Taxonomy and Knowledge Organization

Jan Herd
Science, Technology & Business Division
The Library of Congress

Slide Presentation given at The Library of Congress
West Diningroom
Thursday, June 7, 2001

First I would like to thank Susan Tarr and the staff of FLICC for inviting me to speak to you today. It is similiar to the presentation I gave at the E-Gov Knowledge Management Conference in April 2001on "Knowledge Organization and Taxonomies." The handout is a list of Web resources on standards related to taxonomies and taxonomy builder software which represents various methods of content management.

Today's presentation deals with:
The impact of the Web
Librarians work in dot coms and cooperative cataloging
Web-based traditional cataloging tools
Importance of controlled vocabulary on the Web
Metathesauri and subject correlations with some examples
Mapping of standard and specialized information systems
And taxonomies which are the "new" tools on the Web.

During this presentation you will see that I have made use of a metaphor to draw an analogy between the current situation of knowledge organization and some of the challenges we face to create better retrieval systems for the Web.

[slide 2 Picture of tree by an expanse of ocean]
"Water, water everywhere and not a drop to drink"[1]

[1] Coleridge, Samuel Taylor. Rime of the Ancient Mariner.

We are drowning in a deep sea of information. It is very much like the Dead Sea which (due to heavy salt content) does not allow us to sink into its depths but keeps us afloat with no drinkable water.
What's that? You say you would prefer to drink bottled water? Well, I guess you found the answer to your problem. You have to filter the water and put it in a container. We all know bottled water is expensive but would you prefer to drink salt water?

This analogy describes the reality we face with the ocean of information that is the Web.

[Slide 3 Is the Web too big to organize?]
We have all heard gargantuan statistics on the number of Web pages that exist and that are produced daily. The Web has simply grown too big for human editors to classify. A recent study numbered the Web at a billion pages and some analysts estimate that 1.5 million pages are added to the Web each day.

So why should we care how many Web pages are created each day?
Are we interested in all of them? Certainly not.
Are we interested in some of them. Yes, definitely.
OCLC libraries are using CORC or Cooperative Online Resource Cataloging software to record the selection decisions of subject specialists and to catalog Web sites which become part of the OCLC WorldCat database.

Can we continue using CORC as a Web-based selection and cataloging tool in the future?
Is it adequate for our needs?
Does it need to evolve and use some of the tools that are currently being developed for the Web?

All these questions will be answered in time. I do not presume to have the answers to all of them this morning. I do believe, however, that it would be useful to take stock of how the problem is viewed and some tools that purport to have solutions.

While library catalogers are using CORC to organize Web sites other librarians are contributing in different ways.

[Slide 4 Librarians work in corporate settings, list of organizations]

Given the long history of library science organizing information it is not at all surprising that librarians are currently organizing the Internet for such

companies as Yahoo, NorthernLight.com search engine, Amazon.com , Microsoft, etc. Some of the most important organization for our global e-commerce rests on the shoulders of librarians who are quietly organizing the world of knowledge as they have for so many thousands of years. "Amazon.com has 50+ catalogers on their staff and Microsoft's internal portal has three full-time taxonomists on the payroll."[2]

[Slide 5 OCLC Library Corporation cooperative cataloging stats]
OCLC libraries, along with some foreign participating libraries, cooperatively catalog vast numbers of works. The WorldCat database includes more than 45 million bibliographic records for all formats. That's a lot of metadata! This database also includes bibliographic records for approximately 350,000 Web sites which were cataloged by OCLC participating libraries using the Cooperative Online Resource Catalog or CORC interface . As you can see by these statistics librarians are very busy putting order to not only the traditional resources but also the electronic ones simultaneously by using some new tools to help create metadata.

[Slide 6]
Traditional cataloging tools have been made Web compliant in recent years.
Medical subject headings first appeared in 1996
WebDewey appeared in 2000
and this year LC developed the ClassificationWeb software

The Web compliance of these traditional cataloging tools allows us to utilize a common interface which is known to most catalogers. Using their knowledge organization skills, catalogers not only assign subjects and classification which becomes metadata on the Web but we are also responsible for creating and maintaining controlled vocabularies and library classification systems up to date in many and varied subject domains. Maintaining controlled vocabularies and classification are expensive and are manual forms of "concept mapping." Is it really necessary for us to continue to support these tools? Is there an easier way to simplify content access in the Web world?

[Slide 7 Importance of controlled vocab. As metadata]

---

[2]Crandall, Mike. "Taxonomies for the Real World: the Business Imperative to Simplify Content Access" Paper presented at the Taxonomies for Business Conference in London in Oct. 2000.

These questions were addressed in recent years by the ALA SAC recommendations on metadata and subject analysis. The complete recommendations are located at http://www.ala.org/alcts/organization/ccs/sac/metarept2.html

I will review a select few. The recommendations read:

" Trained catalogers may choose to continue to apply LCSH to the metadata records in the same manner as those assigned to MARC records."

"Classification data should be included in the metadata record by those who have the expertise to do so. "

Under the recommendations dealing with the Application of metadata the first recommendation reads
"In the Dublin Core metadata record, the Subcommittee recommends the inclusion in the SUBJECT element of both free-text and controlled terms, where appropriate and feasible, in order to achieve optimal recall and precision in retrieval."

The third recommendation under Application of metadata reads
"The adoption or adaptation of Library of Congress Subject Headings or Sears List of Subject Headings (for subject representation on a broader level) as the basis for subject data in the Dublin Core metadata records for a general collection is recommended."

"With regard to syntax, the use of full LCSH subject strings, if feasible (i.e., if time and trained personnel are available), particularly in the OPAC environment, should be encouraged."

The last two recommendations deal with classification systems, they recommend using
"Classification data at the most exhaustive or specific level" and state that
"Classification notation should be included."

As you can clearly see by these ALA recommendations. There is very strong sentiment in the library community that controlled vocabularies and classification are mandatory as metadata in the future.

[Slide 8]
Controlled vocabularies have been used "behind" search engines in many online subscription databases which provide excellent search and retrieval systems. The new adherents to controlled vocabularies are coming from an elite group of Web content managers who utilize controlled vocabularies whenever possible because they work! Just this year Kiplinger, a leading financial publisher, has invested in the development of a thesaurus to improve retrieval of its publications on the Web. Of all the investments this company could make it is fascinating to think that the development and maintenance of a thesaurus is considered to be a primary goal. If a financial information provider invests in thesaurus creation and maintenance and other Fortune 500 companies are investing in taxonomies for their Web portals, don't you think we should continue to maintain knowledge organization systems that have done what the e-commerce world is just now beginning to apply?

[Slide 9]
During the recent LC Conference on Bibliographic Control in the New Millenium, Prof. Sherry Vellucci, Associate Professor at St. John's University stated that in a network environment "authority control is not only wonderful, but critical. Controlled vocabulary mediating tools should cover subjects, genres, gazetteers, names and titles, etc."

We have reviewed ALA recommendations and heard witness to the fact that controlled vocabularies are an essential part of our Web work. Now we need to accept that in a networked environment we need to hook together controlled vocabularies to make our systems even more useful.

[Slide 10]
This is exactly what was done by the National Library of Medicine when they mapped over 60 medical and health care thesauri to create the Universal Medical Language System (UMLS)
Pulling all the concepts from these various thesauri together was no small task!
If you wish to learn more about UMLS go to:
http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

Another more recent initiative by LC is the correlations of the LCSH and Library of Congress Classification schedules. They are particularly important since one or both of these systems are used by most American libraries and increasingly by more foreign libraries. To learn more about this initiative go to the home page of LC ClassificationWeb at:

[slide 11 ClassWeb page]
http://classweb.loc.gov

The following slides will be a short demonstration of Classification Web.

[Slide 12 Electronic commerce subj. heading in LCSH]
Here you see the results of a search in the LC controlled vocabulary, that is, the Library of Congress Subject Headings, for the term, "Electronic commerce." Notice the hyperlink to the LC classification number for this heading. The captions, links and hierarchies in the LCSH and LC Classification systems for this concept were developed by a subject cataloger at The Library of Congress.

[Slide 13]
On this slide you see the subject heading and LC classification correlations. Notice that we have the large majority of the works about electronic commerce classed in HF Commerce. The other two classification numbers reflect other places that electronic commerce was classed, namely with a book that dealt with both Internet marketing and electronic commerce and in the KF number which reflects the area of law. This information helps a classifier choose the proper classification number to assign to the Web site he is cataloging. Since 139 works have already been classed in the first HF number he can be fairly certain that this number accurately reflects this topic. However, if he wants to check the HF classification schedule he can click on the class number and he is put into the hierarchy of the HF classification schedule were that concept resides...

[Slide 14]
On this slide he sees the hierarchical arrangement of concepts which complements but differs from the LCSH hierarchies. He is also given a reference for taxation of electronic commerce which refers him to the HJ Public Finance schedule where the the controversial concept of "byte tax" or taxing commercial transactions on the Internet resides. Plus he is given a geographic breakdown for electronic commerce in case his Web site reflects the development of electronic commerce in a particular country or region.
It is very possible that while he is assigning the proper number to this web site on electronic commerce....

a business reference librarian is answering a digital reference question using the new Cooperative Digital Reference Service, a 24X7 worldwide reference service of LC partnering with OCLC with over 100 libraries participating. Let's pretend she received the question from a public library reference librarian in North Dakota who is trying to answer a request for information on electronic commerce made by her local Chamber of Commerce. She has a very small collection of materials in that area, plus her educational background deals with Child development and Children's literature. She chooses to refer the question to the CDRS digital reference service which could use the subject correlations in ClassificationWeb, the CDRS member library profiles and special algorithms written for the CDRS Request Manager. It finds an English speaking business reference librarian in the same time zone with lots of works on electronic commerce in her collection. The business reference librarian searches the concept using the ClassificationWeb software, so she can see the concept and related topics in LCSH, she sees that the concept is present in the database and clicks on the "B" for bibliographic records. She is taken directly to her library OPAC where she views all of the cataloged works.

[Slide 15 OPAC with Electronic resource works]
After reviewing the printed works in her collection and viewing the Web sites on the topic, whose citations are also contained in the OPAC, she selects some excellent sources of information on Electronic commerce. She cuts and pastes the brief citations with the URL for the Web site that was just assigned by her coworker, the cataloger, and includes them in her email response to the requesting library. The Chamber of Commerce of a small town in North Dakota is given some excellent references to printed works and a good Web site which serves their purpose within a very short time.

We just saw how the LCSH and LCC correlations were able to extract some very useful comparative concepts. To date these have always been the intellectual work of the subject cataloger and classifier. In the future we need to capture the intellectual process in concept maps in order to utilize them in AI programs. Speaking of AI, let's quickly search AI programs in ClassificationWeb.

[Slide 16 AI subject search in ClassWeb]
I start by searching the acronym AI.

[Slide 17 AI subject browse]

The subject heading browse allows me to see that there are actually three concepts that use this acronym, Artificial insemination, Artificial intelligence and the name of an extinct city which is currently found on the West Bank. You notice that LCSH disambiguates the concepts by using parenthetical qualifiers. I choose the second one which represents the concept I had in mind.

[Slide 18]
Here you can see the top half of the authority record for Artificial intelligence in its full hierararchy. I scroll down and I am given Artificial intelligence with authorized subdivisions.....

[Slide 19]
Here is Artificial intelligence subdivided by Computer programs. In the hierarchy for this authorized string of headings I am shown a narrower concept, Intelligent agents. It turns out that when I started I had in mind this narrower concept, if you recall, because I mentioned I wanted to search AI programs. With well formulated hierarchical links I was lead to the concept. I click on it and ...

[Slide 20]
am given the authority record in LCSH for Intelligent agents. I note the classification number and then choose to

[Slide 21]
look at the works classed in that number in the OPAC.
I wanted to demonstrate the ClassificationWeb to you not only because it is our newest product at LC but also because it maps concepts across two distinct systems , LCSH and LC Classification, which are used by many libraries. The mapping of concepts is an extremely important part of our work in the network environment. It is not currently, but could also be used in the Request Manager of the CDRS system.
The LC Network Development and Standards Office have also been hard at work creating mappings of our traditional systems and the new ones. For example, they have mapped:

[Slide 22]

The Dublin Core's 15 elements have been mapped to MARC and viceversa which makes it easy to share metadata across many types of metadata records. Likewise MARC has been mapped to XML in order to translate into the latest Web markup language. If you are interested in seeing how this works just go to: http://www.logos.com/marc/default.asp
and paste in any MARC record to convert to XML.

[Slide 23]
Specialized systems of mappings are also important to particular fields of learning. For example, the Standard Industrial Classification system was developed under the Commerce Dept. and maintained by Census. With the advent of NAFTA and the Web there were many changes to this classification system and to new high tech industries which were not reflected in the Standard Industrial Classification system. It was necessary to create a new coding system called the North American Industrial Classification System or NAICS. The SIC and NAICS have been mapped in order to provide a classification system and continued access to new types of industries. The advantage of a numercial classification is that it is not language dependent. It can be used by the French Canadians as well as the Mexicans.

[Slide 24 screenshot of NAICS/SIC codes]
Here we see the crosswalks or bridges of the Standard Industrial Classification codes and the North American Industrial Classification System found at: http://www.census.gov/epcd/ec97brdg/
These industrial codes are used by every U.S. business and many foreign businesses to self-identify the type of industry to which they belong. They are formulated with numbers which represent a hierarchical relationship and use many cross references to guide the user.

[Slide 25 for SIC code example for COTS software industry]
This slide represents the changes for the Commercial off-the-shelf or COTS software industry sector between the SIC codes and the NAICS codes which were officially accepted in 1997.

Anyone who has done research for investment in a company should be familiar with SIC and NAICS codes since they are used by the majority of the investment databases. According to a Security and Exchange Commission spokesman, the S.E.C. assigns codes to a company if none are given in the S1 and S4 filings. He also stated that IRS and Labor Dept. both require the Standard

Industrial Classification codes for their information and research studies.[3] Increasingly the NAICS codes are being incorporated into the systems of business information providers. They will likely become more prevalent in the future. This is one example of how information exchange is becoming increasingly more prevalent and useful on the Web to government agencies and commercial firms alike.

Now we will cover a new area of knowledge organization on the Web. Customized search engines using taxonomies are the foundation for maturing Web portals or Enterprise information portals (EIPs) used by businesses today.
To continue with the metaphor of the Web as an ocean of information and the need to filter salt water to make bottled water. This is where we make water into wine!

[Slide 26 Water into wine] Businesses are finding that in order to get their web portals organized they need to use taxonomies. So how are they defining taxonomy?
[slide 27 Definition of taxonomies]
Taxonomies are high level information search devices constructed to provide a means of understanding, navigating and gaining access to intellectual capital. They are tools for information management and discovery devices for knowledge management.

Taxonomies are not new. Man has been trying to put order into his world ever since his first attempt to understand nature. The written word allowed man to express his thoughts and concepts. So intellectual assets were born. With more thoughts and concepts recorded there was a need to organize the written works.

[slide 28 History of taxonomies, pictures of AlexLib, Aristotle and Linnaeus]
So great thinkers, such as Kallimachos (305-240 BC) of the Alexandria Library in Egypt, put themselves to the task of devising systems and libraries. Aristotelian classification of knowledge was the basis for medieval philosophy and science. Modern scientific taxonomy progressed with the writings of Linnaeus (1707-1778) who developed a system for classification of biological organisms with its familiar breakdown of kingdom, phylum, class, order, family, genus, species.

---

[3]According to a SEC spokesperson, the Security Exchange Commission does not use the NAICS codes as of 3/01. The SEC accommodates new industries, such as cellular telephones, by classing them with older telephone technologies in 4813.

[Slide 29 the word "classification" used more than taxonomy]

Dictionary and encyclopedic definitions of taxonomy still refer primarily to biological taxonomy with classification being a synonym. Taxonomies or classification systems exist for every conceivable discipline and topic. When I did a search of The Library of Congress online catalog I discovered that the words "taxonomy" or "taxonomies" was generally used for biological and other science disciplines whereas the word "classification" was used by all fields.[4] There were many fields represented including physical anthropologists who classify mankind, geologist who classify rocks, soils, etc., biologists who classify plants and animals, medical personnel who classify diseases, etc. Librarians, on the other hand, have inherited the work of classifying the entire world of knowledge.

[Slide 30]

We know that numerous formal taxonomies are maintained by government and commercial enterprises as knowledge organization tools.

[Slide 31]

Taxonomies are use in customized search engines and they are used as an integral part of web portals to enhance resource discovery.

[Slide 32 screenshot of homepage of CBDnet] http://cbdnet.access.gpo.gov

An example of a government taxonomy is found in the CBD. For years the Commerce Business Daily was only in print. On the screen you see the Web site equivalent. CBD is an invaluable source of information for U.S. government procurement invitations, contract awards, subcontracting leads, sales of surplus property and foreign business opportunities. It is now online thanks to the award winning alliance between the U.S. Department of Commerce and the U.S. Government Printing Office. Commerce has management responsibility for the CBD and GPO provides the day-to-day operation of CBDNet.

[Slide 33 screenshot of the CBD alphanumeric codes]

On this slide you see the taxonomy used are alphanumeric codes which are representative of the categories used in the original print version. The alpha codes represent Services and the numeric codes represent Supplies.

---

[4]The Library of Congress Online Public Access Catalog. Search preformed on 2/20/01. Taxonomy and taxonomies as keyword in titles=681 hits. Classification as keyword in titles=6690 hits. These statistics do not include works in foreign languages. A subject search for classification produced over 10,000 hits in the catalog.

[Slide 34 screenshot of the CBD alpha codes with categories]
Here you can see the alpha codes (only) with their corresponding categories for Services. This taxonomy works for the users of this Web site and its archive.

Taxonomies are living structures that must grow and change. The maintenance of them can be a labor of love.

Another example of a taxonomy used in a Web search engine is....

[Slide 35 screenshot of NorthernLight.com homepage]
NorthernLight.com built the foundation of its taxonomy in 1995 and launched its service in 1997. NorthernLight.com has mapped existing controlled vocabularies such as the Library of Congress Subject Headings , the Medical Subject Headings and others to formulate a tightly woven knowledge base. An artificial intelligence system uses these controlled vocabularies to classify pages for this search engine. A combination of automatic indexing using the controlled vocabularies with ever vigilant librarians as the human component is what make this search engine valuable.[5]
Notice the statement on their initial page under Special Editions

> "In-depth coverage of major news stories,
> edited and compiled by our team of
> librarians and updated weekly"

[Slide 36 screenshot of results from a search with folders]
Here you can see the interface technique used is "folders" which disambiguate the search. The user has a visual presentation of the various aspects of the topic he searched.

Needless to say we could review thousands of taxonomies both large and small from informal lists of words to formal controlled vocabularies and classification systems representing every conceivable domain or field.
The basic principle to remember is that the more tightly coupled the taxonomic relationships are defined, the more efficient and focused the data retrieval will be.

---

[5]Ward, Joyce. Indexing and Classification at NorthernLight. Presentation given at the CENDI conference on Controlled Vocabulary and the Internet, Sept. 29, 1999.Can be found at: http://www.dtic.mil/cendi/presentations/ward.ppt Also, spoke with Lynn Wojcek, NL spokesperson via telephone March 2001.

[Slide 37]
How do we Define Taxonomies in a Wired World?

Taxonomy: A classification of elements within a domain
- Domain: a sphere of knowledge, influence, or activity
- Classification: the operation of grouping elements and establishing relationships between them (or the product of that operation)
- Relationships: a defined linkage between two elements
- Element: an object or concept[6]
-

[Slide 38]
What are Taxonomies Good For?

Taxonomies are applied to:
- Items (aka resources) individual pieces of information (documents, people...

By the use of:
- Metadata: (aka properties, attributes) information describing types of data

Which may or may not use values from a:
- Vocabulary: selection of terms, classified or sorted

To create:
- Content: an item and its associated metadata

[Slide 39]
What are some of the Challenges to Creating Taxonomies?

The major challenges connected to structuring content and providing subject access are:
- Information management across divisions of your agency
- Agency global intranets/Internet portals
- Global or national document management including technical documentation
- Incorporating taxonomy technology into agency technology +info. policies
- Cost of building a taxonomy
- Moving a taxonomy from overhead to being a core part of your agency's information management.

---

[6] Crandall, Mike."Taxonomies for the Real World: The Business Imperative to Simply Content Access" TFPL Taxonomies for Business Conference, London, Oct.23, 2000.

[Slide 40 More challenges]

- Certification of the taxonomy by an authoritative body,
- Finding common ground across multiple taxonomies or schemas with similar terms and different meanings.
- Ensuring the ongoing integrity of the taxonomy with constant maintenance.
- Acceptance by developers of tagging tools.
- Integrating with a legacy system and external content.

[Slide 41]

What core expertise is required to create a Taxonomy for the Web?

- Systems Analyst who understands specifications for creating taxonomies
- Domain expert/Subject expert in the area of the taxonomy
- Computational linguist, Artificial intelligence/Knowledge engineer
- Linguist of human languages/Lexicographer
- Database/Application Development Expert
- Administrative Support
- Review Support

A librarian could fit in a number of these positions depending on his background.

[Slide 42 Custom taxonomy in XML]

This is an example of a custom taxonomy written with its own dialect of XML. The American Institute of Certified Public Accountants (AICPA), joined by Reuters and thirty-some other financial organizations created an XML-based specification for the preparation and exchange of financial reports and data.

[Slide 43]

This screenshot gives you the location of the XBRL taxonomy viewer which you can download. I don't want to take the time to review this tool but it is worth looking at.

[Slide 44]

Some Recommendations to keep in mind are:

- Actively seek out existing taxonomies in the target discipline or subject area. If your needs are met in part by an existing taxonomy use it and build on it.
- Look at the intended purpose of the taxonomy and select appropriate software tools.
- Consider scalability of the taxonomy. Look at the big picture and see how the taxonomy will be able to hook into others.

- Consider utilizing numerical taxonomy as a schema in the metadata in order to merge documents in foreign languages. Your current audience might be your agency and clients, however, the entire world looks at your Web presence. Consider planning for multilingual access since English will not be the predominate language on the Web in a few years.
- Accommodate new standards whenever possible.
- Document “Best Practices” while creating the taxonomy and review them regularly.
- Maintain and update the taxonomy continually
-

[Slide 45 Flowchart of creating a taxonomy utilizing metadata]

This flowchart represents the underpinnings of the completed process. Web content providers and designers should be proactive in finding developed taxonomies and utilizing the framework of the hierarchies which have been established and used for providing access to information in the target subject area.

[Slide 46]

Efficient Web information retrieval systems in the form of search engines or Web portals require continued support and improvement of:

[Slide 47]

Web-based classification and numerical taxonomic tools to use in....
Web-based cataloging tools such as CORC, which provides metadata based on....
Taxonomies such as controlled vocabularies/thesauri which will be hooked together using...
Metathesauri and standard information exhange systems such as MARC-XML

[Slide 48]

And this is the house that Jack built.....with a wine cellar
To end with our water into wine metaphor!

[Slide 49:Presentation title]

Thank you for you kind attention.
Hopefully you will build a house with a wine cellar in your governemnt agency or business organization!