# A Digital GeoLibrary:

# Integrating Keywords And Place Names

Mathew Weaver[1], Lois Delcambre[1], Leonard Shapiro[2], Jason Brewster[2], Afrem Gutema[2], Timothy Tolle[3]

[1]Computer Science and Engineering Department, OGI School of Science and Engineering, Oregon Health & Science University
{mweaver, lmd}@cse.ogi.edu

[2]Department of Computer Science, College of Engineering & Computer Science, Portland State University
{len, jbrew, gutemaa}@cs.pdx.edu

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

**Abstract.** A digital library typically includes a set of keywords (or subject terms) for each document in its collection(s). For some applications, including natural resource management, geographic location (e.g., the place of a study or a project) is very important. The metadata for such documents needs to indicate the location(s) associated with a document - and users need to be able to search for documents by keyword as well as location. We have developed and implemented a digital library that supports – but does not require - georeferenceable documents (i.e., documents with reference to geography through the use of a textual place name). Because of their implicit spatial footprint, place names benefit from spatial reasoning and querying (e.g., to find all documents that describe work performed within a five-mile radius of a certain point) in addition to traditional keyword-based search. This paper presents the architecture for a digital library that combines spatial reasoning and selection with traditional (non-spatial) search. The contributions of this work are: (1) the use of a traditional geographic information system (GIS) for spatial processing rather than a specially tailored GIS system or a separate gazetteer and (2) the seamless integration of GIS with our thesaurus-based Metadata++ system, so users can easily take advantage of the strengths of both systems.

## 1 Introduction

Our work is motivated by the needs of natural resource managers in the USDA Forest Service to develop a distributed digital library to provide easy access to various documents such as Decision Notices, Environmental Impact Statements, and Watershed Assessments [2] produced as a normal part of their work processes and decisions. We seek to provide easy access to these documents for other natural

resource managers who might be doing similar work or facing a similar decision. In our system, called Metadata++ [12], we describe these documents with a rich set of controlled vocabularies for twenty-eight domains of interest such as: air, climate, fire, vegetation, and so forth (as shown in Figure 1).

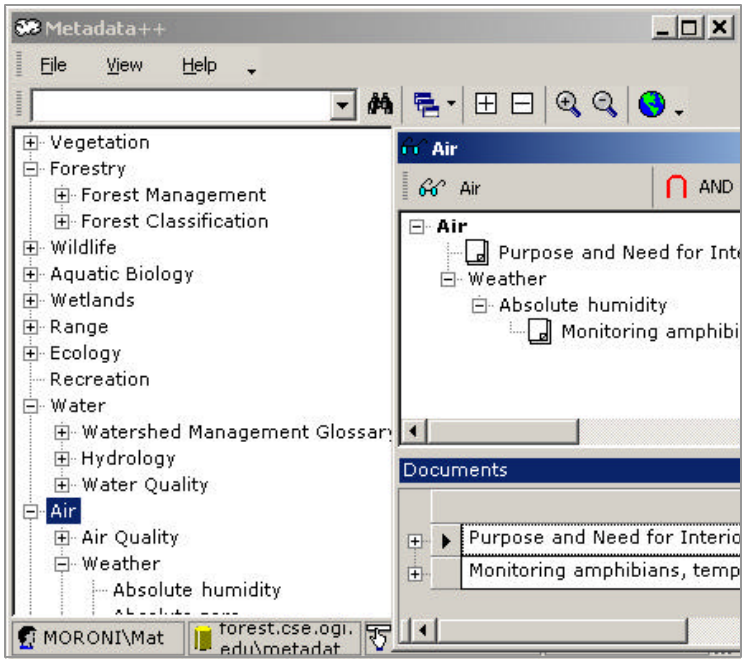



Figure 1: Screenshot of Metadata++ with CVs on the left, and search result (search term "Air") on the right.

In addition to related keywords, the majority of these documents are also associated with one or more geographic areas, e.g., to indicate the area of land where a study was done or where a proposed project will take place. Locations are typically described using standard locations, such as the National Forests/Ranger Districts and Hydrologic Unit Codes as shown in Figure 2. The location schemes can be represented in a controlled vocabulary where spatial containment and equivalence of geographic footprint are represented by narrower/broader term and synonymy. In this application, about a dozen such location schemes are in common use.

The user (quite naturally) often wants to combine search terms from both spatial and non-spatial vocabularies – which presents both challenges and opportunities. We want the benefit of the rich metadata structure and search capability of Metadata++ plus the benefits of spatial reasoning of a standard GIS system. Our approach is to use Metadata++ and a standard GIS system running independently, with communication as shown in Figure 3. GIS datasets containing place names are used to generate controlled vocabularies in Metadata++.

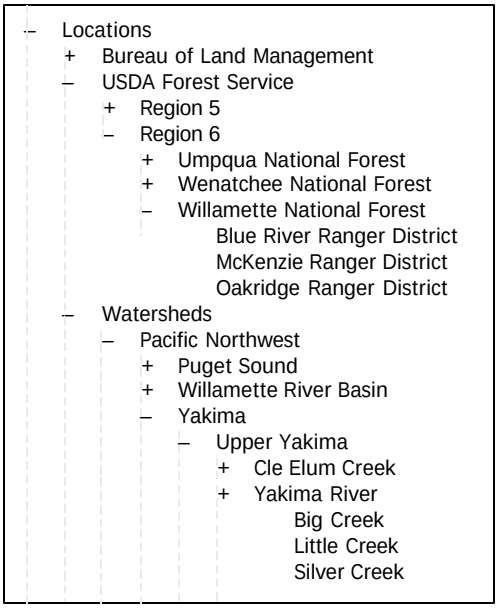



Figure 2: Excerpt showing two location CVs: the USDA Forest Service National Forest System, and Watersheds (as named in the Hydrologic Unit Code scheme).

The user may use the standard GIS system to browse and search GIS datasets, e.g., with rainfall or temperature coverages, soil types, and so forth. The user may also, at any time, select a region of interest (in the GIS system) and send the locations to Metadata++ for inclusion in a document search. Additionally, when Metadata++ has a set of documents in a search result, the documents that are associated with locations can be sent to the GIS system for display on a map. Metadata++ may also ask the GIS system to compute broader and narrower terms (i.e., containing and contained places) and synonyms (i.e., significantly overlapping places) for any location term.

Other approaches to this problem include: augmenting a document retrieval system or a GIS in various ways or introducing a gazetteer to sit between the two systems to translate between place names and geographic footprints [6]. More recently, so-called Geographic Information Retrieval (GIR) systems seek to support both types of search directly [4].

The organization of this paper is as follows. Section 2 provides a quick introduction to Metadata++. Our combined architecture is presented in Section 3, with a discussion of the detailed interaction between the two components. Section 4 discusses how our work compares to other approaches in more detail; and the paper concludes in Section 5.

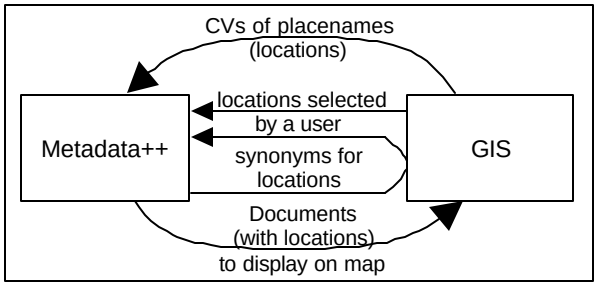



Figure 3: Functions supported by our integration of Metadata++ and a GIS

## 2 Metadata++

Metadata++ is a document retrieval system built on a framework of related terms [13]. *Terms* are words or phrases frequently used in the application domain, that may be related using relationships similar to those in a thesaurus [1], including hierarchy, synonymy, and association. Metadata++ allows a user, usually the author or librarian, to select terms (as shown on the left side of Figure 1) to associate with a document.

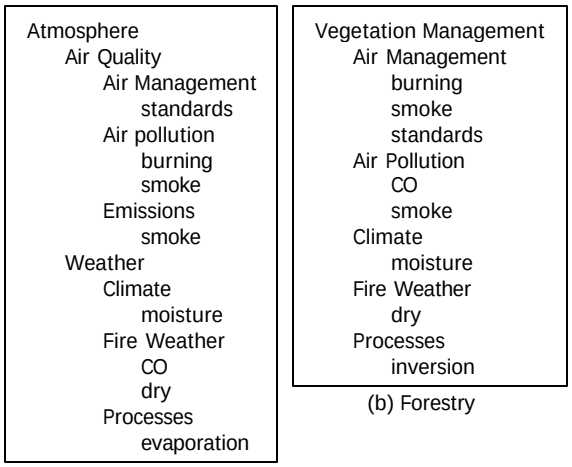
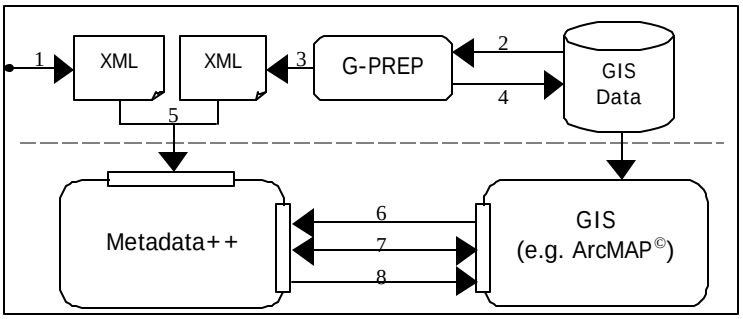



Figure 4: CVs showing terms that appear multiple times, in different arrangements

One significant difference between Metadata++ and a typical thesaurus is that a term can appear in multiple locations in the broader/narrower term hierarchy. As illustrated in Figure 4, users with different expertise (e.g. climatology and forestry) use some of the same terms but organize them differently within vocabularies. Thus Metadata++ distinguishes between a term, such as *smoke* or *air pollution* and the *node(s)* of the hierarchy where it appears. Each node in the hierarchy is thus identified by its trail where a trail is the path from the top of the hierarchy to the selected term. In Figure 4, we see four different nodes for the term *smoke.* One of the trails is *Atmosphere~Air Quality~Air pollution~smoke* and another is *Vegetation Management~Air Management~smoke.* Thus, Metadata++ uses trails to disambiguate terms that appear in multiple nodes.

## 3 Architecture

This section explains our architecture for integrating Metadata++ and GIS to build a digital geolibrary. The architecture enhances retrieval of geographic information by exploiting functionality of both systems, and communicating information between the two systems (as shown in Figure 5).




1. (Non-spatial) CVs are prepared in XML
2. GIS datasets processed by G-PREP
3. Location CVs are prepared in XML
4. Place ID attribute added to GIS datasets
5. XML CV files are loaded into Metadata++
6. User-selected locations are passed to Metadata++
7. Metadata++ asks GIS to compute synonyms
8. Metadata++ passes documents (with locations) to GIS for display

Figure 5: Architecture

### 3.1 Vocabulary Extraction

The first requirement for integrating Metadata++ with GIS is extracting controlled vocabularies of place names from GIS datasets. Geographical places are often naturally hierarchical based on spatial containment. Larger places, such as states or national forests, contain smaller places, such as counties or ranger districts, respectively. In GIS datasets, this hierarchy is represented implicitly by the fact that some spatial footprints are contained within others. For example, the polygon representing the State of Oregon geometrically contains the polygons that represent the counties within Oregon. A process called G-PREP implements vocabulary extraction by taking GIS datasets and generating hierarchical controlled vocabularies of place names expressed in XML (shown in Figure 5, Steps 2 & 3). The controlled vocabularies are then loaded (along with non-spatial vocabularies from other sources) into Metadata++.

While implementing G-PREP, we encountered a number of problems. GIS datasets are usually separated into layers (also called themes) based on similar feature type. For example, USDA Forest Service ranger districts are contained in one layer and national forests are contained in a separate layer. Because of this separation of layers, G-PREP cannot generate the complete hierarchy within the context of a single layer. Instead, the G-PREP needs to know which layers correspond to which levels in the hierarchy. With the datasets available to us, we could not easily automate this process. We had to process each layer in the proper hierarchical order.

The nature of geographic footprints presents another, more significant, obstacle to vocabulary extraction. A person might imagine a layer consisting of precise, disjoint polygons – such as political county boundaries. However, many geographic places cannot be represented by a simple polygon. For example, a national forest may

consist of many non-contiguous regions of varying shapes and sizes – resulting in a complex geographic footprint. All of these polygons must be mapped to one conceptual place name (i.e. the name of the national forest). G-PREP implements this mapping by generating a conceptual place identifier and adding the identifier as a new attribute for each polygon. This place identifier attribute is added to the original GIS dataset (as shown in Figure 5, Step 4) and included in the XML hierarchy that is loaded into Metadata++.

The place identifier generated by GPREP is also used to disambiguate place names. For example, the State of Oregon contains twenty six places that are all officially named "Salmon Creek" – three of which are not even creeks! A simple keyword search for "Salmon Creek" would likely yield many irrelevant documents. Using a place identifier in both the GIS and Metadata++ allows the user to precisely select a location of interest – whether from a map or from a controlled vocabulary.

Another significant problem arises because various footprints are generated at different times, using different instruments, for different purposes – and quite often have different precision and accuracy. Because of complex and imprecise footprints, calculating spatial containment is non-trivial. For example, the footprint of a ranger district may slightly extend beyond the footprint of the national forest – even though the ranger district is under the administrative jurisdiction of (and contained within) the national forest. G-PREP uses heuristics to handle these calculations but occasionally requires user interaction.

Representing places as terms in Metadata++ supports the use of unofficial place names – places that are commonly referred to but do not have an official geographic footprint. Searching for unofficial places names is defined as a necessary feature for digital geolibraries [10]. For example, most people in the State of Washington know about Snoqualmie Pass – many people could take you there without any difficulty. However, Snoqualmie Pass does not have an official geographic footprint – and may not even appear on some maps. By representing Snoqualmie Pass as a term in Metadata++, the user may still use it for document retrieval – even though it may or may not appear in a GIS dataset. Furthermore, if an unofficial place does have a footprint, we can easily assign a place identifier – as with official places – and include the place name in a controlled vocabulary within Metadata++.

### 3.2 Place Selection

A primary user task is selecting places within the GIS and communicating those selections to Metadata++ – to associate with a document or specify a search. Place selection may be as simple as pointing to a region on the map and clicking the mouse to select the polygon. However, one advantage of using the GIS is the ability to do more advanced spatial analysis. For example, the user may issue a query to find all counties with geographic area less than two million acres. The GIS will answer this query by automatically selecting those regions within the active dataset that satisfy the specified query. In addition to maps, the user may choose to use other types of GIS data to assist in finding relevant regions – such as a dataset of annual precipitation to find those counties that receive more than one hundred inches of rain per year. The GIS will answer this query by intersecting the qualifying regions of the

annual rainfall dataset with the map of counties. This query could be combined with the previous example to find all counties with area less than two million acres that receive over one hundred inches of rain annually.

After using GIS functionality to select the desired place(s), the place names are sent to Metadata++. Selecting a place in GIS is equivalent to selecting the corresponding place name in the Metadata++ hierarchy. The selected place may be used for any Metadata++ function – such as document search or metadata attachment. Users have the flexibility of using either Metadata++ or the GIS – or a combination of both systems – for selecting places. Because all terms are presented to the users in Metadata++, they can easily mix geographic terms (e.g. place names) with any other (non-geographic) terms.

Our current implementation adds a new button to ArcMAP© (a popular GIS application produced by ESRI®). The user may use all of the functionality in ArcMAP© to analyze geographic datasets. At any point in that process, the user may select one or more locations (using any of the selection mechanisms in ArcMAP©) and click the button to send the selections to Metadata++. The integrated architecture could easily be extended to work with other GIS applications – transforming any GIS application into a powerful tool for geographic information retrieval.

### 3.3 Synonym Discovery

Metadata++ uses synonyms for query expansion during document retrieval. For non-geographic controlled vocabularies – such as wildlife species or climate terms – synonymous terms are represented explicitly in Metadata++. For geographic controlled vocabularies of place names, Metadata++ uses the GIS to find synonyms. We could explicitly represent synonymous places in Metadata++, but we chose not to do so for two reasons. First, a large number of synonyms exist among places. Every place within every location vocabulary may be considered a synonym with one or more places in many other location vocabularies. For example, the land within Clackamas County is also within the Lower Willamette River basin. All of the land in the State of Oregon belongs to some county and also belongs to some watershed and may also be part of a ranger district – resulting in a large number of possible synonyms.

The second, more significant, reason is the ambiguous semantics of spatial synonymy. If the spatial footprints of two different places exactly coincide, then those places would likely be defined as synonyms. However, that rarely – if ever – occurs in real geography. Clackamas County and the Lower Willamette River Basin are in the same geographic place, but their spatial footprints do not coincide. Some points within Clackamas County are not within the Lower Willamette River basin and vice versa. This type of situation makes it difficult to explicitly represent spatial synonyms as related terms in Metadata++.

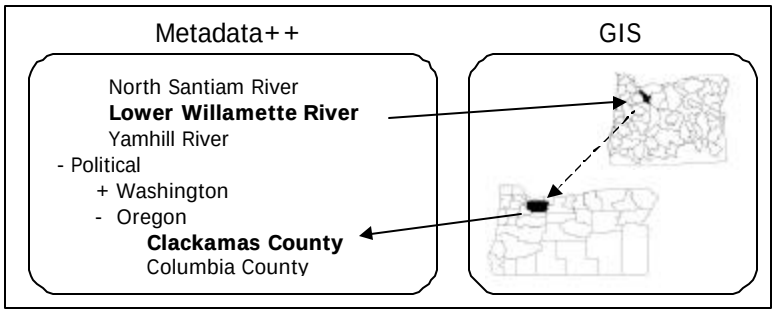



Figure 6: Synonym Discovery

Spatial synonyms are determined in the GIS by computing a percentage of overlap between the polygons based on a user-specified threshold. Because the GIS computes synonyms (instead of representing them explicitly in Metadata++), the user may adjust the threshold to achieve the desired results. In addition to percentage of overlap, the user may also wish to include other GIS computations (such as adjacency or proximity) while discovering synonyms. These computations are performed on a default set of datasets – or on one or more specific datasets chosen by the user.

Figure 6 shows an example of synonym discovery within the GIS. The user selects Lower Willamette River from the Watershed controlled vocabulary. Metadata++ then sends that selection to the GIS. Using the established threshold, the GIS determines which place(s) in other vocabularies (such as political regions) overlap with the selected region. In the example, the GIS determines that the Lower Willamette River basin is a spatial synonym of Clackamas County, and returns that place to Metadata++. In addition to finding synonyms for places selected in Metadata++, synonym discovery may be combined with place selection. If a user selects an area on a map, the GIS can compute the spatial synonyms and send all of the corresponding places (the selected place and its synonyms) to Metadata++ for processing.

### 3.4 Document Display

Because many natural resource documents are attached to one or more place names, it is useful to display the documents on a map. When such a document is part of a query result, Metadata++ can send the appropriate place name(s) to the GIS. The GIS will then display the documents on the appropriate map. For example, consider a document about hazardous tree removal in Clackamas County. Perhaps the user searched for documents about tree removal in the State of Oregon. Metadata++ would show the document in the context of the hierarchy and would notify the GIS that the document was attached to Clackamas County. The GIS would then locate the polygon corresponding to Clackamas County and display an icon representing the document in or near the polygon.

## 4 Related Work

Geographic information retrieval (GIR) focuses on document retrieval based on geographic references within documents. Each document can be uniquely classified as georeferenced, georeferenceable, or non-georeferenceable. Georeferenced documents (such as a digital map) contain explicit reference to geography in the form of a spatial footprint (coordinate, polygon, etc.). Georeferenceable documents (such as an Environmental Impact Statement) do not contain an explicit spatial footprint, but do contain implicit reference to geography in the form of one or more textual place names. Non-georeferenceable documents (such as a scientific report describing the feeding process of spotted owls) are not related – implicitly nor explicitly – to any geographic location. Each type of document must be considered in a GIR system.

Some GIR systems use only spatial queries for document retrieval. Spatial queries execute over geographic footprints – so georeferenced documents are easily retrieved. However, before a georeferenceable document can be considered in a spatial search, the system must somehow associate a footprint with the document [14] – at which point it becomes georeferenced. Automatically associating footprints with georeferenceable documents is non-trivial. Georeferenceable documents contain place names – but place names are often ambiguous. For example, a document about the first president of the United States may be mistakenly associated with the town of George in the State of Washington (which is a real town). Similarly, the city of Portland exists in Oregon and Maine. A related problem is the use of alternate or informal names for places. For example, an older or more informal document may refer to Boston, Massachusetts as "bean town" or Portland, Oregon as "stump town". Metadata++ searches for documents within the context of the term hierarchy – not in context of the spatial footprints of documents. Our architecture uses the GIS to execute spatial queries to extract vocabularies of place names and discover spatial synonyms, but document retrieval occurs in Metadata++ (which does not require that documents be associated with spatial footprints).

The ADEPT [6] Digital Library Architecture provides support for geographic information retrieval using search buckets. A search bucket is an abstract metadata category with defined search semantics. Collections may provide metadata for items (e.g., documents) using various buckets. For example, the "Geographic location" bucket contains coordinates describing a document's spatial footprint and supports three spatial search operators: contains, overlaps, and is-contained-in. The "Assigned term" bucket contains subject-related terms from controlled vocabularies and supports three text-based search operators: contains-all-words, contains-any-words, and contains-phrase. The ADEPT architecture searches multiple heterogeneous collections by specifying queries using the various bucket search operators. By treating place names as terms in controlled vocabularies, Metadata++ uses a single, unified search operation that exploits hierarchical relationships and synonyms. Synonyms discovered by the GIS are handled uniformly with synonyms explicitly represented in Metadata++.

Recent GIR research [3,7,8] relies on ontologies to facilitate information retrieval. SPIRIT [8] attempts to address proximity relationships as well as alternate and informal place names by developing a geographical ontology that "models both the vocabulary and the spatial structure of places". OASIS [3] represents places as

explicit objects in an ontology with specific attributes (latitude, longitude, standard name) and relationships (meets, overlap, partOf) to other place objects. These relationships are used to explicitly represent and query spatial relationships between places. Instead of building an explicit ontology, Metadata++ focuses on faithfully representing the controlled vocabularies (both spatial and non-spatial) that are commonly used in the application domain. The hierarchy among place names in Metadata++ is similar to the partOf relationship, but Metadata++ relies on the GIS to discover spatial synonyms instead of representing those relationships (e.g. meets, overlap) explicitly.

Our work is similar in spirit and proposes a similar architecture to that of GeoVSM [4]. The authors argue convincingly that GIR systems must support two kinds of description (keyword as well as spatial) as well as two kinds of search, although they assume that both sides of the system are providing search capability over the same set of documents. They also recommend that the user interface available in a GIR system include two different user interfaces for the two components because they correspond to distinct ways of representing and organizing information. Our architecture follows the same philosophy, with a separate interface for Metadata++ and the (standard) GIS system. Key differences in our work compared to GeoVSM are that: (1) our system explicitly accommodates non-georeferenceable documents; (2) our GIS component is a standard GIS system that is used to browse various kinds of maps and layers and to select locations (but is not explicitly used to search for documents); (3) our document system, Metadata++, does not use a spatial metaphor to display non-spatial keywords, rather we provide a hierarchical display of terms; and (4) because of the explicit use in this domain of controlled vocabularies to describe places, we are able to easily combine place names with any other (non-spatial) terms in our description and search of documents in Metadata++.

G-Portal [9] is a map-based digital library architecture for georeferenced resources. Like Metadata++, G-Portal provides a map-based interface and a classification interface (to support non-georeferenceable documents). The authors emphasize synchronization between the interfaces – documents selected in one interface will be automatically selected in the other interface. Our work differs from G-Portal in the primary purpose of the map-based interface. In Metadata++, the map-based interface (i.e., the GIS) is not used to specify searches for documents. Rather, it is intended to search for geographic places so that the place names can be combined with non-geographic search terms in Metadata++ – as well as displaying georeferenceable documents from search results.

## 5 Conclusions and Future Work

We have implemented a prototype of our architecture (including vocabulary extraction and place selection). Initial user feedback is very positive – and testing will continue with additional users. Future work will include implementation of synonym discovery and document display. We will also consider georeferenced documents that are associated with unnamed places (e.g., a coordinate). Such

documents may be handled using spatial synonyms and/or spatial searches within the GIS.

We are continuing our work on Metadata++. This includes implementing templates as a way for users to pre-select certain terms in the hierarchy for search or metadata attachment. For example, a botanist that works in the Wenatchee National Forest may define a template that will automatically select Wenatchee National Forest and Douglas Fir trees. We have also begun a detailed comparison of Metadata++ searches based on trails with vector space searches based on keyword vectors. A detailed comparison will help us better evaluate the overall effectiveness of Metadata++.

By combining a rich, vocabulary-based document retrieval system with traditional GIS tools, we enable users in natural resource management to effectively and precisely retrieve geographic (and non-geographic) information. Combining vocabularies of place names with other controlled vocabularies, allows us to provide a simple, unified mechanism for creating and searching metadata. Our architecture lets users easily specify searches involving spatial and non-spatial search terms. The architecture provides a clean separation between the two components: Metadata++ is unaware of spatial footprints and the GIS is unaware of documents. Instead, our architecture simply provides for the GIS to deliver place names to Metadata++, on user request. And such terms in Metadata++ are known to have an external identifier. A term with an external identifier in Metadata++ can have synonyms computed by the designated external system. Also, documents associated with such a term can be passed to the external system for display or other processing. This architecture permits the use of various GIS software tools as well as other external systems, such as an ontology, with Metadata++.

*Acknowledgements*:

We acknowledge the helpful collaboration of the USDA Forest Service under the direction of Tim Tolle. We also thank Dale Guenther of the USDA Forest Service for his help in gathering and understanding several GIS datasets.

## 6 References

[1] ANSI/NISO Z39.19 – 1993. *Guidelines for the Construction, Format, and Management of Monolingual Thesauri.* NISO Press, 1994.

[2] Delcambre, Lois, and Timothy Tolle. "Harvesting Information To Sustain Forests". *Communications of the ACM*, January 2003 Volume 46, Number 1, pp. 38-39.

[3] Fonseca, F., M. Egenhofer, P. Agouris, and C. Câmara. "Using Ontologies for Integrated Geographic Information Systems". *Transactions In GIS* 6(3), 2002.

[4] Guoray Cai, "GeoVSM: An Integrated Retrieval Model For Geographical Information". Lecture Notes on Computer Science (LNCS) 2478: *Geographic Information Science*, M. Egenhofer and D. Mark, (ed) Springer-Verlag (2002)

[5] Guoray Cai, :GeoVIBE: A Visual Interface For Geographic Digital Libraries", Lecture Notes on Computer Science (LNCS) 2539: *Visual Interfaces to Digital Libraries*, K. Borner and C. Chen, Eds.: Springer-Verlag (2002)

[6] Janée, Greg and James Frew. "The ADEPT digital library architecture". JCDL 2002: *Proceeding of the second ACM/IEEE-CS Joint Conference on Digital libraries*, 2002, Portland, Oregon, USA., ACM Press, pp. 342-350.

[7] Jones, Christopher B., Harith Alani, and Douglas Tudhope. "Geographical information retrieval with ontologies of place". Lecture Notes in Computer Science (LNCS) 2205: *Spatial Information Theory Foundations of Geographic Information Science*, D. Montello (ed), COSIT 2001, Springer-Verlag (2001)

[8] Jones, Christopher B., R. Purves, A. Ruas, M. Sanderson, M. Sester, M. van Kreveld, R. Weibel. "Spatial Information Retrieval and Geographical Ontologies: An Overview of the SPIRIT project" in 'SIGIR 2002: *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, August 11-15, 2002, Tampere, Finland', ACM Press, pp.387 - 388.

[9] Lim, Ee-Peng, Dion Hoe-Lian Goh, Zehua Liu, Wee-Keong Ng, Christopher Soo-Guan Khoo, Susan Ellen Higgins. "G-Portal: a map-based digital library for distributed geospatial and georeferenced resources". JCDL 2002: *Proceeding of the second ACM/IEEE-CS Joint Conference on Digital libraries*, 2002, Portland, Oregon, USA., ACM Press, pp. 351-358.

[10] National Research Council, "Distributed Geolibraries: Spatial Information Resources, Summary of a Workshop". Panel on Distributed Geolibraries, National Academy Press, 1999.

[11] Seaber, P.R., F.P Kapinos, G.L. Knapp. "Hydrologic Unit Maps". *U.S. Geological Survey Water-Supply Paper* 2294, 1987, 63 p.

[12] Weaver, Mathew, Lois Delcambre, David Maier. "A Superimposed Architecture for Enhanced Metadata". *Proceedings of the DELOS Workshop on Interoperability in Digital Libraries*, Darmstadt, Germany, September 2001.

[13] Weaver, Mathew, Bill Howe, Lois Delcambre, Timothy Tolle, David Maier. "Representing, Exploiting, and Extracting Metadata using Thesaurus++". *National Conference for Digital Government Research* (dg.o 2002), Los Angeles, California, May 2002.

[14] Woodruff , A. G. & Plaunt, C. "GIPSY: Geo-referenced Information Processing System". *Journal of the American Society for Information Science*, (1994a), 45, 645-655.