Using Controlled Vocabularies as a Knowledge Base for Natural Resource Managers

Tim Tolle[1], Patty Toccalino[2], Lois Delcambre[3], Julia Norman[2], Fred Phillips[4], David Maier[3]

[1]Strategic Planning, Region 6, ttolle@fs.fed.us, USDA Forest Service

[2]Department of Environmental Science and Engineering, toccalino@ese.ogi.edu
[3]Department of Computer Science and Engineering, {lmd, maier}@cse.ogi.edu
[4]Department of Management in Science and Technology, fphillips@admin.ogi.edu
OGI School of Science & Engineering, Oregon Health & Science University

1 Introduction

Figure 1: The ten Adaptive Management Areas

This project seeks to provide access to forest information for both its producers and consumers, e.g., local, state, and federal agencies and many other parties. This project focuses on the Adaptive Management Areas of the Pacific Northwest, ten areas of forestlands set aside for experimentation in all aspects of forest management (Figure 1). The goal of the project is to build a web-based, forest information portal that preserves the autonomy and local focus of each Adaptive Management Area.

2 Overview of the Project

Figure 2 shows the overall structure of our project. The boxes at the top of the figure represent the engagement of the natural resource managers in careful consideration and articulation of the focus for this project including the target user(s) and the information content of most importance. Much of this work has been conducted by the USDA Forest Service and other agencies involved in the Adaptive Management Area Program (indicated with yellow shading). A graduate student in Management in Science and Technology joined with Eric Landis, a consultant to this project, to conduct extensive interviews and documentation of user scenarios involving searching for information. As shown in the figure, this extensive requirements gathering effort led to our choice of targeting the needs of the natural resource manager. Some of the most important information requests for natural resource management include:

a. learn what is known about a certain place, be it a watershed, a county or an ecological province or zone.

b. learn what is known about places that have a similar climate, topography, or forest type as the place in question. Often documents for specific places do not exist but do exist for places that are similar, in one or more key ways, to the place of interest.

c. find documents about specific forestry topics, such as forest type, plant associations endangered species, or rural communities and public participation.

d. find documents about specific management activities such as planting, pruning, inventorying, monitoring, researching, or assessing.

1. National Resources Conservation Service PLANTS Database; http://plants.usda.gov/plants/

e. find documents about recreation activities, such as fishing, camping, and hiking.

f. find documents written by specific people, e.g., to answer questions such as "I wonder what Jerry Franklin is working on?" or "I wonder what he has to say about this situation?"

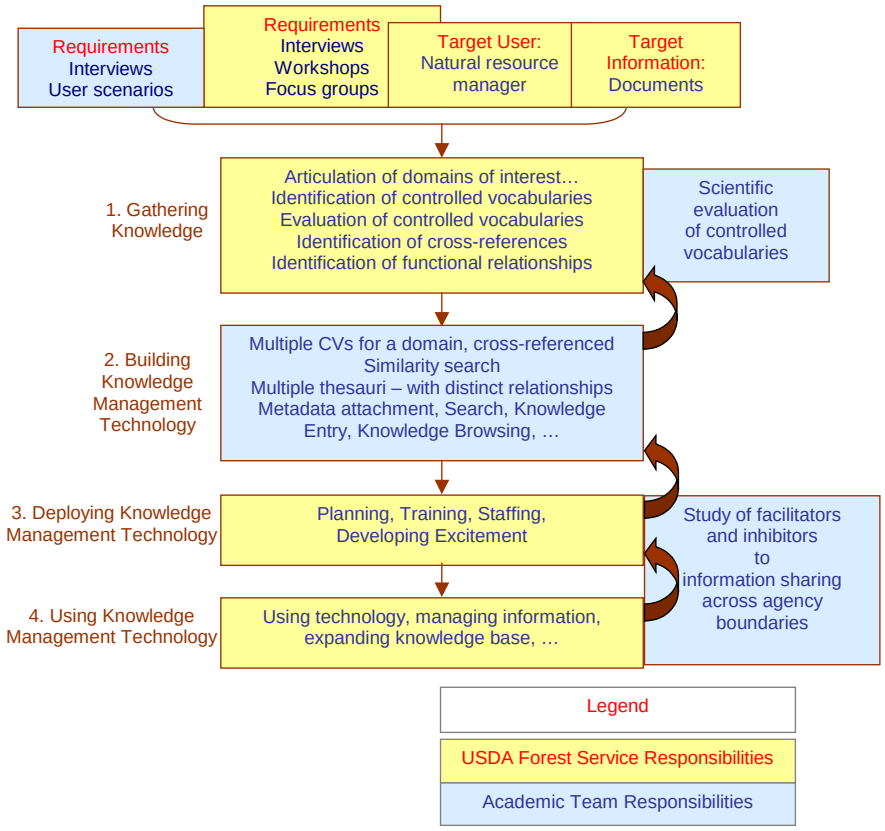

Figure 2: The Knowledge Management Cycle

In order to meet these needs, we adopted a knowledge-based approach with two equally important kinds of knowledge:

(1) **Documents**. The USDA Forest Service produces various documents containing knowledge insomuch as they convey assessments and judgments. These assessments and decisions articulate the impact or severity or "so what" of any inventory, survey or analysis. The credibility of these documents derives from the subject matter expertise, scientific practice, and interpretation of values of the natural resource managers who prepare these documents. The goal of this project is to reuse this fundamental source of agency knowledge.

(2) **Terms and their relationships**. Based on significant engagement and intellectual effort of natural resource managers across various agencies and organizations involved in this project, we identified approximately twenty-six domains of interest such as: location, climate, vegetation, wildlife, hydrology, and so on. For each domain, we are identifying, evaluating, and selecting controlled vocabularies to place in the automated system. The basic approach is to facilitate the attachment of terms from the controlled vocabularies to documents and to provide a rich search and browsing capability that exploits the terms and structure in the controlled vocabularies. In addition to terms in the controlled vocabularies, we include structured relationships including broader-term-narrower-term, synonym, and general, functional relationships.

1. National Resources Conservation Service PLANTS Database; http://plants.usda.gov/plants/

Thus, the main activity of our project is now focused on the four, numbered steps in Figure 2. The box labeled "1. Gathering Knowledge" represents the considerable effort invested within the agency to identify, consider, and, in some cases, invent controlled vocabularies of interest to describe the domains. The box to the right, shaded in blue, represents the efforts of Environmental Science and Engineering graduate students, led by Dr. Patty Toccalino, to evaluate some of the controlled vocabularies from a neutral, scientifically informed point of view.

The next box ("Building Knowledge Management Technology") represents the computer science work to define and build the technology for this system.

The box labeled "3. Deploying Knowledge Management Technology" is the responsibility of our USDA Forest Service – to introduce the technology into the workplace and to provide appropriate management and policy guidance so that the technology can be sustained over the long term.

Finally, the box labeled "4. Using Knowledge Management Technology" recognizes the continued evolution of the knowledge and the technology as well as the practical issues associated with its use. A box on the right of box 3 and 4 represents a companion project that is studying the inhibitors and facilitators of information sharing, with a focus on the interaction across agency boundaries, led by Drs. Nicole Steckler and Marianne Koch.

In this paper, we focus on the structure of the controlled vocabularies and cross-vocabulary relationships that provides a rich description of the important concepts in natural resource management and also serves as the conceptual framework to attach and retrieve documents. The paper concludes with a brief discussion of the technical challenges presented by this project.

3 The Controlled Vocabularies

The first part of step 1 in Figure 2 required the identification of the domains of interest to the natural resource agencies, especially the Forest Service, Bureau of Land Management, Park Service and Fish and Wildlife Service. The agencies identified twenty-six domains of interest, including vegetation, location, climate, wildlife, and so forth. Each domain has one or more controlled vocabularies (CVs) associated with it. CVs are composed of terms that are commonly understood among targeted portal users. Existing terminology standards, as appropriate, have been utilized. We are also exploring keyword lists for describing database or WWW-site subject content in a controlled way, thesauri, classification schemes, and glossaries and other reasonable sources of terms in the domains of interest.

3.1 The Controlled Vocabulary Evaluation Process

Identification of candidate controlled vocabularies followed these steps:

- Initially, the domains of interest were identified from a) agency program areas, b) elements of the ecosystem, and c) the Dublin Core.
- Subject matter experts searched the web and sources with which each was familiar to locate key word lists, glossaries and indexes.
- Then, criteria were applied which had been developed and tested on one controlled vocabulary, names of plants, selecting the terms for use in that specific CV.
- Third, relationships between and among terms were identified initially from the literature by the subject matter experts.
- More relationships are being identified daily by using the terms which define the initial terms, if glossaries or data dictionaries are used to identify terms.

1. National Resources Conservation Service PLANTS Database; http://plants.usda.gov/plants/

- Last, terms and the relationships are expressed in XML within the metadata ++ model.

Each CV recommended for consideration in the forest information portal is being evaluated according to ten primary criteria as shown in Figure 3. After the USDA Forest Service staff reviewed and approved the criteria, the evaluation of the CVs, beginning with the vegetation CVs, began.

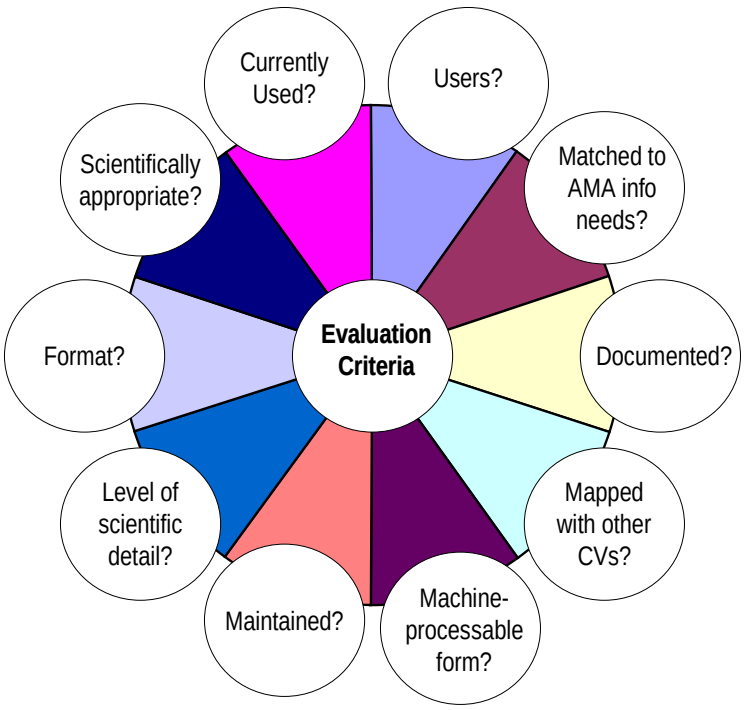

Figure 3. Criteria used to evaluate CVs

We determined how often each CV is used, when it is planned to be used, and by whom. There are a variety of user categories to consider, including the Forest Service, Bureau of Land Management , researchers, the public, etc. We ensure that each selected CV is well matched to the information needs of the Adaptive Management Areas and that each CV includes documentation such as bibliographies of contributing sources, a description of the development of CV terms (e.g., via a committee), etc.

Then, we determine whether each CV is "mapped" or cross-referenced with any other CV(s); this mapping may include simple or complex associations such as using terms from other accepted glossaries. When CVs exist in machine-processable form such as delimited text documents, XML, etc., incorporation of the CV into the portal is facilitated. It is desirable to use CVs that are maintained, kept up-to-date, and that have someone with authority over the CV. The level of scientific detail and validity of each CV is evaluated, and the format of each CV (such as a glossary, hierarchical thesaurus, etc.) is identified. We also determine whether each CV is appropriate to use on a scientific basis.

A subset of these criteria, listed approximately in order of importance, allows us to judge the scientific appropriateness of CVs that contain scientific information.

- Quality of the data sources used in the CVs. Data from peer-reviewed literature are preferred, but other data sources may include agencies such as botanical societies, Nature Conservancy, etc.
- Level of documentation. Complete bibliographies of contributing sources are preferred.
- Audience to whom the CV was directed. The intended audience (scientists, land managers, public, etc.) provides some indication of the level of scientific detail included in the CV.
- National or international recognition of the CV. Widely used and recognized CVs provide an indication of the scientific validity of the CV.

1. National Resources Conservation Service PLANTS Database; http://plants.usda.gov/plants/

- Amount of scientific detail and usefulness of information to scientists. This criterion may be subjective, but includes a review of the scientific terms and definitions used in the CV, and a general estimate of the usefulness of the CV to scientists.
- ·Status of data review. CV data that are fully reviewed are preferable. In some cases, data are currently undergoing review or have not been reviewed.

3.2 The Relationships in the Knowledge Base

The use of carefully selected CVs allows the users of the forest information portal to enhance their searches for relevant information. These enhancements promote browsing for terms and subjects that are related to their initial inquiry. Several types of relationships exist among or between words and groups of words, and these relationships promote browsing, in the sense of one browsing at a library. These relationships are important for automatically augmenting the user's search (e.g., by adding synonyms) and to support browsing in the traditional library shelf sense. The use of relationships in the system will also inform searchers as to why various documents might result from their search. The three kinds of relationships supported in the knowledge base, among terms from the controlled vocabularies, are: broader-term/narrow-term, synonym, and functional relationships. These are discussed in turn.

Broader-Term/Narrower-Term
The intuitive nature of the broader-term/narrower-term relationship can be conveyed using spatial relationships. A county (narrower term) is located within a state (broader term) that is located within a country (broader term) as shown in Figure 4. Taxonomic relationships, such as a particular plant species (narrower term) being part of a genus (broader term), is also represented using a broader-term/narrower-term relationship as shown in Figure 5.

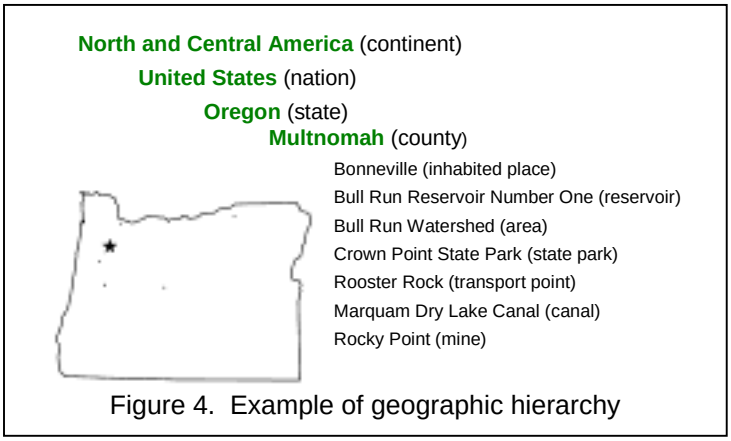

Figure 4. Example of geographic hierarchy

Synonyms (Terms Used Interchangeably)
The synonym relationship is obvious, but particularly powerful. For example, one can look for a plant with its common name (e.g., Douglas Fir), and find scientific documents that were key worded by the scientific name (e.g., *Pseudotsuga menziesii*) alone, as shown in Figure 5. The correlation between one common name and two scientific names is also of interest. This correlation can happen for any number of reasons, but most commonly, when one species is mistaken for another and that document is coded by the mistaken identity.

A more interesting case happens with competing taxonomies. For example, the U.S. Environmental Protection Agency (USEPA) and the Forest Service use different land classification systems. When people familiar with the Forest Service classification system, but not the USEPA system, enter documents and their meta-data, the documents could be lost to those who are only familiar with the USEPA classification system. By noting the correlations between and among terms from different classification systems 'behind the scenes,' a user can find documents that were not key worded with words familiar to that user. Thus we are identifying and exploiting cross-reference schemes using the synonym relationship.

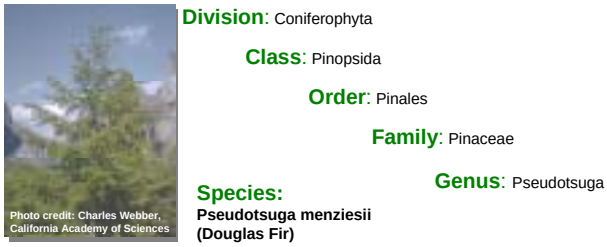

Figure 5. Example of a taxonomic vegetation hierarchy (1)

Functional relationships

Functional relationships correlate one term with another when their relationships are associated according to some function such as behavior, soil type, etc. These relationships may involve one domain or multiple domains. The characteristics of each organism, such as eating, migrating, breeding, with the frequencies of those activities, can be related to each other. For example, a black bear (organism) hibernates (activity) annually (frequency).

In an example using multiple domains (soil type and vegetation), ultramafic soils have strong correlations with specific plants and plant communities. Therefore one can relate ultramafic soils to fern species, in this example, and find fern species when one is looking for documents about ultramafic soils and vice versa. Further, serpentine soils are ultramafic, so one could look for serpentine soils and be alerted to documents about ferns and ultramafic soils.

4 Summary

The computer science focus for this project is to embrace the complexity of the controlled vocabularies and the relationships that are useful for natural resource management. This includes: multiple controlled vocabularies for a domain, terms appearing in multiple controlled vocabularies, often with different, even contradictory broader-term/narrower-term relationships, and a broad range of functional relationships. The computer science contribution to this project is first to define the underlying model for terms and relationships that effectively accommodates the scale and complexity of the terminology of interest. And, secondly, to permit the terms and their relationships to serve as the basic frame or structure to attach the various documents of interest to the natural resource managers. Finally, we seek to exploit the knowledge to assist the user in all aspects of the system. For example, when a term is selected from a controlled vocabulary to be attached to a document, the system automatically attaches all broader terms (according to the broader-term/narrower-term relationship in the CV). Similarly, if the user requests documents related to a particular term, the system has the capability to augment the search to include all narrower terms (within the current CV) plus all synonymous terms. We are building a system, called Metadata++, for this purpose.

Our project and our technology takes advantage of knowledge already developed (in the form of terms and relationships) to enhance the utility of new information (the documents to be produced). The system also easily accommodates new controlled vocabularies, new terms, and additional relationships so that it can easily grow and evolve as the scientific pursuits and the public policy shifts and expands over time.

The relationships among terms, including cross-reference relationships among distinct terms to describe the same concepts, are the most challenging and the most valuable part of the system. By capturing the intellectual capital of these relationships among terms, one can make possible electronically the kinds of document searching that humans naturally do in libraries: first, they look up something, often very specific. Then, they browse similar topics found “on the shelf” next to where they originally looked. The nature of similarity varies by user needs; so we are working toward making the basis of similarity be quite flexible in this design and controllable by the user.

This system provides the framework or “velcro” to capture much of the knowledge that is typically lost and, therefore, never reused or, at best, reused by only a limited audience. Examples of the knowledge include watershed assessments, environmental analyses, and monitoring reports. According to the

KM.GOV website, knowledge management is “a discipline dedicated to more intentional means of people creating and sharing knowledge, data, information, and understanding in a social context to make the right decisions and take the right actions.”

Thus, this project adheres to the notion that:

> knowledge management recognizes employee experience and expertise (intellectual capital) as the organization's most valuable asset. Knowledge management maximizes value of what we already know and helps manage risks from what we do not know. It serves to continuously improve effectiveness and performance by optimizing; organizing and segmenting information and data flow internally and externally. Knowledge management […] sets the stage for the continual creation of new knowledge and best practices through sharing, feedback, and information exchange back into the system. It can provide balance to the organization's diverse organizational structures and challenges, such as: matrix management, enterprise organizations, decentralized organizations, and coordination across boundaries. (Knowledge Management Team, USDA Forest Service, 2001)