Rule #3

It must be flexible and constantly updated

Taxonomies may be tools from our past, but they definitely are not written in stone. Attitudes change and new objects appear on the scene. All groups are constantly reviewing their taxonomy and debating the questions: how to fit each bit into the whole and how to add new bits. Therefore, the structure must allow for this change. Many groups simply leave "to be allocated" levels in their classification schemes to accommodate new material. No one wants to start over with new unique identifiers, although it has been done. NAICS was developed to replace an older version called SIC (Standard Industry Classification), whose unique identifiers were too limited for today's technological needs.

As new concepts, vocabulary, and products come into use, the taxonomy must reflect these changes. The simplest case is when one can add an alternate term: moving from "horseless carriage" to "automobile" to "car". In other cases, this may require adding a new subcategory, splitting a category or rearranging a whole section. In summary, a standard taxonomy will be is a hierarchical classification system that reflects the subjective perceptions of its authors, generates a unique identifier and is constantly evolving. It is a tool in managing information that is very simple but powerful.

The State and Value of Taxonomy Standards

Trudy Levy

Image Integration //www.DIG-mar.com

Standard taxonomies: implementing taxonomy rules

A standard taxonomy is one that is widely accepted as an authoritative classification system for a subject area. Because they have the authority, government agencies and professional groups are the prime source of standard taxonomies. Some quite old taxonomies are still in use, such as those developed by Linnaeus and the Dewey Decimal System. This is because they are widely accepted and fit people's needs.

Another such standard taxonomy is SIC [Standard Industrial Classification][<http://www.osha.gov/pls/imis/sicsearch.html>]. While it has been replaced by NAICS and has not been upgraded since 1987, it can still be found on such business information sites as Hoovers. The reason is partly inertia. The change of unique identifiers was extreme. It also is because the terms used to describe the hierarchical order are more familiar. There is no need for a dictionary to understand the big picture. Our rose in a SIC search reveals the following:

Division A: Agriculture, Forestry, And Fishing

Major Group 01: Agricultural Production Crops

Industry Group 018: Horticultural Specialties

0181 Ornamental Floriculture and Nursery Products

Establishments primarily engaged in the production of ornamental plants and other nursery products, such as bulbs, florists' greens, flowers, shrubbery, flower and vegetable seeds and plants, and sod. These products may be grown under cover (greenhouse, frame, cloth house, lath house) or outdoors.

Bedding plants, growing of

Bulbs, growing of

Field nurseries: growing of flowers and shrubbery, except forest

Florists' greens, cultivated growing of

Flowers, growing of

Foliage, growing of

Fruit stocks, growing of

Greenhouses for floral products

Mats, preseeded: soil erosion-growing of

Nursery stock, growing of

Plants, ornamental: growing of

Plants, potted: growing of

Rose growers

The State and Value of Taxonomy Standards

Trudy Levy

Image Integration //www.DIG-mar.com

NAICS is a better cataloging tool. It is more precise, allows for greater depth in accessing the correct item, and has greater flexibility. SIC will probably fade away, but it is a more comfortable tool, which is probably why it is still used so actively, especially by students doing their market research projects.

Of course both NAICS and SIC are broad classification systems meant to manage a wide variety of information for a broad audience. When an industry has specialized needs or culture, they have had to develop their own taxonomies, such as MeSH (Medical Subject Headings), the National Library of Medicine's controlled vocabulary and the MasterFormat, which the Construction Specification Institute maintains. CSI is a professional organization of primarily specification writers and construction product vendors.

The CSI taxonomy, MasterFormat is a great example of how a good taxonomy can be adopted as the foundation of industry processes. All trades and professions involved in the construction sector are familiar with it, even if they may not know who the authoring agency is. In fact, many, especially those new to the field think McGraw-Hill is the authoring agency, because it uses it as the index of Sweets catalog, its construction product catalog. It is entirely worthless to anyone else, but those involved in construction use it extensively in all aspects of their business, from contracts to filing systems.

Both the order and the unique identifiers are fundamental in communication between the different participants: architects, engineers, contractors and vendors. Also, while it appears to be primarily a classification system of building products, it actually establishes a clear perception of the building process, from ground up as it were. Whether it reflects the construction sector's view of the process or created that view is hard to tell, but it is intrinsic to that view. How does a rose fit into this picture?

In the section: Construction Products and Activities, our rose would be in Division. 2

Division 2 — Site Construction

02050 Basic Site Materials and Methods

02100 Site Remediation

02200 Site Preparation

02300 Earthwork

02400 Tunneling, Boring, and Jacking

02450 Foundation and Load-bearing Elements

02500 Utility Services

02600 Drainage and Containment

02700 Bases, Ballasts, Pavements, and Appurtenances

02800 Site Improvements and Amenities

02900 Planting

02950 Site Restoration and Rehabilitation

The State and Value of Taxonomy Standards

Trudy Levy

Image Integration //www.DIG-mar.com

Last, but not least of the currently accepted standard taxonomies is the basic classification scheme after Dewey's Decimal, the Library of Congress's Classification System [<http://www.loc.gov/catdir/cpsolcco/lcco.html>] Granted this is designed for classifying printed material, but attempts universal coverage, defining a unique location for each work while grouping by the main topic. The LCCS has a two-level letter hierarchy and number ranges for deeper categories. There are facets such as region and country that can be applied wherever appropriate. It also leaves space for future growth and as detailed a hierarchy, as needed. For example, in the LCCS, we do not find specifically roses, but there are several classifications that we could use. If we stay with agriculture, we might use the following:

Class S Agriculture

Subclass SB – Plant Culture

SB412- 439.8 Classes of plants

Or

SB442.8-443.4 Marketing. Cut flower industry. Florists

One can sink your teeth into this taxonomy.

However, because there is one and only one place a book can be on library shelves, users can be fooled into thinking that all the works on a topic are grouped together. Looking for books on roses shows the problem with this: some rose books are in the agriculture section, some in art, some in home decor, some in perfume, and so on.

The Library of Congress Subject Headings (LCSH), [<http://lweb.loc.gov/cds/lcsh.html>] is a comprehensive list of topics. It is the source for definitive subject headings in card catalogs and library computer systems. When it was all cards, librarians could only assign one or two headings, or the card catalogs would run out of space. Now they can and do assign three or more subject headings, attempting to define the various topics covered by a book. It is relatively shallow and has facets such as "charts", "correspondence" and "case studies" which can differentiate within the broad categories.

How to benefit from standards.

The rules of standard taxonomy: order by relationships, a place for everything, and flexibility, apply to many situations. Even without creating the taxonomy, organizing material into classifications at whatever level increases awareness of the big picture and how each part fits. It is a good method of approaching any chaotic information; especially as these days, little information is going to stay in individual files. Most information is being distributed to everybody and may well get lost.

"Taxonomies and full-text search are like peanut butter and jelly: they just go together. Search is more flexible, more ad-hoc, more dynamic. It's easy to get interesting synergies when searching, and to find items using new terms and jargon. Taxonomy is more reliable and standard, expresses the information architecture, providing context and

The State and Value of Taxonomy Standards

Trudy Levy

Image Integration //www.DIG-mar.com

grouping even when the words used are different. Taxonomy makes search results more accurate and understandable; search provides alternate access, uses different vocabulary and crosses boundaries.” Avi Rappoport, Search Engine Consultant
<mailto:avirr@searchtools.com> Complete Guide to Search Engines for Web Sites and Intranets<http://www.searchtools.com>

Since there are not many Melvyl Deweys in every organization, an easy step is to use a widely accepted taxonomy, which will allow information to merge into the larger pool of information without much effort. The size of that pool and its diversity will depend on the organization, but remember even a single business includes many diverse users, corporate, accounting, promotion, product development and production. A standard taxonomy facilitates information exchange among a diverse group.

If there does not exist a standard taxonomy that meets an organization’s needs perfectly, then a standard taxonomy can be customized to fit an organization’s needs. It may be simply a procedure to translate the taxonomy into a standard’s terms. For example, if the term floriculture is not suitable, substitute “growing ornamental flowers” but maintain the unique identifier. There are many consultants and software developers, eager to help an organization create the perfect standard for its purposes. Look for those who are basing their work on an accepted standard taxonomy. No organization is an island, especially in this day where mergers are happening in such multitudes. A standard taxonomy can integrate the merged information more smoothly.

In addition, after a taxonomy has been adopted, there are software programs that can then “read” unmanaged (unstructured) information and by virtue of word usage etc, organize it according to that taxonomy to whatever level is desired.

Therefore, taxonomy is a useful tool for managing information. Librarians and Information Scientists are constantly refining and defining taxonomy structures and terms to better serve everyone’s needs, but a standard taxonomy, by being widely adopted, will get material in better shape to usefully merge with the other data on the information highway.