Chapter 1

What Are Taxonomies?

Taxonomies? That's classified information.
—Jordan Cassel

The first step in discussing the role and work of the taxonomist is to clarify what a taxonomy is. Even if you already have some understanding of the concept, there are multiple meanings and various types of taxonomies that require further explanation. The descriptions provided here are not strict definitions, and the range of knowledge organization systems should be thought of as a spectrum.

Definitions and Types of Taxonomies

The word *taxonomy* comes from the Greek *taxis*, meaning arrangement or order, and *nomos*, meaning law or science. For present-day information management, the term *taxonomy* is used both in the narrow sense, to mean a hierarchical classification or categorization system, and in the broad sense, in reference to any means of organizing concepts of knowledge. Some professionals do not even like to use the term, contending that it is too often ambiguous and frequently misused. Yet it has gained sufficient popularity, and a practical alternative term does not seem to exist. In this book, taxonomy will be used in its broader meaning and not limited to hierarchical structures.

In the broader sense, a taxonomy may also be referred to as a *knowledge organization system* or *knowledge organization structure*. This designation sometimes appears in scholarly discussion of

the field and in course titles at graduate schools of library and information science. The designation *knowledge organization system* was first used by the Networked Knowledge Organization Systems Working Group at its initial meeting at the Association for Computing Machinery Digital Libraries Conference in Pittsburgh, Pennsylvania, in 1998. Gail Hodge further expanded on it in an article in 2000 for the Digital Library Federation Council on Library and Information Resources. In Hodge's words:

> The term *knowledge organization systems* is intended to encompass all types of schemes for organizing information and promoting knowledge management. Knowledge organization systems include classification schemes that organize materials at a general level (such as books on a shelf), subject headings that provide more detailed access, and authority files that control variant versions of key information (such as geographic names and personal names). They also include less-traditional schemes, such as semantic networks and ontologies.[1]

Although she does not mention taxonomies per se in this paragraph, Hodge goes on to list the various types of knowledge organization systems, which include[2]:

1. Term lists (authority files, glossaries, dictionaries, and gazetteers)
2. Classifications and categories (subject headings, classification schemes, taxonomies, and categorization schemes)
3. Relationship lists (thesauri, semantic networks, and ontologies)

Needless to say, the designation *knowledge organization system* has not caught on in the business world and is not likely to do so. We are even less likely to hear of a *knowledge organization system creator/editor*; that would be a good description of a taxonomist.

While this book uses the term taxonomy broadly (as a synonym for knowledge organization system), most of our discussion focuses on taxonomies that have at least some form of structure or relationship among the terms (types 2 and 3 in Hodge's list) rather than mere term lists. Indeed, people do not usually call a simple term list a taxonomy. Let us turn now to definitions and explanations of some of these different kinds of knowledge organization systems or taxonomies.

Controlled Vocabularies

The term *controlled vocabulary* may cover any kind of knowledge organization system, with the possible exclusion of highly structured semantic networks or ontologies. At a minimum, a controlled vocabulary is simply a restricted list of words or terms for some specialized purpose, usually for indexing, labeling, or categorizing. It is "controlled" because only terms from the list may be used for the subject area covered. If used by more than one person, it is also controlled in the sense that there is control over who may add terms to the list and when and how they may do it. The list may grow, but only under defined policies.

The objective of a controlled vocabulary is to ensure consistency in the application of index terms, tags, or labels to avoid ambiguity and the overlooking of information if the "wrong" search term is used. When implemented in search or browse systems, the controlled vocabulary can help guide the user to where the desired information is. While controlled vocabularies are most often used in indexing or tagging, they are also used in technical writing to ensure the use of consistent language. This latter task of writing or creating content is not, however, part of *organizing* information.

Because controlled vocabulary has this broader usage when applied to content creation, not merely information organization, the term *controlled vocabulary* should not be used as a synonym for knowledge organization system.

Most controlled vocabularies feature a *See or Use* type of cross-reference system, directing the user from one or more "nonpreferred" terms to the designated "preferred" term. Only if a controlled vocabulary is very small and easily browsed, as on a single page, might such cross-referencing be unnecessary.

In certain controlled vocabularies, there could be a set of synonyms for each concept, with none of them designated as the preferred term (akin to having equivalent double posts in a back-of-the-book index instead of *See* references). This type of arrangement is known as a *synonym ring* or a *synset* because all synonyms are equal and can be expressed in a circular ring of interrelationships. An example of a synonym ring, as illustrated in Figure 1.1, is the series of terms applications, software, computer programs, tools. Synonym rings may be used when the browsable list of terms or entries is not displayed to the user and when the user merely accesses the terms via a search box. If the synonyms are used behind the scenes with a search engine and never displayed as a browsable list for the user, the distinction between preferred and nonpreferred terms is thus moot. Though these types of controlled vocabularies are quite common, they are often invisible to the user, so the terminology (synonym ring and synset) is not widely known.

Sometimes controlled vocabularies are referred to as *authority files*, especially if they contain just named entities. Named entities are proper-noun terms, such as specific person names, place names, company names, organization names, product names, and names of published works. These also require control for consistent formats, use of abbreviations, spelling, and so forth.

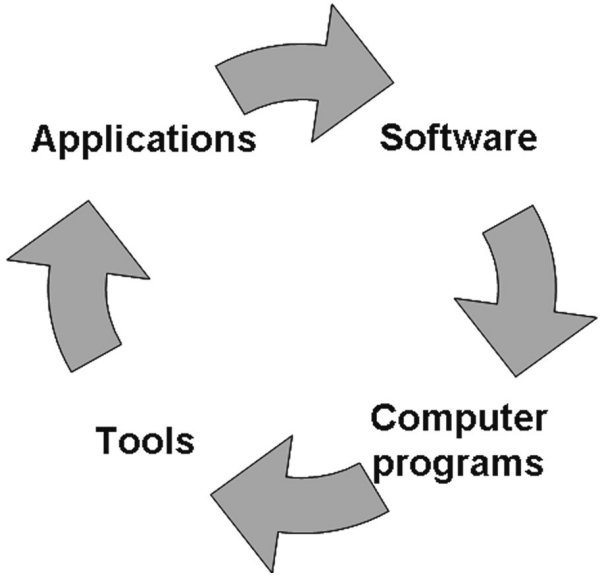


Figure 1.1 Example of terms in a synonym ring

Controlled vocabularies may or may not have relationships among their terms. Simple controlled vocabularies, such as a temporary offline list created by an indexer to ensure consistent indexing or a synonym ring used behind the scenes in a search, do not have any structured relationships other than preferred and nonpreferred terms. Other controlled vocabularies may have broader/narrower and related-term relationships and still be called controlled vocabularies rather than thesauri or taxonomies. This is often the case at periodical and reference index publishers, such as Gale, EBSCO, and H.W. Wilson, which maintain controlled vocabularies for use in their periodical indexes. In some cases, the publisher maintains multiple kinds of controlled vocabularies, some being more structured than others, and controlled vocabulary is the more generic designation for all of these.

Hierarchical Taxonomies

When we think of taxonomy, hierarchical classification systems are what typically come to mind. However, as explained in the previous section, we are using a broader definition of taxonomy that encompasses all kinds of knowledge organization systems. So taxonomies that are structured as hierarchies we will refer to specifically as *hierarchical taxonomies*.

A hierarchical taxonomy is a kind of controlled vocabulary in which each term is connected to a designated broader term (unless it is the top-level term) and one or more narrower terms (unless it is a bottom level term), and all the terms are organized into a single large hierarchical structure. Taxonomy in this case could apply to a single hierarchy or a limited set of hierarchies. This type of structure is often referred to as a *tree*, with a trunk, main branches, and more and more smaller branches off the main branches. Actually, if the taxonomy is displayed as a tree, it is an upside-down tree, with multiple smaller branches for narrower terms lower down on the page or screen. Another way to describe such structure is a taxonomy with *nested categories*. The expression *to drill down* is often used to describe how a user navigates down through the branches. An example of an excerpt from a hierarchical taxonomy appears in Figure 1.2.

The classic example of a hierarchical taxonomy is the Linnaean taxonomy (named after Carolus Linnaeus) of biological organisms, with the hierarchical top-down structure: kingdom, phylum, class, order, family, genus, and species. The Dewey Decimal Classification system for cataloging books can also be considered a hierarchical taxonomy (although, like the Linnaean taxonomy, it has the drawback that each item can be classified in only one place). Other well-known examples of hierarchical taxonomies are the Standard Industrial Classification (SIC) and North American Industrial Classification Systems (NAICS) codes for classifying industries. Hierarchical taxonomies are also common

Top Level Headings
Business and industry
Economics and finance
Education and skills
Employment, jobs, and careers
Environment
Government, politics, and public administration
Health, well-being, and care
Housing
Information and communication
International affairs and defence
Leisure and culture
Life in the community
People and organisations
Public order, justice, and rights
Science, technology, and innovation
Transport and infrastructure
Leisure and culture

Leisure and culture
. Arts and entertainment venues
. . Museums and galleries
. Children's activities
. Culture and creativity
. . Architecture
. . Crafts
. . Heritage
. . Literature
. . Music
. . Performing arts
. . Visual arts
. Entertainment and events
. Gambling and lotteries
. Hobbies and interests
. Parks and gardens
. Sports and recreation
. . Team sports
. . . Cricket
. . . Football
. . . Rugby
. . Water sports
. . Winter sports
. Sports and recreation facilities
. Tourism
. . Passports and visas
. Young people's activities

Figure 1.2 Terms in an expandable hierarchical taxonomy; top categories (left) and the expansion of one category (right) from the Abridged Integrated Public Sector Vocabulary, Version 2.00

in geospatial classification, as for regions, countries, provinces, and cities. While hierarchical taxonomies tend be used mostly for generic things or concepts, they can be used for proper nouns that naturally fall into a hierarchy, such as place names, product names, government agency names, or corporate department names.

The structure of a hierarchical taxonomy often reflects an organization of nested categories. Some hierarchical taxonomies permit a term to have multiple broader terms, thus appearing in multiple places in the taxonomy, whereas other hierarchical taxonomies do not permit this "polyhierarchy" structure. Hierarchical taxonomies may or may not make use of nonpreferred terms. Finally, nonhierarchical related-term relationships may exist but usually are not present in such hierarchical taxonomies

In contrast to the other types of taxonomies described subsequently in this chapter and this book, the hierarchical taxonomy is actually not a defined type of taxonomy. Rather, it is my designation for the narrower, standard definition of taxonomy: "A collection of controlled vocabulary terms organized into a hierarchical structure."[3] It is a kind of taxonomy that is commonly seen in countless real-world applications. And it is the type of taxonomy that the accidental taxonomist is probably most likely to create.

Thesauri

The classic meaning of a thesaurus is a kind of dictionary, such as *Roget's*, that contains synonyms or alternate expressions (and possibly even antonyms) for each term entry. A thesaurus for information management and retrieval shares this characteristic of listing similar terms at each controlled vocabulary term entry. The difference is that a dictionary-thesaurus includes all the associated terms *could potentially* be used in place of the term entry in various contexts; the user (often a writer) needs to consider the specific context in each case because in certain contexts some of the alternate terms would not be appropriate. The information retrieval thesaurus, on the other hand, is designed for use in *all* contexts within the domain of content covered, regardless of any specific term usage or document. The synonyms or near synonyms must therefore be suitably equivalent in *all* circumstances. An information retrieval thesaurus must clearly specify which terms can be used as synonyms (used from), which are more specific (narrower terms), which are broader terms, and which are merely related terms.

A thesaurus, therefore, is a more structured type of controlled vocabulary that provides information about each term and its relationships to other terms within the same thesaurus. National and international standards that provide guidance for creating such thesauri include the following:

- International Organization for Standardization (www.iso.org/iso/iso_catalogue.htm)
  - ISO 2788 (1986): Guidelines for the Establishment and Development of Monolingual Thesauri
  - ISO 5964 (1985): Guidelines for the Establishment and Development of Multilingual Thesauri
  - ISO 2788 and 5964 are to be replaced in 2011 by ISO 25964: Thesauri and Interoperability With Other Vocabularies
- American National Standards Institute and National Information Standards Organization (www.niso.org/kst/reports/standards)
  - ANSI/NISO Z39.19 (2005): Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies
- British Standards Institution (www.bsigroup.com)
  - BS 8723-1 (2005): Structured Vocabularies for Information Retrieval: Definitions, Symbols and Abbreviations
  - BS 8723-2 (2005): Structured Vocabularies for Information Retrieval: Thesauri
  - BS 8723-3 (2007): Structured Vocabularies for Information Retrieval: Vocabularies Other than Thesauri
  - BS 8723-4 (2007): Structured Vocabularies for Information Retrieval: Interoperability Between Vocabularies

Although the ANSI/NISO standard refers to controlled vocabularies, a document created in accordance with these guidelines is usually called a thesaurus.

The standards explain in detail the three types of relationships in a thesaurus: hierarchical (broader term/narrower term), associative (related term), and equivalence (use/used for). Additional information about a term, such as a *scope note*, may be included to clarify usage. An example of a term and its details from a thesaurus is shown in Figure 1.3. The consensus is that if a controlled vocabulary includes both broader/narrower and related-term relationships between terms, along with nonpreferred terms that redirect to the accepted term, then it is called a thesaurus.

In comparing a thesaurus with a hierarchical taxonomy, a thesaurus typically includes the features of a taxonomy plus the additional feature of associative relationships, for a greater degree of structural complexity. However, while all terms must belong to a limited number of hierarchies within a hierarchical taxonomy, this is not a strict requirement for a thesaurus. Although most thesaurus

materials acquisitions
UF acquisitions (of materials)
library acquisitions
BT collection development
NT accessions
approval plans
gifts and exchanges
materials claims
materials orders
subscriptions
RT book vendors
jobbers
subscription agencies
subscription cancellations

Figure 1.3 A term in the *ASIS&T Thesaurus* with its various relationships to other terms (BT: broader term. NT: narrower term. RT: related term. UF: used from)[4]

entries will list a broader and/or a narrower term, such relationships are not necessarily required for every term. If there is no appropriate broader term, that relationship may be omitted. In a thesaurus, the focus is more on the individual terms than on the top-down structure. Thus a thesaurus might include multiple small hierarchies, comprising as few as two or three terms, without the strong overarching tree structure typical of a hierarchical taxonomy.

If you had to force all the terms in a thesaurus into a single hierarchical tree, some of the hierarchical relationships would probably be imperfect. Thesaurus guidelines, however, mandate that each term's hierarchical relationships be accurate and valid. In addition, having multiple broader terms for an entry is never a problem in a thesaurus, whereas such "polyhierarchies" may be prohibited in a given hierarchical taxonomy. Some thesauri do in fact have a significant hierarchical structure, and thus the distinction between a hierarchical taxonomy and a thesaurus may be blurred. Finally, recursive retrieval by a broader term (explained in Chapter 9) is not as common in a thesaurus as in a hierarchical taxonomy.

The greater detail and information contained in a thesaurus, compared with a simple controlled vocabulary or a hierarchical taxonomy, aids the user (whether the indexer or the searcher) in finding the most appropriate term more easily. A thesaurus structure is especially useful for a relatively large controlled vocabulary that involves human indexing and/or supports a term list display that the end user (searcher) can browse. In contrast to a hierarchical taxonomy, which is designed for user navigation from the top down, a thesaurus with multiple means of access can more easily contain a greater number of terms. Thus, a thesaurus may be able to support more granular (specific) and extensive indexing than a simple hierarchical taxonomy can, especially if the hierarchical taxonomy lacks nonpreferred terms. As thesauri explain relationships among terms, they are more common in specialized subject

areas, where the purpose is not merely to aid the user in finding information but also to aid the user in obtaining a better understanding of the terminology. In some cases, thesauri have even been published and printed as stand-alone works, separate from any indexed content.

Examples of thesauri include the Getty Art & Architecture Thesaurus, the ERIC (Education Resources Information Center) Thesaurus for education research, and the NASA Thesaurus of aeronautics and space terminology. The periodical and reference index publisher ProQuest also refers to its topical controlled vocabulary as a thesaurus.

Ontologies

An ontology can be considered a type of taxonomy that has even more complex relationships between its terms than does a thesaurus. Actually, an ontology is more than that; it aims to describe a domain of knowledge, a subject area, by both its terms (called *individuals* or *instances*) and their relationships. This objective of a more complex and complete representation of knowledge stems from the etymology of the word *ontology*, which originally meant the study of the nature of being, existence, or reality. Tom Gruber provides a current definition of ontology:

> An ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. … ontology can be viewed as a level of abstraction of data models, analogous to hierarchical and relational models.[5]

The relationships between terms within an ontology are not limited to broader/narrower and related. Rather, there can be any number of domain-specific types of relationship pairs, such as owns/belongs to, produces/is produced by, and has members/is a member of. The creator of the ontology also creates

these relationship types. Thus, not only do the terms have meanings, but the relationships themselves have meanings as well. Relationships with meanings are called semantic relationships.

The terms within an ontology have not merely simple descriptions, such as scope notes in a thesaurus, but are also accompanied by specific attributes in a more structured format, such as properties, features, characteristics, or parameters. The terms also have assigned classes, which the ontologist defines, as an additional kind of classification. All of these components of an ontology—semantic relationships, attributes (for each of the terms/instances), and classes—contribute to making an ontology a richer source of information than a mere hierarchical taxonomy or thesaurus. A schematic representation of part of an ontology dealing with retail management appears in Figure 1.4.

While not considered standards, there are guidelines of specifications for constructing ontologies in machine-readable format

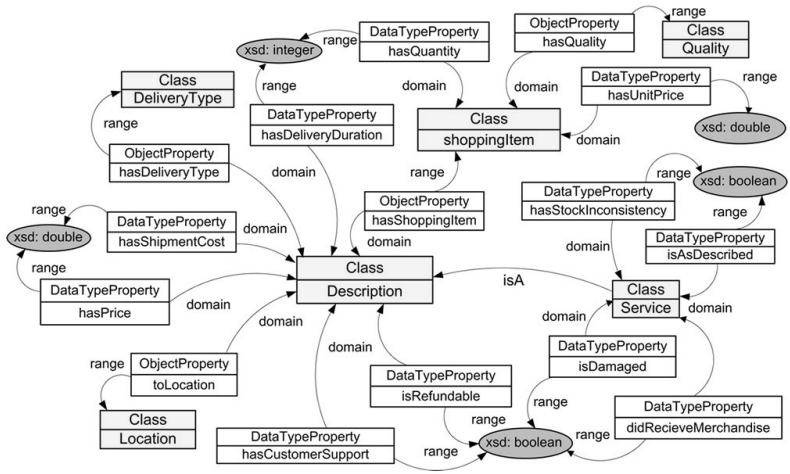


Figure 1.4 Example of a domain ontology dealing with retail management (reproduced with permission of the creators, Murat Sensoy and Pinar Yolum)[6]

for the web, which has become the most common implementation of this type of taxonomy. The World Wide Web Consortium (WC3) has published the RDF (resource description framework) Schema and the Web Ontology Language (OWL) recommendation. There is also a presentation format for ontologies called Topic Maps, which is the ISO 13250 standard. Topic Maps are implemented more in Europe than in North America. A looser structure of knowledge organization that does not attempt to adhere to such guidelines might be called a semantic network instead of an actual ontology.

Ontologies are suitable for any subject area, but a significant percentage of those currently published have been in the biological sciences, such as the Gene Ontology, Protein Ontology, Systems Biology Ontology, and Purdue Ontology for Pharmaceutical Engineering. It is an interesting irony that taxonomies, which got their start in biological classification, are now widely used for any form of knowledge, while ontologies, which originally applied to the broad scope of existence, are now used most often in the field of biology. As other scientists find a need to express more complex relationships among terms in their disciplines, the spread of ontologies to other subject areas, however, will likely increase. There is also a growing importance of ontologies in semantic search engine deployment in specialized industries, and building ontologies could be a growth area for experienced taxonomists. In 2009, a new organization for supporting ontologies, the International Association for Ontology and Its Applications (www.iaoa.org), was founded.

The designation given to a knowledge organization system—controlled vocabulary, taxonomy, thesaurus, ontology, and so on—depends largely on the complexity of the structure, but complexity is not the only factor to be considered. As all these designations have ambiguous meanings, the choice of what to call a set of terms also depends on what is most clear and understandable to the contributors, stakeholders, or end users. Depending on the display of the

knowledge organization system, the end users may not even need to know what it is called. The confusion in terminology is why we default to using the single designation of taxonomy in most contexts.

Applications and Purposes of Taxonomies

As we have seen from the various definitions, there are different kinds of taxonomies or controlled vocabularies, based on their complexity. However, that is only one way to classify them. A more practical approach is to categorize taxonomies by their application and use. While one taxonomy can certainly serve multiple functions, there tends to be a certain emphasis in its design, use, and purposes. As such, taxonomies serve primarily one of the following three functions, although there certainly can be combinations of the different types:

1. Indexing support
2. Retrieval support
3. Organization and navigation support

Indexing Support

For indexing or cataloging support, a taxonomy, better known as a controlled vocabulary in this context, is a list of agreed-on terms for the human indexing or cataloging of multiple documents and/or for indexing performed by multiple indexers, to ensure consistency. If multiple documents, especially by different authors, will be indexed over time, the indexer is apt to forget exactly which index terms were assigned and perhaps inadvertently use different synonyms when the same topic comes up in a different document. Similarly, different indexers will also choose different index terms for the same topic if not forced to use a controlled vocabulary.

Thus, the taxonomy's initial purpose is to serve the people doing the indexing, although a second, equally important purpose is to serve the end users, who, of course, benefit from consistently indexed content and may also have access to the taxonomy. This type of controlled vocabulary is used for cataloging entire works and for indexes to periodical articles, image files, database records, multivolume printed works, webpages, etc. Because indexers must always choose the most accurate terms, they often use a more structured thesaurus type of controlled vocabulary. The broader, narrower, and related term relationships guide the indexer to the best term, and scope notes further clarify ambiguous terms. Named entities are often indexed, too, and these are managed in an authority file. An authority file lacks the interterm relationships of a thesaurus but may have many synonymous nonpreferred terms for each preferred term, such as variations on an individual's name.

Controlled vocabularies for indexing support have been around the longest, and their format may be electronic or print. Such controlled vocabularies are used by reference and periodical article database publishers, including H.W. Wilson, ProQuest, Gale, and EBSCO; in more specialized subject databases such as Chemical Abstracts and PsycINFO; and in the internal documents of large companies, especially those in the sciences. The fact that some of these controlled vocabularies are offered for sale/license illustrates the fact that they serve the purpose of indexing and not just specific content retrieval.

While controlled vocabularies for indexing are quite widespread, those that are publicly available on the web are limited and tend to be those published by public agencies. You may search or browse them, and in some cases, you may also access linked content. Library of Congress Subject Headings and Medical Subject Headings are two such examples.

Library of Congress Subject Headings (LCSH; authorities.loc.gov) contains both subjects and names, and covers all subject

areas. LCSH was originally established for cataloging library materials but has also been adapted by various publishers for indexing articles. The terms are called *authorities*, as in authority file, even those that are not named entities. The purpose of the website is to aid catalogers of library materials in finding the approved subject heading in the Library of Congress controlled vocabulary. It is not aimed at the end user looking for a book, although consistently cataloged books will, of course, benefit the user. The subject headings can be searched and the results browsed alphabetically. Nonpreferred terms are included in the alphabetical list along with preferred terms. Nonpreferred terms are prefaced by a button labeled References, which provides a cross-reference to the preferred term. Preferred terms are called *authorized headings* (see Figure 1.5).

Medical Subject Headings (MeSH; www.nlm.nih.gov/mesh/MBrowser.html) is the thesaurus of the U.S. National Library of Medicine, which is considered the authority for medical terms. Users can search terms, or they can browse by selecting the button Navigate from Tree Top. The browse display is hierarchical rather than alphabetical. Clicking once on a term expands the tree and reveals its narrower terms; double-clicking on a term displays its details (see Figure 1.6).

Other examples of thesauri that aid indexing and are publicly available include the ERIC Thesaurus (eric.ed.gov), sponsored by the Institute of Education Sciences of the U.S. Department of Education, and the various controlled vocabularies of the Getty Research Institute of the J. Paul Getty Trust: the Getty Art & Architecture Thesaurus, Getty Thesaurus of Geographic Names, Cultural Objects Name Authority, and Union List of Artist Names (www.getty.edu/research/conducting_research/vocabularies).

Retrieval Support

A taxonomy that serves indexing also serves end-user retrieval. Searchers benefit from nonpreferred terms, as their search terms

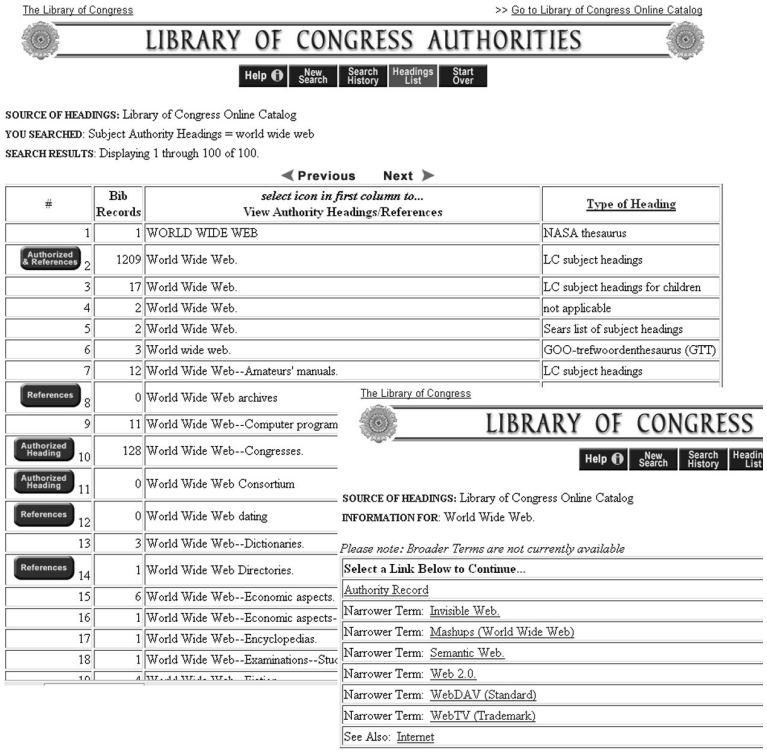
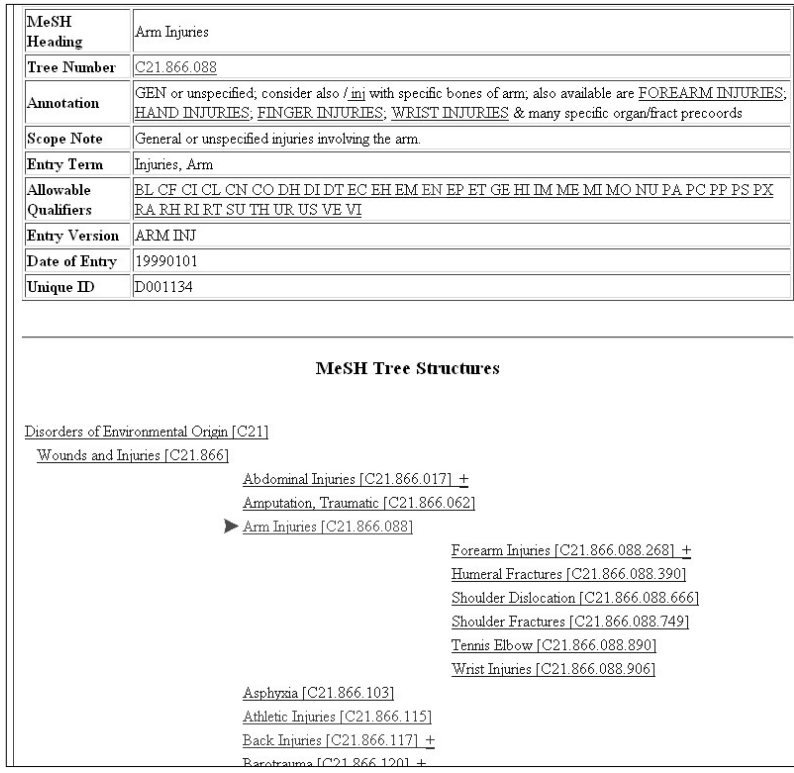


Figure 1.5 Two successive screenshots from Library of Congress Subject Headings, searching on the term *World Wide Web* and displaying its details

may be different from the terms used to index the document. For example, a user may type in **doctors** for articles that are about physicians. Users can also take advantage of broader and narrower term relationships or hierarchies to broaden or narrow their search. These relationships, and also the related-term relationships, may suggest to users other possible terms of interest. In such cases, the end-user searcher is seeing an explicit representation of the taxonomy to navigate.

There are also taxonomies designed to aid search retrieval without supporting human indexing. These taxonomies are typically



MeSH Heading	Arm Injuries
Tree Number	C21.866.088
Annotation	GEN or unspecified; consider also / inj with specific bones of arm; also available are FOREARM INJURIES; HAND INJURIES; FINGER INJURIES; WRIST INJURIES & many specific organ/fract precoords
Scope Note	General or unspecified injuries involving the arm.
Entry Term	Injuries, Arm
Allowable Qualifiers	BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH UR US VE VI
Entry Version	ARM INJ
Date of Entry	19990101
Unique ID	D001134

Figure 1.6 Searching Medical Subject Headings for the selected term *arm injuries*

mapping-tables of terms and their synonyms/variants designed to aid online retrieval. These might be synonym rings or synsets, especially if the terms are not even displayed to the user; or, if there is a display, it may designate preferred terms.

Depending on the user interface display, there may or may not be a hierarchical structure to the taxonomy. A hierarchical arrangement allows users to browse and locate narrower (more specific) subjects of interest. Thus, users find out what is included in the taxonomy and what is not, saving themselves the trouble of repeatedly typing in terms that yield no results. Users may also find related subjects of interest by browsing the hierarchies.

These types of controlled vocabularies are often used with site search engines, enterprise search systems (used internally within a large organization), online databases, and large commercial directories (such as online "yellow pages" or classified ads). The format is always electronic, and a form of automated indexing is usually involved.

Examples of taxonomies aiding retrieval include the Verizon SuperPages yellow pages directory site (www.superpages.com/yellowpages) and the Amazon.com ecommerce site (www.amazon.com/gp/site-directory), as shown in Figure 1.7. While a hierarchy can be selected for browsing in each, the synonyms in the case of Verizon SuperPages and the related subject links in the case of Amazon.com are not displayed to the user, although the links are evident in the display of results.

Figure 1.7 Top-level taxonomy of Amazon.com

Faceted Taxonomies for Retrieval Support

One way to better serve specifically the retrieval of data is to construct a controlled vocabulary that is divided into multiple subsets, lists of terms of different types representing different aspects of information. These aspects are often called *facets,* and this type of controlled vocabulary is therefore called a *faceted taxonomy*. Examples of facets might be people, places, events, products, and laws. Facets can also reflect metadata other than subject categories, such as document type, author, and audience. The search interface for a faceted taxonomy is designed for the user to search on a selected combination of multiple facets.

Faceted taxonomies are commonly used for online databases and ecommerce sites, such as the shoe-retailing site Shoebuy.com. In Shoebuy's advanced search (www.shoebuy.com/s.jsp/r_as), the facets are Brand, Category, Type, Size, Width, Color, Country, Price, and, additionally for women's shoes, Heel Height. Another example of a faceted browse interface is on the Microbial Life Education Resources site (serc.carleton.edu/microbelife/resources), where facets are Subject, Resource Type, Extreme Environments, Ocean Environments, and Grade Level (Figure.1.8).

Faceted taxonomies, or *faceted browse* systems, make use of the electronic format. Depending on the size of the vocabulary in each facet, these taxonomies may not make use of synonyms and may or may not have hierarchies within them. Some facets can be quite small. Facets will be discussed in more detail in Chapter 6.

Organization and Navigation Support

A taxonomy, as a hierarchy, can provide a categorization or classification system for things or for information. For the organization of information, we often see taxonomies applied in website information architecture (structural design), online information services, intranet content organization, and corporate content management systems. In such website or enterprise taxonomies,

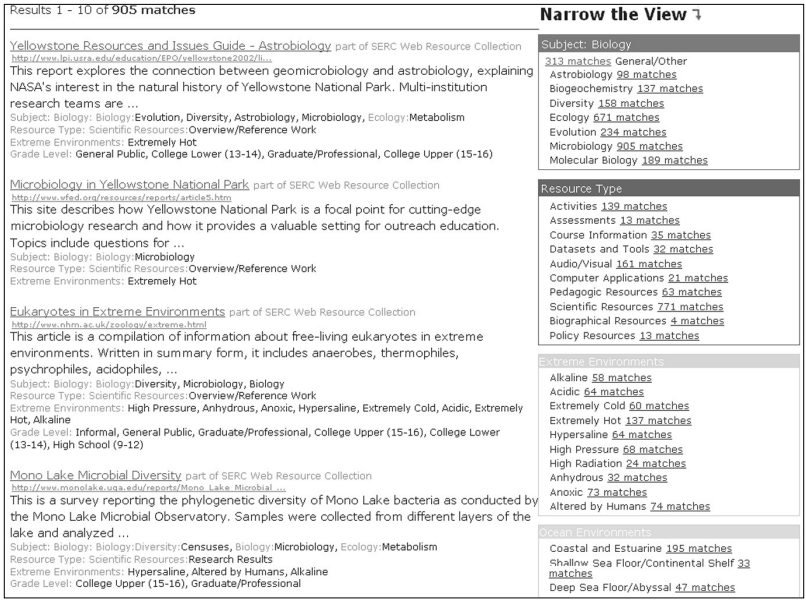


Figure 1.8 Faceted taxonomy in the margin of the Microbial Life Education Resources search site

the emphasis is on classification and guided user navigation rather than on search and retrieval of specific information. *Navigation* means finding one's way around, whereas *retrieval* means going after specific information. The taxonomy for a website is a lot like a table of contents, organized by topic. It can be reflected in the navigational menu and in the site map. As such, it might be called a *navigational taxonomy*. These types of taxonomies tend to be relatively small and can coexist with additional, more detailed taxonomies elsewhere on the website.

An example of navigational website taxonomy that is present in both the site map and the navigational menu can be found on the Information Architecture Institute site map (iainstitute.org/en/site-map.php), where the top-level categories of the taxonomy and the navigation are Member Services, IA Network, Learning IA, and

About Us (Figure 1.9). Another example of a navigational taxonomy is MyFlorida.com, the State of Florida site map (www.myflorida.com/taxonomy), where the top-level categories of the taxonomy, which also are the main navigation menu items, are Visitor, Floridian, Business, and Government. It is interesting to note that the file name for this site map page has been named taxonomy.

Enterprise taxonomies can be very large, but the top levels typically demonstrate some form of information organization for the enterprise. The purpose is not merely to retrieve documents but also to help users better understand the organization of the enterprise and its intranet and thus make better use of it.

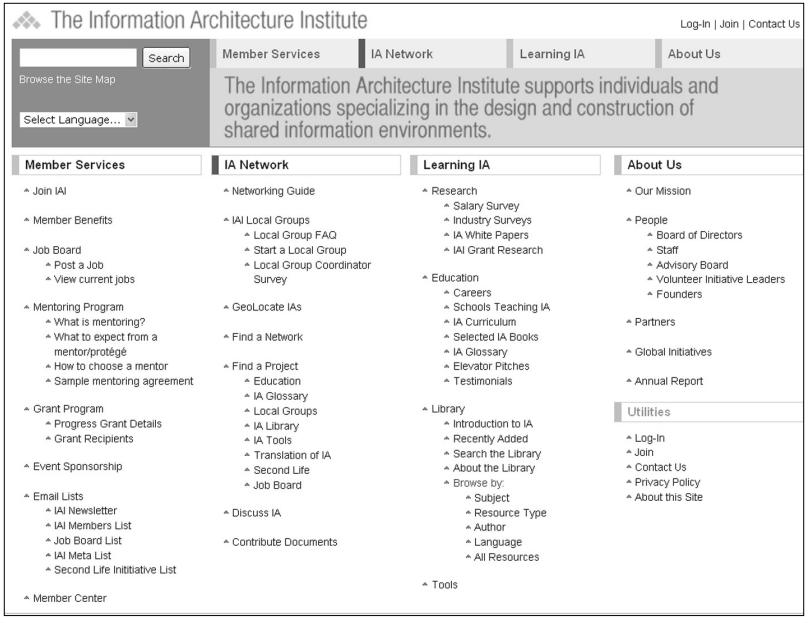


Figure 1.9 The Information Architecture Institute site map, a navigational taxonomy

Taxonomies for License

Although the primary objective of this book is to provide instruction on building taxonomies, it is not always necessary to build an entire taxonomy from scratch. Some or all of a taxonomy could be acquired from another source. While navigation taxonomies for websites and intranets and taxonomies for enterprises and content management systems should be custom-created, a taxonomy for the indexing of documents or files in a given subject area could be purchased or licensed. Furthermore, taxonomies for license not only serve the purposes of indexing and content retrieval but may also provide an outline of a domain of knowledge. Many subject areas are already covered by existing published taxonomies. There are generic taxonomies for geographic places, industry types, product types, and so forth. In addition, lists of named entities are available from various sources. You might consider licensing an external taxonomy if the right taxonomy already exists or if creating one from scratch would be too great a task due to size, specialty subject area, and limited time. A licensed controlled vocabulary could be used for merely a single facet or for part of a larger set of taxonomies.

Taxonomies or controlled vocabularies that are available for license come from all kinds of sources: government agencies, professional associations, other nonprofit organizations, and a few commercial enterprises. Governmental published taxonomies available for license (or even without a license) include LCSH, Library of Congress Thesaurus for Graphic Materials, MeSH, USDA National Agricultural Library Thesaurus, and the U.K.'s Integrated Public Sector Vocabulary. The Getty Research Institute (part of the J. Paul Getty Trust) is a reputable nonprofit provider of controlled vocabularies, including the Art & Architecture Thesaurus, Getty Thesaurus of Geographic Names, and Union List of Artist Names. Commercial vendors of taxonomies, both pre-built and custom, include Dow Jones Client Solutions, which specializes in business

and finance, and WAND Inc., which has strengths in products and services.

The largest (and the only multisource) directory of taxonomies available for use is Taxonomy Warehouse (www.taxonomywarehouse.com). The list was started by the taxonomy software vendor Synapse and is now managed by Dow Jones Client Solutions. The database includes hundreds of taxonomies, including most of those mentioned previously. Some are simple controlled vocabularies or glossaries, but others are full-featured thesauri. Although some are hosted on the web, the data files (usually in CSV or XML formats) can be obtained for most of them. Figure 1.10 shows the information that Taxonomy Warehouse provides for a specific taxonomy. A single source/publisher may also offer numerous taxonomies on different subjects.

Formats may vary, but typically, taxonomies or thesauri that are made available for other uses are formatted in some kind of XML whereby all terms, relationships, nonpreferred terms, scope notes, and so forth are retained when they are imported into other taxonomy management systems. The use of XML and other interoperable taxonomy formats is described in greater detail in Chapter 10.

If you acquire a taxonomy, however, you will likely want to modify or enhance it for your own needs, and in any case it will require some maintenance over time. The following is an example of how a generic taxonomy taken as-is may not be ideal. A large-scale historical digitization project that coded early American election results used the Getty Thesaurus of Geographic Names. Even though the thesaurus includes historical place names, it was still found to be insufficient for the project's needs. It does not include all the towns and boroughs that were named in the elections project and does not indicate exactly when various historical names were used or when boundaries were redrawn.

Licensing agreements may allow use of a taxonomy without a fee in some cases but may prohibit for-profit use or require statements

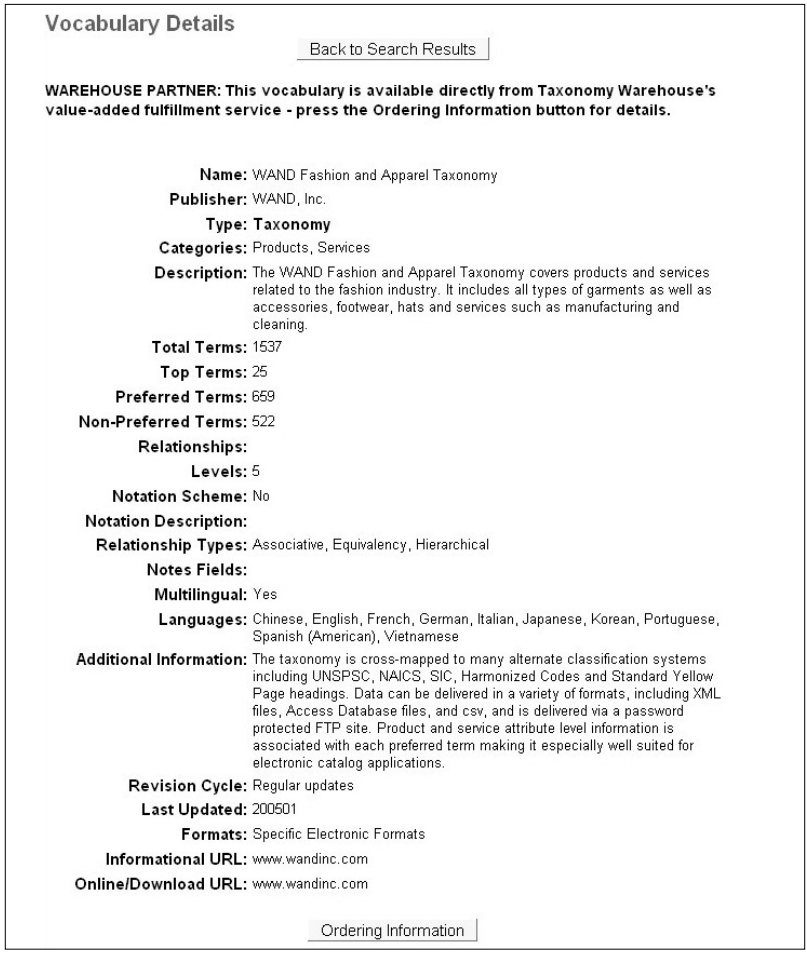


Figure 1.10 Example display of information for a single taxonomy offered for sale through Taxonomy Warehouse

referring to the original copyright holder. If the taxonomy is treated as a published copyrighted work, whether free or for a fee, then there will also be restrictions on making changes to it. The policy for modifications to the Getty vocabularies is as follows:

> The license for the Getty vocabularies, AAT, TGN, ULAN, and CONA (in development) does not restrict additions or alterations to the vocabulary, provided that—if the product is visible to the public or other end users—the terms and other information that come from the Getty vocabulary are labeled with a citation for the vocabulary and the copyright of the J. Paul Getty Trust. Likewise, any additions and alterations must be clearly indicated as NOT being from the Getty vocabulary. If the vocabulary is only used behind the scenes for retrieval and not visible to the end user, labeling which bits come from Getty vocabulary and which are added is irrelevant.[7]

The policy for using and modifying MeSH is as follows:

> If the use is not personal, (1) the U.S. National Library of Medicine must be identified as the creator, maintainer, and provider of the data; (2) the version of the data must be clearly stated by MeSH year, e.g., 1997 MeSH; and (3) if any modification is made in the content of the file, this must be stated, along with a description of the modifications.[8]

Often you will want to make changes to the acquired taxonomy, so make sure the license permits changes. Also be aware that you are responsible for continued updating. Thus, a solid understanding of how to create terms and relationships, as discussed in Chapters 3 and 4, is still necessary to manage pre-built taxonomies. Therefore, acquiring a taxonomy from an external source does not eliminate the need for a taxonomist. Starting with a pre-built taxonomy, though, is much easier for the less experienced taxonomist. You can follow examples of term formats and relationships as you build out the taxonomy further. Licensed taxonomies, both

those that prohibit and those that permit changes, typically offer updates through an annual subscription.

History of Taxonomies

Taxonomies are both new and old. "Both librarians and indexers were doing 'taxonomy' long before it became a hot topic in the 1990s," wrote taxonomy trainer Jean Graef of the Montague Institute.[9]

Taxonomies in Cataloging and Indexing

The earliest taxonomies were for classification, such as for organisms or for books, but each item could only go in one place in the taxonomy. For example, a book gets a single call number for its location on the shelf. In the field of library science, by the end of the 19th century more practical taxonomies emerged that supported supplemental descriptive cataloging, which is not limited to one descriptive term per book. The leading controlled vocabularies for cataloging books have been the American Library Association Subject Headings (1895), LCSH (1898), and the Sears List, published originally as the List of Subject Headings for Small Public Libraries (1923). These were simple controlled vocabularies lacking broader/narrower and related term relationships. LCSH used *See also* references for every kind of relationship and began to introduce broader term, narrower term, and related term references only in 1985.[10]

The LCSH, still in its simpler form, was adopted by various periodical index publishers for the indexing of articles from multiple newspapers, magazines, and journals. These publsihers include the H.W. Wilson Company, which is as old as the LCSH, and in the 1970s Information Access Company (now Cengage Learning) and ABI Inform (now ProQuest). H.W. Wilson still uses modified LCSH,

whereas Cengage's and ProQuest's controlled vocabularies have diverged over the years based on the work of their taxonomists.

Meanwhile, professional societies developed their own controlled vocabularies for indexing periodical literature in their fields since at least the early 1900s. These included the American Chemical Society's Chemical Abstracts Service founded in 1907. The word *thesaurus* was first used to refer to a controlled vocabulary for information retrieval purposes by Peter Luhn at IBM in 1957. Early published thesauri included the Department of Defense's *ASTIA Descriptors* in 1960 and the American Institute of Chemical Engineers' *Chemical Engineering Thesaurus* in 1961.[11] Standard thesaurus relationships emerged over time, and guidelines were developed that reinforced them, including UNESCO's 1967 guidelines, which formed the basis of the ISO 2788 standard of 1986.[12] Since the 1960s, various companies, government agencies, and professional associations have published dozens of specialized thesauri. In 1972, the Dialog began offering the first publicly available online research service, providing access to multiple bibliographic citation databases indexed with controlled vocabularies.

Corporate Taxonomies

Up through the 1980s, however, taxonomy (thesaurus) development was mostly limited to large index or literature-retrieval database publishers and to a few large companies, especially in the sciences (such as DuPont), or government agencies. The companies and government agencies that developed taxonomies did so mostly within specific subject areas. Taxonomies for an entire organization, that is, enterprise-wide taxonomies, first began to appear in the late 1970s, but their adoption was limited. According to taxonomy and knowledge management consultant Lynda Moulton, it was not so much a lack of interest but simply the limitations of software tools at the time that hindered a wider adoption of enterprise taxonomies.

Moulton recalls teaching a number of thesaurus construction workshops during 1982–1984, attended by librarians and indexers from such companies as Liberty Mutual, John Hancock, Fidelity, MITRE, and Digital Equipment Corp.[13]

Contemporary library automation began to emerge in the late 1970s and systems for "special libraries" (corporate libraries and information management) as early as 1980. Although dedicated taxonomy management systems had not yet appeared on the market, these earlier systems included taxonomy management features. These included BiblioTech by Comstow (acquired in 1999 by Inmagic), which was first installed at Polaroid in 1981, and TechLib, released in 1984, which was built on BASIS and acquired by OpenText in 1998. Comstow Information Services held a number of workshops that were devoted to thesaurus development for corporate libraries in the early 1980s.[14]

It was only in the late 1990s that a broader interest in taxonomies, and the corresponding tools to support them, developed. For example, the taxonomy consultancy Earley and Associates started working on classification, categorization, and metadata projects (essentially taxonomy, but not called that yet) to help their clients make the most out of the Lotus Notes application, by building classification structures, forms, and navigation. In 1998, IBM introduced its Lotus Discovery Service, which "really called out the need for a taxonomy," according to Seth Earley, so he and other consultants at the time provided services in creating taxonomies for Lotus Notes.[15]

The Growth of Enterprise and Web Taxonomies

The emergence and growth of the web in 1990s was a major contributing factor in the growing interest in taxonomies, for several reasons. The web enabled smaller publishers to offer online information services. Companies started developing intranets that quickly expanded in size and required better navigation and

search. "With growth of the internet, there was a lot of interest in building to improve search results," explained Synapse co-founder Trish Yancey regarding the start of the company.[16] The proliferation of search engines, and then site search or enterprise search, also led to an interest in taxonomies as it became apparent that search alone was not sufficient. According to Jean Graef, "Taxonomy became hot when IT realized that search engines by themselves couldn't solve the whole retrieval problem."[17] Finally, attention to site design and navigation through the new field of information architecture also put value on taxonomies. Indexer, information architect, and taxonomist Fred Leise wrote, "As the field of information architecture and the influence of Louis Rosenfeld's and Peter Morville's *Information Architecture for the World Wide Web* grew, the knowledge of library science-related information such as faceted browsing classifications and the use of synonym rings as search improvements spread more widely."[18]

The growing interest in taxonomies in the 1980s and 1990s was also reflected in the growth of taxonomy management software. Software for creating and maintaining taxonomies was originally developed internally within the few large organizations that had already developed taxonomies. In 1980 Comstow released BiblioTech, its fully integrated library system for corporate and government libraries, which included a module for thesaurus creation, fully integrated with the cataloging and indexing module. Battelle Columbus Laboratory released similar functionality in TechLib soon after.[19] In the mid-1980s commercial PC software for thesaurus creation became available, including the desktop tools MultiTes, Term Tree, TCS (later a part of WebChoir), and several others that have not survived. Larger-scale client server systems became available in the 1990s, reflecting the growing demand. Synapse Corp. had developed software to maintain taxonomies it was creating for others as a consulting service but soon found a market for the software itself and began selling the Synaptica taxonomy

management system in 1999. Similarly, Access Innovations had been offering indexing services since 1978 but then found demand for its taxonomy management tool and has commercially offered its Data Harmony Thesaurus Master since 1998. Wordmap, another major taxonomy software vendor, was founded in 1998. Content management systems and enterprise search solutions, which only really entered the market in the 1990s, have also begun to offer taxonomy management components or features.

The 1990s also saw the establishment of commercial vendors of taxonomies, including Synapse Corp. and WAND, both of which were founded in 1995, and the automatic taxonomy generator company Intellisophic in 1999.

The rise of the term taxonomy paralleled this growing interest in taxonomies. Taxonomy consultant Ron Daniel, now a partner of Taxonomy Strategies, got his start in the field working for the Department of Energy on its thesaurus. He recounts how around 1997, it was starting to use the word *taxonomy* interchangeably with *thesaurus* and another term that hasn't become quite as popular, *synonymy*.[20] Earley recalls starting to use the word *taxonomy* with clients around 1996 or 1997. Moulton recalls the adoption of the term taxonomy:

> Throughout my professional career, first as a technical librarian, then as a software developer and consultant, the operative terminology for my work was thesaurus. … I first heard the term taxonomy applied to "organization maps," in the early 1990s. … In the late 1990s I began to see the term "taxonomy" routinely used to describe "terminology maps," "topical hierarchies," and "terminology relationships." Before long, taxonomy became the de facto label for topical navigation schemes on commercial websites that had a focus on text content retrieval. … At some point I recognized that the term

> thesaurus was not understood by IT and business management professionals. So, about 2000, I adopted taxonomy to cover any controlled vocabulary being developed or applied in any indexing, metadata management or retrieval situation. … To this day, I use thesaurus and taxonomy interchangeably depending on which word will most likely resonate with my audience.[21]

Our online survey completed by 65 taxonomists in November and December 2008 also confirmed the recent trend toward increased use of the term taxonomy. Whereas 17 (26.2 percent) of the respondents had been involved in taxonomy work as we define it (taxonomies, controlled vocabularies, metadata for classification or tagging, thesauri, or authority files) for more than 15 years, only four of them, or 6.3 percent of the total, reported that their work was specifically called taxonomy as long ago as 1993 (15 years prior to the survey). This response contrasts with the answers of those who had been working in the field less than a year: Six out of nine of these respondents call their work taxonomy (see Appendix A, Questions 3 and 4).

Another way to track the growing popularity of taxonomies is to count the magazine and trade journal articles (excluding scholarly journals) in literature retrieval databases with the plural word *taxonomies* appearing in their texts. While many of these articles may be about specific-subject taxonomies, rather than information taxonomies in general, searching on the word *taxonomies* (rather than *taxonomy*) focuses the results more on the creation of generic information taxonomies. Looking at Gale's InfoTrac PowerSearch of 12 databases of magazine and newspaper articles and at HighBeam Research's database of journals, newspapers, and press releases, occurrence of the word *taxonomies* shows a marked increase especially in the period of 1998 to 2002, as shown in Table 1.1. (HighBeam's numbers are higher because High Beam includes

Table 1.1 Number of periodical articles including the word *taxonomies*

Year	Gale InfoTrac	HighBeam
1997	6	88
1998	14	74
1999	31	128
2000	67	242
2001	85	269
2002	205	413
2003	134	382
2004	171	401
2005	151	506
2006	127	504
2007	125	613

scholarly journals, which occasionally mention scientific nomenclature taxonomies.) Although the periodical collections in both database services also grew over time, the collection did not grow at such a fast rate.

A similar more focused search on the truncated string *taxonom** in the industry journals of Information Today, Inc. (specifically *ONLINE Magazine*, *EContent*, *Information Today*, *KM World*, *Computers in Libraries*, and *Searcher*) shows a similar trend: 0 results through 1989, a significant increase in articles on the subject just before and after 2000, and then a more recent slight decline (Figure 1.11).

The turning point came around 2000. In the summary of the European Business Information Conference (EBIC) conference in 2000, Tom Koulopoulos, president of the Delphi Group and renowned writer and public speaker on knowledge management, declared, "Taxonomies are chic." Since then taxonomies have been a popular topic in conference presentations and workshops. The Montague Institute held its first taxonomy roundtable in 2000. A

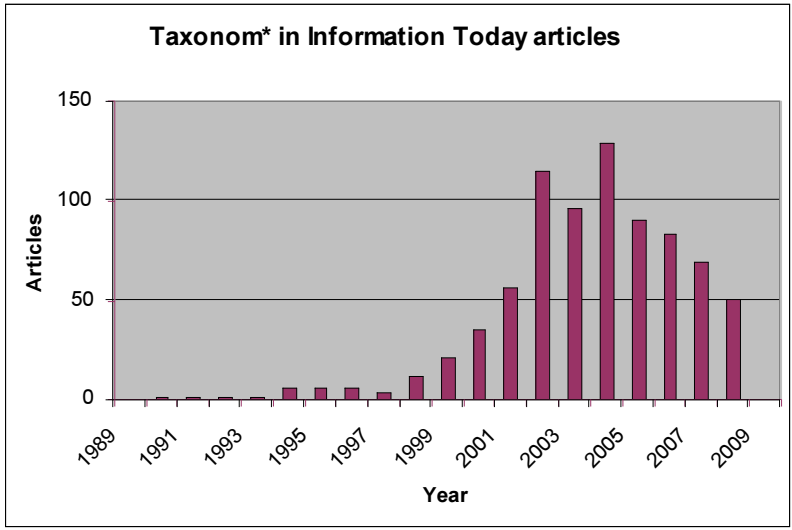


Figure 1.11 Numbers of trade journal articles returned by the search string *taxonom**

significant number of taxonomies had become available publicly (usually for licensing), so in 2001 Synapse Corp. launched its Taxonomy Warehouse website directory of taxonomies. Taxonomy consultant Marcia Morante recalled:

> The year 2000 was probably the very beginning of the commercial taxonomy wave. That was the year that I started with Sageware, and we still had to do a lot of explanation. But by that time, there were definitely a few companies whose business was built around taxonomies.[22]

Although newer buzzwords, such as *folksonomy*, *social networking*, and *Web 2.0*, have superseded taxonomy in the 2000s,a sustained interest in taxonomy and taxonomists continues. Search industry analyst Steve Arnold analyzed web traffic on Google from 2002 to 2008 on the term *taxonomy* and found it continuing to

remain strong, stronger than *CMS* (content management systems). He concluded that "taxonomy is a specialist concept that seems to be moving into the mainstream."[23]

Endnotes

1. Gale Hodge, *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files* (Washington: The Digital Library Federation Council on Library and Information Resources, 2000), 1, www.clir.org/pubs/reports/pub91/pub91.pdf
2. Ibid. 4–7.
3. National Institute of Standards Organization, *ANSI/NISO Z39.19-2005 Guidelines for Construction, Format, and Management of Monolingual Controlled Vocabularies* (Bethesda, MD: NISO Press, 2005), 166.
4. Alice Redmond-Neal and Marjorie M. K. Hlava, eds., *ASIS&T Thesaurus of Information Science, Technology, and Librarianship*, 3rd ed. (Medford, NJ: Information Today, 2005).
5. Tom Gruber, "Ontology," tomgruber.org/writing/ontology-definition-2007.htm
6. This image, reprinted with permission of the authors, first appeared in Murat Sensoy and Pinar Yolum, "Ontology-Based Service Representation and Selection," *IEEE Transactions on Knowledge and Data Engineering* 19, no. 8 (2007). It is also available at mas.cmpe.boun.edu.tr/project/AgentBasedSemanticWebServices.htm
7. Patricia A. Harpring (Managing Editor of the Getty Vocabulary Program, Getty Research Institute), email to author, August 20. 2009.
8. MeSH Memorandum of Understanding, www.nlm.nih.gov/mesh/termscon.html
9. Jean Graef, email to author, November 21, 2008.
10. Alva Stone, "The LCSH: A Brief History of the Library of Congress Subject Headings, and Introduction to the Centennial Essays," *Cataloging & Classification Quarterly* 29, no. 1–2 (2000),1.
11. Jean Aitchison and Stella Dextre Clarke, "The Thesaurus: A Historical Viewpoint With a Look to the Future," in *The Thesaurus: Review, Renaissance, and Revision*, eds. Sandra K. Roe and Alan R. Thomas (Binghamton, NY: Haworth Press Inc., 2004), 7.
12. Ibid. 8.
13. Lynda Moulton, telephone interview with the author, October 19, 2009.

14. Lynda Moulton, email to author, October 19, 2009.
15. Seth Earley, telephone interview with author, November 22, 2008.
16. Kimberly S. Johnson, "International Information Provider Buys Franktown, Colo., Taxonomy Company," *Denver Post*, June 30, 2005.
17. Jean Graef, email to author, November 21, 2008.
18. Fred Leise, email to author, December 2, 2008.
19. Lynda Moulton, email to author, October 19, 2009.
20. Ron Daniel, telephone interview with author, December 1, 2008.
21. Lynda Moulton, email to author, November 9, 2009.
22. Marcia Morante, email to author, November 21, 2008.
23. Steve Arnold, "Taxonomy: Silver Bullet or Shallow Puddle," *Beyond Search* blog, September 27, 2008, arnoldit.com/wordpress/2008/09/27/taxonomy-silver-bullet-or-shallow-puddle