NSF Project: Harvesting Information to Sustain our Forests

Lois Delcambre
Computer Science Department
Portland State University
Portland, OR 97225 USA
+1 503 725-2405

lmd@cs.pdx.edu

Marianne Lykke Nielsen
Dept. of Information Studies
Royal School of Library & Info.
Science
Denmark
+45 98 90 18 54

mln@db.dk

Timothy Tolle
Consultant
Portland, OR
+1 360 573-2291

timtolle@starband.net

Mathew Weaver
Computer Science & Eng. Department
OGI School of Science & Eng.
OHSU; Portland, OR 97006
+1 435 245-9354

mweaver@cse.ogi.edu

David Maier
Computer Science Department
Portland State University
Portland, OR 97225 USA
+1 503 725-2406

maier@cs.pdx.edu

Susan Price
Computer Science Department
Portland State University
Portland, OR 97225 USA
+1 503 725-4036

prices@cs.pdx.edu

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: Systems Issues, User Issues.

General Terms

Management.

Keywords

Thesaurus model, controlled vocabularies.

1. INTRODUCTION

This project seeks to build a domain-specific digital library for use in natural resource management in conjunction with our partners, the USDA Forest Service, the USDI Bureau of Land Management, and the USDI Fish and Wildlife Service. We focus on the documents in use in natural resource management such as *Decision Notices*, *Environmental Assessments*, *Environmental Impact Statements*, and the many documents that provide the detail behind these documents. Our contributions include a generalized thesaurus model (with it's associated software) called Metadata++ that accommodates the extensive terminology used in this fundamentally multi-disciplinary field. Key features of the Metadata++ model are: (1) terms can appear in multiple locations in the broader term/narrower term hierarchy (*polyhierarchies*), (2) terms are distinguished by their path within the hierarchy, and (3) multiple terms that are used interchangeably in a given context can be represented as a single node in the hierarchy (*polyterms*).

We have explored a series of implementation strategies for our comprehensive collection of terms, with more than 100,000 terms in the current system [1]; we have developed a software architecture, with an associated implementation that allows the Metadata++ digital library system to interact with a GIS system to facilitate the search for documents that pertain to a specific location or place [2]; and we have recently completed a usability test [3,4] with natural resource managers at two national forests in the Pacific Northwest. In this paper, we briefly describe the Metadata++ model and system, note the highlights of our usability test, and mention our plans to exploit the Metadata++ model to assist the users when indexing documents.

2. The Metadata++ Model and System

The terminology of interest in this application domain spans a number of subject areas, as shown in Figure 1, and includes terms from a large number of well-established, well-known classification schemes and terminologies used by multi-disciplinary experts including soil scientists, fish biologists, wildlife biologists, fire specialists, and hydrologists. Most of the terms come from existing sources – published glossaries, terminologies, and taxonomies. These specialists frequently use the same terms with often subtle (and occasionally significant) differences in meaning. One of the primary objectives of the Metadata++ digital library is to faithfully represent and utilize multiple controlled vocabularies from a variety of discourses within the natural resource management domain.

The Metadata++ system provides a number of features that allow users to interact directly with the terms and their associated paths, as listed here.

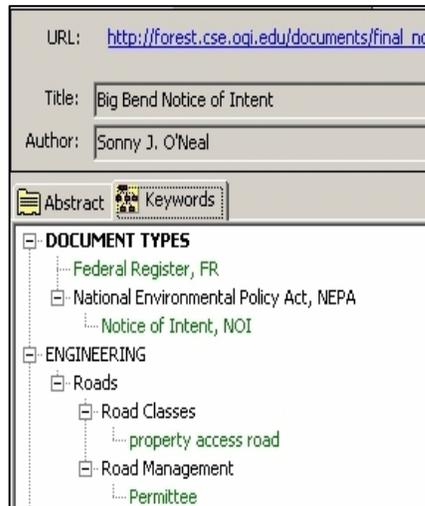**Figure 1**: Top-level Subject Areas

Figure 2: Path-based keywords

- Users can browse the terminology directly by browsing the broader-term/narrower-term hierarchy using the "+/-" convention for expanding and contracting a node in the *Browse* window.
- At any point, the user can easily see all multiple occurrences, as well as related terms and synonyms, of a term simply by right-clicking on the given term. In addition to the related terms, right-clicking on the selected term will also display all *explicitly referenced documents* (documents for which a human indexer selected the path-based term as a keyword) and *implicitly referenced documents* (documents that contain the term based on computer automated indexing).
- The user can search for terms at every level in the hierarchy, including the use of wildcard matches, using the *Find Term* function. Each term found is displayed with its complete path.
- The user can double-click on any term in the *Find Term* window in order to see it in context, within the full hierarchy of terms in the *Browse* window.
- The user can select path-based terms, to index a document or to formulate a search, in either the *Find Term* or the *Browse* window and drag it to the document or to the search window.

The path-based term allows the user to understand the meaning or connotation intended for that particular term. Figure 2 shows keywords that a human indexer selected for a particular document.

3. Usability Test

In our usability tests, we wanted to determine if domain experts could easily understand and use the Metadata++ path-based

terminology. During the various tasks of the usability test, participants encountered a total of 657 terms with multiple occurrences. The majority of these terms (78%) had exactly 2 occurrences, but several terms had 3 or more occurrences. The maximum number of occurrences for a single term was 27. When presented with multiple occurrences, participants knowingly selected the paths that were most appropriate for the particular task. In some cases, they selected a path, then later removed the path and selected a different occurrence of the same term – one that more accurately described the concept that they wanted to express.

During the user tests, participants encountered 120 polyterms; with about two-thirds of the polyterms consisting of the full term and its acronym (e.g., *Watershed analysis, WA*), and the other poly-terms were interchangeable sibling terms (e.g., *buoy*, *mooring pile*). Participants easily understood the polyterms they encountered.

4. Extensions

Based in part on our collaboration with Nielsen and Price (M.D.), we are exploring the use of Metadata++ in medical terminology and we are beginning work to suggest index terms based on the structure of the Metadata++ terminology.

5. Acknowledgements

6. REFERENCES

[1] M. Weaver, L. Delcambre, T. Tolle, Metadata++: A Scalable Hierarchical Framework for Digital Libraries, In *Proc. International Symposium on Digital Libraries and Knowledge Communities in Networked Information Society (DLKC '04)*, (Tsukuba, Ibaraki, Japan, March 2004).

[2] M. Weaver, L. Delcambre, L. Shapiro, J. Brewster, A.Gutema and T. Tolle, A Digital GeoLibrary: Integrating Keywords And Place Names, In *Proc. of the 7th European Conference on Research and Advanced Technology for Digital LIbraries (ECDL 2003)*, (Trondheim, Norway, August 2003).

[3] M.L. Nielsen, L. Delcambre, T. Tolle and M. Weaver, Indexing and Retrieval Challenges in Digital Government Systems - Summary of an Empirical Research Project. *2nd Scandinavian Workshop on eGov*. (Copenhagen: CBS. January 2005).

[4] M. Weaver, L. Delcambre, T. Tolle, M.L. Nielsen, Using a Path-Based Thesaurus Model to Enhance a Domain-Specific Digital Library. Submitted to *European Conference on Digital Libraries* (Vienna, Austria, 2005).