Workshop on Thesauri and Metadata
July 26, 1997
Notes

Workshop Leader: Joseph Busch, Program Manager, Getty Information Institute.
Notes submitted 9/10/97, by Jane Greenberg, Teaching Fellow, University of Pittsburgh, School of Information Sciences. Student volunteer for workshop.

I. INTRODUCTION

Joseph Busch introduced the workshop and indicated that he wanted all participants to contribute. Joe also stated that different points of view were welcomed and that participants should make note of the questions which ought to be part of the larger research agenda. Joe then introduced the workshop participants, workshop goals, and workshop plan (this information follows).

A. Workshop participants:
1. Individuals who responded to the initial call for presentations
2. Invited participants
3. Individuals who registered as part of the DL conference

B. Workshop goals:
1. To gather information about research and products related to thesauri
2. To investigate how metadata elements need to be completed
3. To come up with a list of key issues, questions, concerns, etc. focusing on the use of thesauri as metadata content tools

C. Workshop plan:
1. Part 1 was to focus on using thesauri for searching metadata.
2. Part 2 was to focus on using thesauri for naturally generating metadata.

The Dublin Core (DC) was introduced along with a series of questions about the DC and general metadata standards. These questions follow:
1. Does it matter if the terminology source is qualified in the DC?
2. How can we take advantage of thesauri when using the DC?
3. How useful are the DC element qualifiers?
4. What terminology guidelines are needed for metadata standards in the digital environment?
5. How might terminology guidelines be used with the DC or other metadata standards?
6. What enumerated lists are needed for the DC or other metadata standards?
7. Is metadata needed for terminology resources?
8. What mechanisms are needed for extending/sustaining terminology?
9. How are we going to deal with terms in the digital world? (centralized? decentralized?)

II. PRESENTATIONS

PRESENTATION 1: *Thesauri for Knowledge-Based Assistance in Searching Digital Libraries*, Dagobert Soergel, College of Library and Information Services, University of Maryland, College Park.

Soergel introduced himself as an ontological engineer. His talk outlined, in detail, key metadata challenges and highlighted the role thesauri can play in the digital library environment. The challenges identified both technical/system oriented and human problem solving issues. Soergel discussed how thesauri can assist with learning and assimilating information, assist researchers and practitioners with problem clarification, support information retrieval by providing a knowledge-based support of end-user searching, support information display (especially the presentation of results), provide a tool for indexing, facilitate combinating multiple databases or unifying access to multiple databases through a variety of mechanisms, and support document processing after retrieval. Discussion following Soergel's presentation reiterated that thesauri are more than a tool for indexing. Questions were raised about how indexing languages can be used for capturing the interest of various user communities. (The points listed above were taken from Soergel's presentation overheads, which provided greater detail).

PRESENTATION 2: *Conceptual Searching Using Thesaurus-Based Metadata*, Luray M. Minkiewicz, DuPont Central Research and Development, Corporate Information Services.

Minkiewicz introduced DuPont's online Thesaurus and SCION, a database for proprietary technical and business information. The thesaurus' role as an indexing and searching tool was reviewed, and indexing cost realities were highlighted. Minkiewicz discussed the challenge of getting DuPont employees to index their work with the thesaurus. Discussion following Minkiewicz's presentation focused on getting document producers or lay persons to do their indexing (that is the common Joe (pardon!) who is not trained in indexing). Several workshop participants expressed concern over lay person access to thesauri for indexing. One participant indicated that the lay person will assign the worst index term. Another participant asked if the librarian/indexer will take on a new role in helping the lay person index his/her work in the digital environment. The topic of a mega-thesaurus, which would map a number of different thesauri, was also raised. (Minkiewicz handed out copies of her overheads as well as off-prints of an article on the DuPont Global Technical Information System).

PRESENTATION 3: *The Use of Subject Vocabulary in a Distributed Search Environment*, Ron Davies, Bibliomatics Inc., Ottawa, ON, Canada.

Davies provided an overview of UNIVOC, a hierarchically structured controlled vocabulary that is being developed to provide access to UN system-wide information available over the Web. The stockholders (diplomats, information providers, general public) and the thesaurus applications (classifying agencies, restricting searchers, describing services, and describing resources) were discussed in relation to the thesaurus. Davies put forth a number of metadata questions/issues that have emerged from his UNIVOC work. For example, he asked, Do we index using one or all languages? What is the role of the search engine in the recognition and use of metadata?; and, What future activities will insure interoperability? Goals outlined in Davies' conclusion included testing UNIVOC with practical applications, adapting it to other languages, enriching its entry vocabulary, and expanding it into new topical areas.

ABSTRACT from Ron Davies:
UNIVOC is a hierarchically structured controlled vocabulary designed to improve access to UN system-wide information available over the Web. Developed on behalf of the UN Information Systems Coordinating Committee, it is currently being tested for use in the subject description of selected UN sites.

Traditionally, controlled vocabularies have provided users with a consistent terminology to be used in searching for a particular document. However a controlled vocabulary such as UNIVOC can also be applied to assist users in selecting sites to search in a distributed system, to help search front ends or search clients direct searches to particular servers, or to even to create "virtual servers" on a particular broad subject topic based on information indexed by different local servers. In addition, the multilingual nature of the UN subject scheme allows possibilities for expanding accessibility in other languages, even for documents which are written in, or indexed in, English only.

The testing of UNIVOC will focus primarily on its use with UNIONS, the UN-system wide search system. However the development of such vocabularies highlights the urgent need to define common standard services for accessing information in thesauri and controlled vocabularies so that terms and term relations can be used by the different search engines providing access to Internet-available information.

PRESENTATION 4: *Applications of the Art & Architecture Thesaurus (AAT) to Resource Discover, Searching, and Navigation*, Vivian Hay, Getty Information Institute, Los Angeles, CA.

Hay's talk introduced and discussed in some detail three of the Getty's tools: Thesaurus for Geographic Names (TGN), Union List of Artist Names (ULAN), and the AAT. A chart outlining information about each of these tools was reviewed. The chart covered details, such as, entry terms, variants, terms/name source, types of encoded relationships (links), and so forth. Challenges in constructing, testing, and publishing these three tools were discussed.

PRESENTATION 5: *Metadata for a Dynamic Networking Learning Environment*, Terry Zimmerman, University of Texas, Austin, TX.

Zimmerman's presentation focused on his experiences in developing a metadata architecture for an education-oriented digital library environment. Zimmerman's work deals with K-12 cirriculum, and is part of a commercial, governmental, and academic joint venture project (For confidentiality reasons, specific team players and other information could not be revealed). Zimmerman stressed that content, context, and integrating the community and its resources are all key aspects of the project. Zimmerman raised a number access management questions about individual/corporate rights, quality of materials, and personal security in the digital environment. Zimmerman indicated that the project needs to provide an flexible organization scheme, one which is adaptable to a variety of clientele’s needs. The applicability of various DC elements (e.g., other contributor, date, and resource type) was discussed, along with the need for metadata to support a rich set of relationships among curriculum bearing objects and teachers.

PRESENTATION 6: *Thesauri and Metadata in the Alexandria Digital Library Project*, Linda Hill, University of California, Santa Barbara, CA.

Hill described the Alexandria Digital Library Project, a georeference information system in which every item is represented by a geographical footprint that describes the area that item is about. Hill explained that the georeference information system functions as a gazetteer, and that place names can be described by latitude and longitude coordinates. The thesaurus-like features of linking place name to footprint and vice/versa were demonstrated, and the project’s open contribution status was reviewed. The presentation ended with a discussion of possibilities for georeferencing, such as linking geographical data to MARC records in library catalogs.

PRESENTATION 7: *Automated Classification for Networked Information Retrieval*. Rob Dolin, University of California, Santa Barbara, CA.

Hierarchical collection metadata lends itself very nicely to scalable source selection (106 sources). The talk discussed the process of automated classification.

III. QUESTIONS AND ISSUES:

Throughout the workshop, key issues and questions were marked on a master list. The following is an enhanced version of the original list.

1. What are the functions of thesauri?

2. What issues surround thesaural interoperabiltiy standards? What work has been completed/or is needed to ensure interoperability?

3. What research is needed about thesauri on a practical use level? Is research needed? [We don’t understand how people look for information, how they might look for information, or if they look for information in the most useful way].

4. Who should/is creating metadata? [Joe pointed out that it is becoming a democratic act. Is this dangerous?]

5. Where do/how do registries and various other schema fit in with the topic of thesauri? What potential does Z39.50 offer for thesauri? Could a mechanism be developed where an indexer (even the untrained indexer) could submit a term and retrieve a definition before assigning it as a descriptor?

6. What profile functions for thesauri need to be considered at the system level?

7. How can syntax for multiple languages be successfully dealt with?

8. What tools exist/are needed for facilitating description?

9. What is the potential of automatic classifiers in relation to thesauri?

10. How can we get search engines that recognize core metadata?

11. There is a need develop trusting relationships between creators and users. (This relates to Zimmerman's work noted above).

12. What kind of thesaural guidelines are needed for special user communities (e.g., instructional materials (K-12).

13. How should we deal with hierarchical metadata. Metadata itself is fairly hierarchical.