Superimposing Spatial Enrichments in Traditional Information

[1]Lois Delcambre, [1]David Maier, [1]Mathew Weaver, [2]Leonard Shapiro, [3]Judith Bayard Cushing

[1]Computer Science Dept.
OGI School of Science & Eng.
Oregon Health & Science Univ.
{lmd, maier, mweaver}@cse.ogi.edu

[2]Computer Science Dept.
Portland State University
len@cs.pdx.edu

[3]Member of the Faculty
(Computer Science)
The Evergreen State College
judyc@evergreen.edu

Abstract

There are various tools available to manipulate and reason with spatial information, including GIS tools. And it is quite common for traditional, non-spatial datasets to contain spatial information. In this work, we seek to enable spatial reasoning for such datasets. We describe two applications with a mix of spatial and non-spatial information and describe the additional structure and function that we propose to introduce to enable spatial co-location and spatial reasoning. The first is a forest canopy study where a spreadsheet is used to capture scientific observations of epiphyte coverage. The challenge for this application is to co-locate observations from different studies, to support more complex scientific studies. The second application is a domain-specific digital library for natural resource management where a broad range of place names are frequently used to describe documents. The challenge for this application is to compute spatial relationships among place names to determine spatial "synonyms." Enabling spatial reasoning for spatial data embedded in a traditional dataset is one example of our broader goal to enable the use of appropriate tools for (a subset of the) data that is intermixed with other kinds of data.

1 Introduction

Spatial information is often intermixed with other types of information. For example, a street address is often stored in a relational database with traditional, non-spatial information. As another example, a place name such as the *Wenatchee National Forest* may appear in an Environmental Assessment and also be used as a keyword to describe the document. But traditional database or information retrieval systems are not able to do spatial reasoning, e.g., to geocode an address or determine the extent of the overlap between the spatial footprint for Wenatchee National Forest and the footprint for Chelan County. In our research, we seek to superimposed appropriate data structures and operations over non-spatial information sources to support spatial reasoning.

In Section 2 of this paper we consider a database used by forest canopy scientists [5] to record scientific observations and we suggest how a superimposed "spatial harness" can be used to translate a series of relative measurements into precise spatial locations. The spatial harness makes it easy for the forest canopy researchers to structure their datasets in the way that is most familiar to them while still permitting proper location in space. In Section 3, we describe how we use a GIS to manipulate the spatial footprints associated with place names in conjunction with a domain-specific digital library for natural resource management called Metadata++. This is achieved using a simple software architecture that allows Metadata++ to concentrate (only) on documents and keywords (without awareness of spatial reasoning) and the GIS to concentrate (only) on spatial footprints (without awareness of documents). In both of these systems, the original, traditional information in the database or the digital library is not altered yet the user

receives the benefit of the superimposed spatial enrichment. The paper concludes in Section 4 with a brief discussion of our work.

2 Spatial Harness for Forest Canopy Research

Forest canopy science is a broad, multi-disciplinary endeavor that seeks to understand the various ecosystems in the forest canopy and how they relate. Individual investigations consider a range of scientific questions such as: how epiphyte coverage, by species, correlates with vertical position in the canopy or how rain throughfall relates to canopy structure. Other studies seek to understand the habitat for birds, bats, or other wildlife. There is general interest in using results and datasets from individual studies to support broader, correlative investigations, (e.g., to study possible correlation of owl nests with epiphyte coverage). Such cross-study use of datasets is particularly valuable when the observations and measurements in earlier studies were made at the same site. And the most obvious way to compare and correlate the observations from the earlier datasets is by positioning these observations correctly in 3-D.

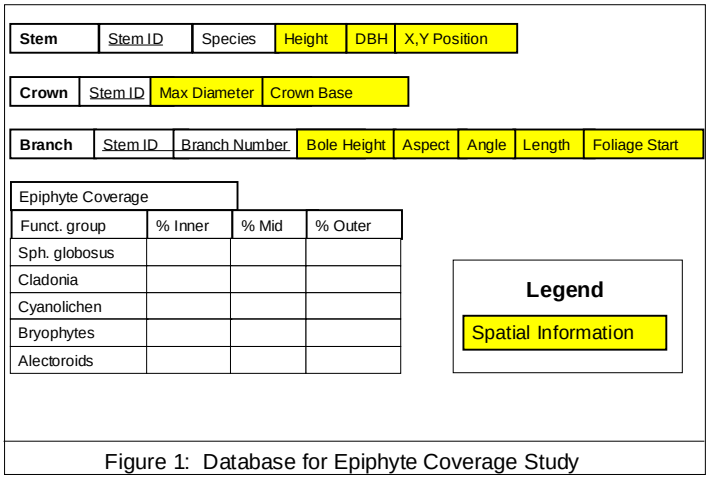

Figure 1: Database for Epiphyte Coverage Study

The simplified structure of a dataset used to study epiphyte coverage [3] is shown in Figure 1. Each tree that was measured has an entry in the Stem table, with associated measurements in the Crown table. Measurements of the percentage of epiphyte coverage, by species, were taken for each branch. Branches are recorded in the Branch table, with a repeating group for Epiphyte coverage showing the percent coverage on the inner, middle, and outer third of each branch for each functional group of epiphytes. We note that the Epiphyte Coverage measurements are the basis of the scientific study, while the Stem, Crown, and Branch data are simply describing the structure of each tree. This dataset was collected at the Wind River Canopy Crane Research Facility[1] and at a stand of young trees, nearby. Such datasets are typically captured in a spreadsheet with fields and repeating groups as shown. This particular study was measuring epiphyte coverage on branches to determine the distribution of epiphytes on trees of a particular species by age class of the trees.

As we examined this dataset, we observed that the scientists were using a basic *stem* (i.e., trunk), *branch* model of forest structure. And some of the collected data fields serve to measure these structural elements (such as *stem height* and *branch length*). More importantly, other

[1] http://depts.washington.edu/wrccrf/

measurements are collected in order to allow spatial connection, in three-dimensions, of the structural elements. For example, the *bole height* of a *branch* indicates where the branch is attached to the stem. And the *branch aspect* and *angle* indicate the angle of the branch in the horizontal and vertical plane. Each stem at the Wind River Facility has a tree tag and the latitude-longitude for each stem is recorded elsewhere. By combining the stem location with bole height, branch aspect, branch angle, and branch length, each branch can be correctly positioned in space. This also provides the spatial position of the scientific measurements (% cover) because *inner*, *middle*, and *outer* refers to the relevant third of the foliage on the branch, where the foliage occurs between the *foliage start* and the *branch length*.

The scientists are taking relative measurements in 1-dimensional coordinate spaces induced by the physical, structural elements of the tree such as bole height, branch length, and foliage start. Each scientist must consider the geometry required for spatial integration, in addition to collecting the scientific observations, so that the structural elements can be properly positioned in space.

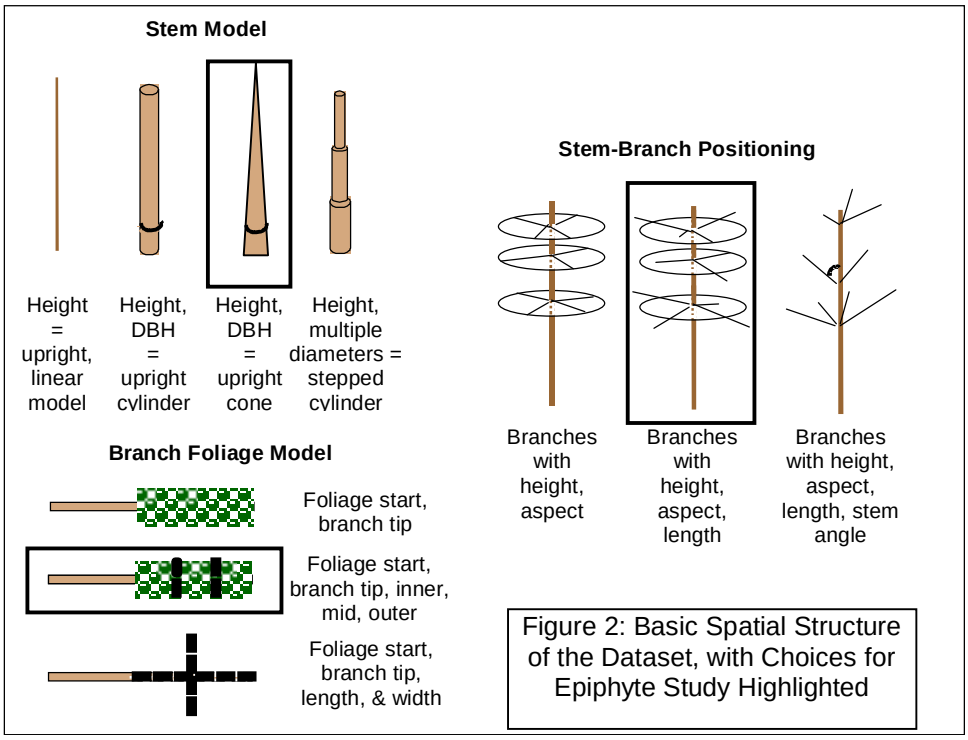

Figure 2: Basic Spatial Structure of the Dataset, with Choices for Epiphyte Study Highlighted

We envision a spatial infrastructure, referred to as a spatial harness [1], that guides their dataset design by helping them understand the spatial implications of the model they choose for structural elements. In Figure 2, we see several choices for modeling stem (i.e., trunk): stem height alone leads to a 1-D upright model, stem height plus DBH ("diameter at breast height," a standard measure of tree diameter) models the stem as a cylinder or cone, and stem height plus multiple diameter readings models the stem with a "wedding cake" structure. The branch foliage models on the lower left side of Figure 2 include: a 1-D model (from foliage start to branch tip), a 1-D model broken into three regions (inner, mid, outer), and a 2-D model (based on length and width of foliage). The branch attachment model, shown on the right of Figure 2, shows various branch attachments that result in various 3-D models. With branch height and aspect, the branch is known to be oriented in the plane perpendicular to the stem. The middle model for stem-branch positioning shown in Figure 2 results when the branch length is known, in addition to branch height and aspect. With branch angle added, we have the third, more realistic model shown.

In the original dataset (shown in Figure 1), the spatial information appears to be ordinary information. But by knowing the spatial semantics, it is possible to automate the geometric calculations for this dataset. Once the tree structure is established in 3-D, one might combine it with other datasets or use a CAD system to compute the wood volume or a visualization tool to render the tree(s).

3 Controlled Vocabularies in Natural Resource Management

In another project [4], we are developing a digital library – called Metadata++ – to satisfy information needs of natural resource managers. This library includes a wide variety of documents, such as Decision Notices, Environmental Impact Statements, Watershed Assessments, research studies, and various other reports. Information contained within these documents is an integral part of natural resource management – and quick, effective access to this information improves the decision making process.

Metadata++ is based on hierarchically structured controlled vocabularies of terms (i.e., keywords) commonly used in natural resource management. Experts from more than two dozen specific domains contributed numerous controlled vocabularies. These vocabularies cover a range of various topics, including climatology, forestry, recreation, vegetation, and wildlife. Figure 3 shows excerpts from a few controlled vocabularies.

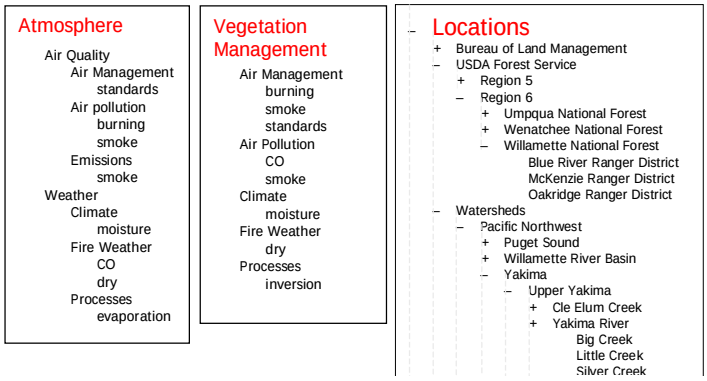

Figure 3: Metadata++: A Digital Library built using Hierarchical Controlled Vocabularies (CVs)

This hierarchy of controlled vocabularies provides an intuitive framework for describing documents, searching for documents, and viewing search results [2, 7]. Keywords for a specific document are defined by relating that document to the appropriate term(s) in the hierarchy. For example, a research study that describes the breeding habitat of Tiger Salamanders would be related to the term "Tiger Salamander" from the Wildlife controlled vocabulary. That same document would also be related to the term "pond" from the Habitat Type controlled vocabulary. By selecting terms from various controlled vocabularies, the author (or librarian) creates keywords for the document.

With documents attached to terms in the controlled vocabularies, users can browse the hierarchy and find documents related to terms of interest. Besides simply browsing the hierarchy to find documents, users may also issue search requests – to find documents attached to one or more terms of interest. A user defines a search by selecting one or more terms from anywhere in the

hierarchy. Searches can include terms from any controlled vocabulary – so a user may include "Tiger Salamander" and "pond" in the same search, even though those terms come from different controlled vocabularies.

When executing a search, Metadata++ uses the structure of the controlled vocabularies to automatically expand the set of search terms to include narrower terms and synonyms as defined in the controlled vocabularies. For example, a search for "Amphibian" would be automatically expanded to include "Tiger Salamander" – since "Tiger Salamander" is a descendant of "Amphibian" in the hierarchy. In addition to query expansion, Metadata++ uses the hierarchy to display the search results. Instead of displaying a linear list of documents ordered by computed relevancy, Metadata++ shows the documents in context of the hierarchy. The user sees an excerpt of the hierarchy, where each relevant document appears beneath the term to which it is attached. The user determines overall relevancy based on where the document appears in the hierarchy.

3.1 Geographic Controlled Vocabularies

Natural resource management is heavily based on geography and many of the documents refer to one or more geographic places – such as an assessment of a particular watershed or a decision notice about a specific national forest. Among the controlled vocabularies defined by domain experts are about a dozen, distinct controlled vocabularies of place names including: political divisions (state, county, city), USDA Forest Service divisions (regions, national forests, ranger districts), watersheds (as described by the Hydrologic Unit Codes), and various eco-regions (described as provinces, under several different schemes). Including multiple controlled vocabularies of place names makes it easier for different users to find information. For example, a forest ranger may want to search by ranger district – whereas a hydrologist may want to search by watershed. Each user may choose the vocabulary that is most familiar to them.

Metadata++ interprets all controlled vocabularies in the same way – regardless of what terms the vocabulary contains. Whether its USDA Forest Service divisions, wildlife species, watersheds, or climatology – the controlled vocabularies all reside in a single hierarchy. By combining all vocabularies in a single hierarchy, Metadata++ makes it easy for users to find relevant information.

3.2 Using GIS Tools with Geographic Controlled Vocabularies

As with the address field of an employee record, the various place names appearing in controlled vocabularies in Metadata++ refer to geographic footprints. Furthermore, the user who is searching for documents may wish to select locations of interest on a map (using a familiar GIS tool) instead of from the hierarchy. The challenge is supporting spatial selection of features (that correspond to place names) freely mixed with non-spatial keywords selected in the user interface of Metadata++. In our solution [6, 8], Metadata++ and a GIS operate independently with specific interactions as shown the Figure 4. Metadata++ is concerned with documents and their descriptive metadata (with no knowledge of spatial footprints) and the GIS is concerned with spatial footprints (with no knowledge of documents).

The geographic controlled vocabularies are generated by processing existing GIS datasets. Feature names found in the datasets become terms, and geographic containment computations determine the hierarchical relationships among the terms. A new attribute is added to each feature that contains an identifier. The generated vocabulary is loaded into Metadata++ just like the other non-geographic vocabularies.

After the GIS datasets are processed and the vocabularies are loaded in Metadata++, users may use the datasets with familiar GIS tools – and do any of the functions supported by the tool (zoom, pan, query, etc.). At any point in this process, the user can select one or more features on the map, and send the selection to Metadata++. The user may then use Metadata++ to select

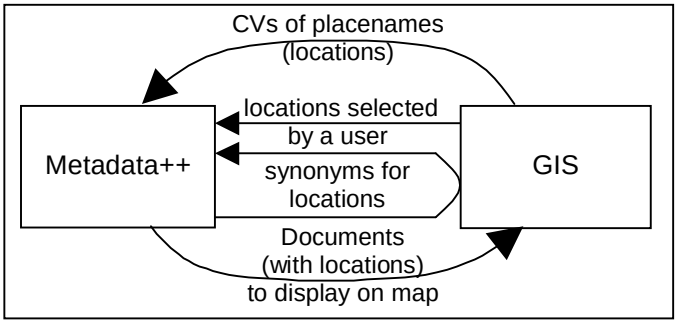

Figure 4: Functions supported by our integration of Metadata++ and a GIS

additional terms (geographic or non-geographic) and issue the search. Figure 5 shows a map where the user has selected a place (Oregon) which is then highlighted in the hierarchy of terms in Metadata++ shown on the left.

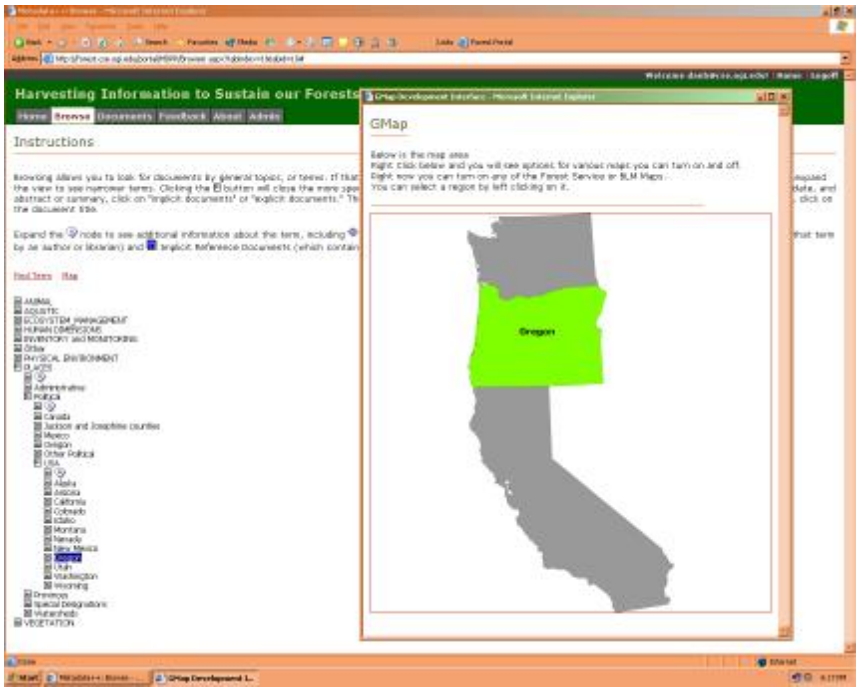

Figure 5: Screenshot of Metadata++ with GIS Displaying Map

During automatic search expansion, Metadata++ asks the GIS to compute spatial synonyms – one or more features that are near or overlap the place of interest. For example, a search for "Clackamas County" may also include "Lower Willamette River" – because those two places are geographically the "same" place. Spatial "synonymy" is not well-defined – because few, if any, geographic footprints exactly coincide. Using the GIS to dynamically compute spatial synonyms

allows the user to define "synonymy," e.g., as overlap by a specified percentage or within a specified radius.

When documents retrieved by a search are related to place names (i.e. terms from geographic controlled vocabularies), Metadata++ will send these place names (and the related documents) to the GIS tool. The GIS tool will display the documents as icons on the map located at the specified place(s). The user can view the search results within Metadata++ (in context of the hierarchy) or on the map (in context of geography).

Superimposing spatial enhancements in Metadata++ makes a valuable contribution to the usefulness of the application. Keeping all (geographic and non-geographic) controlled vocabularies in a single hierarchy allows Metadata++ to provide a unified interface and search engine for combining geographic and non-geographic search terms. The superimposed GIS tool provides a useful, familiar interface for working with the geographic terms.

4 The Vision: Natural Information Processing

The goals of this work are (1) to exploit spatial and other kinds of information that are mixed with traditional information in a single dataset and (2) to superimpose semantic information that highlights the spatial information so that it can be easily used by the specialized tools that understand the spatial aspects. The bole height, and branch length, angle, and aspect appear to be no different in character from the other fields such as branch id or tree species in Figure 1. But with the proper superimposed semantics, these fields can be processed geometrically and then the dataset can be visualized, analyzed, or processed spatially. In a similar manner, the place names in Metadata++ appear as ordinary keywords. But with the proper superimposed semantics, they can be conveyed to a GIS system for display and spatial reasoning.

We envision technology where information of various sorts can be co-mingled and where selected parts of the information can be highlighted, augmented with additional semantics, and easily manipulated in useful ways. The goals of the research are to

- leave the original information where it is, as it is;
- specify the particular information that is of use externally and supply the additional information or processing that is needed to prepare it for external processing;
- use existing tools to process the information, without losing the connection to the original information.

We refer to this style of information processing as *doing what comes naturally* because we are processing spatial, temporal, probabilistic, or any other kind of information with natural, powerful (often familiar) tools.

5 References

[1] Judith Bayard Cushing, Nalini Nadkarni, Lois Delcambre, David Maier, "Spatial Data Infrastructure for Ecological Research," Proc. of the National Conference on Digital Government Research, May 2002, Los Angeles, CA, pp. 157-160.

[2] L. Delcambre, T. Tolle, et al, "Harvesting Information to Sustain Forests," L. Delcambre, *Comm. of the ACM*, 46(1), January 2003, pp. 38-39.

[3] Betsy Lyons, "Crown Structure and Spatial Distribution of Epiphytes on Western Hemlock in an Old Growth Coniferous Forest," M.S. Thesis, The Evergreen State College, Thesis Advisor; Nalini Nadkarni, March 1998.

[4] NSF Award Number EIA 99-83518. Harvesting Information to Sustain our Forests, Digital Government Program.

[5] NSF Award Number 9630316. “Enhancing Researcher Productivity at Shared Research Facilities: Database Tools for Analyzing Forest Canopy Structure.”

[6] L. Shapiro, L. Delcambre, T. Tolle, M. Weaver, D. Maier, D. Guenther, J. Brewster, A. Gutema, “G-Metadata++: Rich Keyword Search Enhanced with a GIS”, Proceedings of GIScience 2002, Boulder CO, September 25-28, 2002.

[7] Mathew Weaver, Lois Delcambre, David Maier, “A Superimposed Architecture for Enhanced Metadata”, DELOS Workshop on Interoperability in Digital Libraries, held in conjunction with European Conference on Digital Libraries (ECDL 2001), Darmstadt, Germany, Sept. 2001.

[8] Mathew Weaver, Lois Delcambre, Leonard Shapiro, Jason Brewster, Afrem Gutema, Timothy Tolle, “A Digital GeoLibrary: Integrating Keywords And Place Names,” Proc. of the 7th European Conference on Research and Advances in Digital Libraries, Trondheim, Norway, Aug. 2003.