Spatial Data Infrastructure for Ecological Research

Judith Bayard Cushing, Nalini Nadkarni
The Evergreen State College
Olympia, WA 98505
judyc@evergreen.edu, nadkarnn@evergreen.edu
http://canopy.evergreen.edu

Lois Delcambre, Dave Maier
Oregon Graduate Institute
Portland, OR 97006
lmd@cse.ogi.edu, maier@cse.ogi.edu

Abstract

The Biodiversity and Ecosystem Informatics Research Agenda (http://bio.gsfc.nasa.gov) notes current inability to compare data across spatial scales as a critical problem: "Biological data from different sources are frequently collected and presented in different scales and resolutions resulting in a loss of detail when multiple data sets are required for data synthesis and analysis." The proposed infrastructure would allow a scientist to define a data set by putting together spatial building blocks represented in conceptually familiar, domain-specific terms. Data sets thus constructed (or recast) would be amenable to the automatic integration of spatial locations and measurements with automatic spatial data transformations. This would allow spatial analysis of individual field data sets and the linking together at same and different spatial scales data sets defined within the spatial infrastructure. We aim to include productivity enhancing research tools such as field data gathering devices and data validation and analysis, and to automatically tag field data with metadata early in the research cycle. We also contribute to the synthesis of metadata and data. This paper describes a recently funded NSF project that aims to build such an infrastructure. We describe the need for the infrastructure, our design approach, work to date and future work.

1. Introduction. Most current ecology research is accomplished through measurement, observation, and study in the physical world. Although most studies now rely on measurements taken by hand in the field, remote sensing and satellite devices are increasing the capacity of the ecologist to gather data. Such new methods do not change a fundamental aspect of ecological data. Determining the location of a physical object in the real world, and its relative position with respect to other objects of interest, is critical not only for the author of a data set, but (because the common elements among different studies are often spatial locations) to others undertaking cross study and interdisciplinary analysis. New methods of data collection and increasing availability of data archives will, however, encourage the development of new modeling applications that will likely increase the need for ecologists to work at different spatio-temporal scales. We believe that the paramount importance of reporting the location of any measured object and an increase in studies crossing spatio-temporal scales, mean that ecologists will need automated tools that support spatial operations.

Although it may seem obvious that determining the precise location of any object simply requires a GPS, our experience with ecology data has convinced us to the contrary. Because of difficulties in conducting field work, nearly all measurements are taken relative to a very local frame of reference. As a result, different measurements even in the same study are sometimes taken relative to different coordinate systems. The analysis and visualization of such a data set as well as its linking with other data sets often requires the resolution of different coordinate frames, each induced by a physical object in the real world.

These different coordinate frames result from the nearly universal practice of taking relative measurements: Where is a branch located in space? It's 15.3 m from the base of the tree, pointing 35 degrees west from north (bole height and aspect: a measurement in a 3-D space induced by the tree stem). Where is the bird nest? It's 29 cm from the tip of the branch (a measurement in a 1-D space induced by the branch). Where is the rain collector located? It's at x=3m, y=4m in the study plot (a measurement in a 2-D space induced by the plot).

To the individual scientist, data set design is a necessary but secondary activity. He or she typically focuses on data collection and subsequent analysis to answer one (or a few) specific, focused question. In-depth consideration of the geometry and spatial aspects of the data to meet current and future needs is usually not the primary focus. Failure to capture even one simple measurement could prevent the data from being properly placed in 3-D and thus defeat the use of various tools and or prevent investigation of questions in unanticipated cross-study correlations. Finally, although there is no need for every scientist to reinvent the spatial structures and transforms needed for their science, individual researchers often spend much time performing spatial transformations – at best time consuming and at worst error prone. Our discussions with experts in geographical systems has confirmed that the most onerous aspect of preparing field data for analysis is refining the spatial reference points. Better database infrastructure for manipulating and relating spatial objects, defining coordinate frames for those objects and capturing measurements relative to these coordinates would increase the productivity of individual ecology researchers and of interdisciplinary teams, allow better integration of new kinds of data and enable more facile data mining and spatial scaling.

Because of the inherent spatial nature of ecological data, coupled with increasing need to increase the productivity of scientists in the field by the use of remote sensing devices, and ultimately to perform cross study analyses, we proposed an NSF planning grant that will culminate in a BDEI proposal to develop a "Spatial Data Infrastructure for Ecological Sciences".

2. The Envisioned Spatial Infrastructure. The infrastructure would allow a scientist to define a data set by putting together the proper "spatial" building blocks – building blocks represented in conceptually familiar, domain-specific terms. Its key contribution would be the automatic integration of spatial locations and measurements with automatic conversion among coordinate frames. The system we envision would articulate appropriate spatial data transformations that would allow spatial analysis of individual field data sets and the linking together at same and different spatial scales of data sets that are defined within the spatial infrastructure. Figure 1 shows our vision for end user data set design with domain-specific constructs that have an underlying spatial infrastructure. This example concerns the understanding of forest and tree structure.

Researchers would select from among "real world" conceptual structures such as those in Figure 1 to represent their data. Each structure would have built in one or more alternative domain-specific coordinate frames that a scientist could select from. An example domain-specific coordinate frame for canopy studies might be: "from bole out [or from tip in] on a branch". These structures would embody mechanisms for spatial transformations and scaling when conducting analysis. In addition, and as important, the structure's coordinate frames provide the cement that holds together the individual spatial objects into a coherent spatial organization: "connectors" that can situate one spatial object relative to another. Ultimately, we anticipate being able to connect other real world objects to any one scientist's collection when data mining. Incidentally, we aim to include productivity enhancing research tools for field data gathering and data validation and analysis, and to automatically tag field data with metadata early in the research cycle.

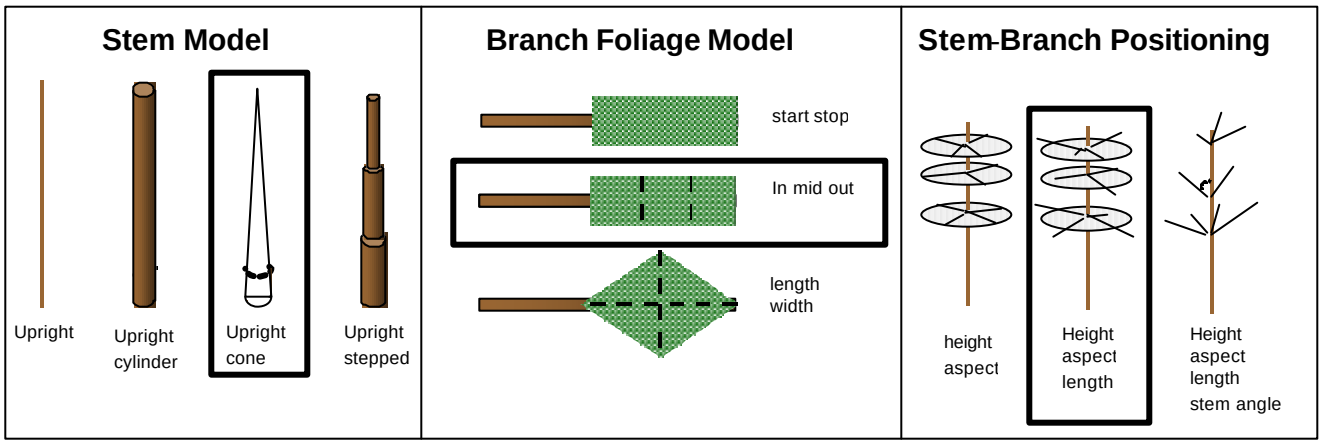

Figure 1. End-user Database Design using domain-specific constructs.

We want to make data set design and data set analysis, as well as metadata acquisition and data validation, very easy for the wide range of ecologists whose data reflect significant spatial components. Using such spatially-enabled building blocks to design a research data set will require less effort than current methods. The project could automate the difficult, labor-intensive step of geo-coding data sets for use in spatial and other sophisticated tools. This capability would go beyond work to date in spatial databases, e.g., where spatial queries are expressed against spatial data (often in a single coordinate frame). This work could lead to the explicit incorporation of such domain-specific spatial frameworks with GIS, spatial, or DBMS tools.

3. Work to Date. In the Canopy Database Project, computer scientists (led by Cushing) work closely with ecologists (led by Nadkarni) are building representative field data sets and a web-accessible database with research reference material and documented field data. Our vision is to increase the capacity of canopy researchers to understand organisms and ecological processes by providing mechanisms to more easily manage and archive field data and to link previously unrelated studies. Our belief is that canopy studies can serve as a model for other ecological fields. The project contains both database development and field work. Our prototype includes a research reference system: study site characteristics, scientific citations, canopy images, researcher contacts, safety protocols.[1] The database behind the research reference system also serves as source tables for field data documentation.

The prototype includes tools to upload, view, query and download the field data sets that we are currently working with as exemplars, and to re-use parts of field data bases (domain specific data types) in creating a design for a new field database.[2] We call these data types *templates* and hypothesize 1) that each template has spatial attributes that link it with others, 2) that a template's spatial attributes allow relative location to be resolved to absolute location and in some cases be scaled up or down, and 3) that the functions might be generalizable across a number of template instances.

The primary focus of the field work has been to develop ecological field data set exemplars that lead to a range of the templates that apply to a wide range of field studies for this one ecology domain. Thus far, we have focused on structure-function relationships of forest canopies in Douglas-fir forests of the Pacific Northwest. We collected within-tree and within-stand forest data to serve as templates and sample data for the database design and development. The ecology team established an 8-site chronosequence between 45 to 1000 yrs. old and is collecting functional data (throughfall volume and light attenuation) to

[1] See http://canopy.evergreen.edu/bcd
[2] See http://canopy.evergreen.edu/databank

quantify relationships between canopy structure and function as stands develop. These templates are being compared to a range of other field studies by other ecologists, and we are building databases from those templates. The use of hand-held computing with laser range finders has been explored with building a small prototype to increase productivity of gathering spatial data in the field.

Because metadata will be a critical aspect in spatial scaling, we also conducted an effort with Dr. Barbara Bond at Oregon State University to investigate how data documentation strategies for within-laboratory data sharing can lead to easier later archiving. We have been working with Gody Spycher of the H.J. Andrews Long Term Ecological Research (LTER) Data Center, to develop a process for sending the data and metadata from our stores to their archive.

We conclude from our current work that 1) these data have significant spatial aspects that require considerable effort to transform and scale, 2) such transformations could be exploited for cross-study endeavors, 3) barriers to individual scientists to archiving and re-using ecological data are considerable and still poorly understood by computer scientists, and 4) because canopy research is highly interdisciplinary ecology done in a particular place (the canopy), our work could likely be generalized to ecology as a whole. Further, while sociological issues are certainly relevant to the data archiving (individuals' desire to hold their data close), we do find ecologists willing to publish their data. They cite as the primary barrier to publishing data database design and description – and, until these tasks become less onerous and more obviously helpful, they will continue to balk at or delay data archiving. Because a primary problem in analyzing data for the individual researcher is its spatial transformation, an easy to use database infrastructure that did such transformations should compensate for the time required to describe data. Such a system would offer incidentally the association of metadata and easier data archiving.

4. Research Issues and Future Work. The primary issue to be determined during the planning period is the extent to which building the spatial infrastructure involves new research, as opposed to applying existing research and technology. We pose five research questions:

1. What end-user tasks require spatial infrastructure and what spatial analysis and transforms would be required for building or linking tools to accomplish these tasks?
2. What are the canonical building blocks (and associated spatial coordinate frames) for forest canopy data structures, and how will we generalize these to other ecological domains?
3. How can we incorporate into the infrastructure feasible connections among data sets represented with those building blocks? Should those connections be part of the spatial objects, or separate entitites?
4. How can we formalize the building blocks and coordinate frame reference system?
5. How can we automate the development of new building blocks and coordinate frame reference systems, along with corresponding spatial transformations and the inclusion of new tools or applications?

In the fall of 2002, we will have determined abstractions of our current spatial data sets and a preliminary review of relevant research and technology, and a survey of canopy ecologists and LTER researchers. A workshop in September 2002 will use survey results and data sets from current work and those identified by new collaborators to determine requirements for a spatial infrastructure.