How Trees and Forests Inform Biodiversity and Ecosystem Informatics

To solve critical biosphere-level problems such as global warming, decreased biodiversity, and natural resource depletion, scientists must integrate data from many researchers. This, in turn, requires better data infrastructure and informatics tools than are currently available. The Canopy Database Project brings together computer scientists and ecologists to develop informatics tools for forest canopy research that meet such ecosystem informatics challenges.

Understanding the biosphere and how human activities affect it nearly always requires the efforts of many investigators, usually from various scientific disciplines. The collective analysis of data originally gathered by individuals but subsequently stored in shared databases can yield insights beyond inquiry of a single data set. To put it all together, ecologists need large, complex data warehouses and data-mining facilities.[1,2] A major barrier to such data warehouses is scientists' inadequate documentation of field data so that others can use that data. Early integration of database technology into the research process would enable more efficient data documentation, but the payback to researchers for any additional work must be real and immediate.

The Canopy Database Project is one of several national efforts building prototype systems that deepen our understanding of how field ecologists could use database technology. It focuses on forest canopy research, an emerging ecological subfield. The canopy is one of the richest but most poorly studied habitats in the biosphere (see Figure 1).[3,4] The field's relative youth, with its lack of entrenched methods, legacy data sets, and conflicting camps of competing groups, provides an excellent opportunity to integrate data management and analysis tools into the research process. And, because canopy research is inherently multidisciplinary, the work is generalizable to other fields of ecology.

This article presents one aspect of our approach to building a data archive for canopy researchers. We show how small, ecologist-centered projects that produce immediate short-term value to participating researchers are essential to achieving long-term ecosystem informatics research and development goals. Such projects keep ecologists interested and involved, provide experience with real data and problems, and increase our ability to use effective software engineering techniques to construct larger systems iteratively.

1521-9615/03/$17.00 © 2003 IEEE
Published by the IEEE CS and AIP

Judith Bayard Cushing and Nalini Nadkarni
The Evergreen State College
Barbara Bond
Oregon State University
Roman Dial
Alaska Pacific University

Obstacles for Ecologists

Our project began in 1993 with a planning grant

from the US National Science Foundation. The first step involved surveying 240 canopy researchers to identify major obstacles that impede their research.[5] The most commonly cited obstacle was not difficulty of physical access to the canopy (as we had expected), but problems in managing, using, and sharing data sets. Researchers associated this with a lack of uniformity in collecting, processing, and analyzing canopy data, a dearth of data archives, and an inability to link data for comparative research. Two years later, with a second grant from the NSF, we launched the Canopy Database Project, pairing ecologists with computer scientists to develop database tools that support data sharing and easier access to analysis tools for canopy research.[6]

Creating data-warehousing and data-mining tools for tomorrow's ecologists requires many improvements to current archival systems.[7,8] We must populate these archives with more and better-documented data than now but this is a daunting task for scientists who are not "power computer users." Although ecologists often consult Web-accessible information, they typically still enter data by hand into private data stores that are rarely published or archived. Despite increasing pressure from funding agencies, availability of several ecological data archives, emerging tools for recording metadata,[9] and opportunities for publishing data in the Ecological Society of America's archives (www.esapubs.org/esapubs/archive/archive_main.htm), ecologists still perceive documenting data for archival purposes to be a time-consuming process and many don't even attempt it.[10,11]

Documentation of field data sets is known as scientific metadata and is essential to retrospective or application use of data set information. These metadata are typically recorded, if at all, at the end of a scientific study, usually after researchers have lost their intimate familiarity with data set details. Using database systems could help, but ecologists tend to prefer flat files, spreadsheets, or (at best) nonrelational (flat) database systems. Individual researchers rarely use modern database management systems, although those who are good programmers tend to use sophisticated statistical programs or write complex mathematical models.[12,13]

Project Goals and Objectives

Our initial objective was to integrate database use very early in the research cycle—ideally starting with a study's initial design. Specifically, we wanted to develop an end-user database-design tool that could reuse domain-specific database components. A field database, designed in this way, and used during the entire research cycle, would include ways to note metadata iteratively as the project evolved. Such metadata-marked data are easier to validate, document, transform, analyze, and archive than data from flat files or spreadsheets. Moreover, the field databases developed with a coherent collection of components would be easier to integrate for collaborative and retrospective study than those built idiosyncratically. This vision is being carried forward in our lab with the development of a database-design tool (called DataBank; http://canopy.evergreen.edu/databank).[14,15]

The project's initial focus on end-user database-design encountered three problems. One, we built a tool that generates field databases from off-the-shelf, domain-specific components, but evolving it as ecological studies grow is beyond the technical skill or inclination of typical end-user ecologists. Current database management systems, even Microsoft Access, are not yet easy enough for most nonprogrammers to manipulate. Two, finding the "right" reusable components is not easy and requires effective tools for the user community to publish, maintain, and organize components. Due to the field's relative youth, articulating the components will require careful documentation of field protocols and more time than we originally envisioned. Finally,

Figure 1. The forest canopy of the Pacific Northwest and the Wind River Canopy Crane Research Facility. The WRCCRF is one of several research sites specialized for studying the forest canopy. A construction crane lets researchers study the canopy from multiple viewpoints. Several researchers working with the Canopy Database Project use this crane site. (Photo courtesy of J. Franklin.)

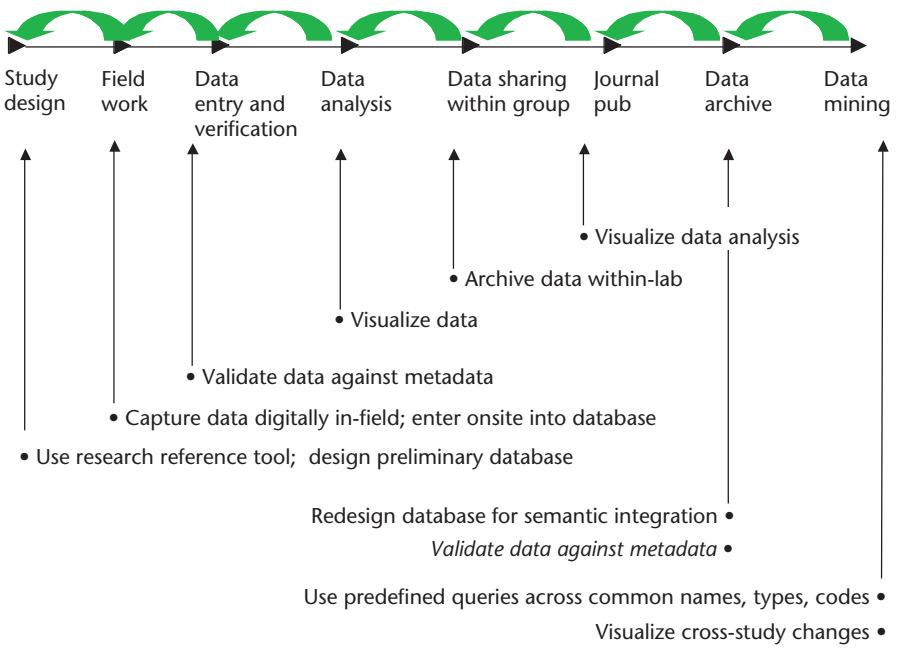

Figure 2. The ecological research process annotated with the Canopy Database Project's efforts to improve researcher productivity. In the first phase of the research process, for example, a research reference tool could facilitate study design and a preliminary database for field data could clarify field protocols. For field work, digital data capture should be made available, and these data should be uploaded onsite to a database. At later stages, visualization, validation and query tools, and processes for intermediate data archiving should be available.

using a database system means changing the way ecologists work. Without clear and immediate benefit, few researchers will spend extra time adopting new technology.

To better understand how to provide immediate benefit from early integration of database tools with DataBank, we needed deeper insight into the research process. To accomplish this, we developed several small, immediately useful tools applicable at various phases of the research process. Two of these smaller efforts, presented in this article, have given particular insight into what a larger tool should look like and what productivity gains might motivate researchers to use database systems and provide metadata.

The Ecology Research Cycle

To learn what ecologists need, we first studied their research cycle. We represent the typical steps in ecology research as a linear process for explanatory purposes, but note that any phase n can progress to phase $n + 1$ or return to any previous phase $n - x$ (see Figure 2).

In the first phase (study design), canopy scientists find studies similar to the one they're designing. They may want scientific articles or field data relevant to the research sites they're considering or accident reports associated with equipment under consideration. To help with this phase, we developed a preliminary domain-specific research reference tool (see http://canopy.evergreen.edu/bcd). Researchers at this stage also typically draft preliminary research protocols, articulate hypotheses, and design preliminary field data intake forms. This phase usually culminates in writing a proposal for funding.

In the second phase (field work), a scientist typically uses the research protocols and field intake forms developed in the first phase. Because field characteristics often differ from those initially envisioned, ecologists often alter the protocols and forms (and hence the database schema) in the field. Thus, any digital data collection and database systems used in the study's design phase must have easy update functionality.

The third phase (data entry) is usually a separate step for the researcher, with data transferred from paper to electronic intake forms, sometimes months after the data's initial acquisition. If data errors are discovered, the researcher cannot always return to the field to gather new data, so data analysis strategies or even research objectives might be rethought or data extrapolated. There is currently little automated data validation at this phase, although early data validation could enhance the research.

During the fourth phase (data analysis), researchers often reformat data for analysis or modeling or place intermediate results into data sets. This is time-consuming, especially if scientists are not experienced with the analysis or modeling program. At this point, the researcher could still discover that key parameters or data are missing and return to the field, use another researcher's data (collected for the same or a different purpose), extrapolate or interpolate existing data, or not conduct an originally envisioned analysis. Database technology would make these problems

obvious earlier and facilitate data transformation and sharing of macros, parameters, or scripts for applications. We are developing visualization programs using the Visualization Toolkit (http://public.kitware.com/VTK) that can display the data represented in our database components. We also are experimenting with a spreadsheet application that finds tree physiology data outside local minima or maxima range.

The fifth phase (collaborator data sharing) brings the research group as a whole into the picture. Subsequent to data verification and preliminary analysis, some laboratories require researchers to save and document data sets to a common within-lab data store in formats readily understood by others in the laboratory. This avoids potential loss of field data if a researcher leaves, for example. Documenting data, even for an audience familiar with the research objectives and protocols, takes time, however. Database technology could facilitate this process, especially in larger laboratories where many researchers focus on related problems and have common research protocols, vocabulary, field procedures, and instrumentation.

The sixth phase (journal publication) is how researchers receive feedback from peers, advance the state of the art, and acquire continued funding. Database tools can help in manuscript preparation in several ways—for example, by easily generating graphics or by gaining access to scientific citations.

Data documentation and validation are the major bottlenecks to the seventh phase (data archiving).[11,16] Deciding which data to archive, writing data descriptions, and validating the data against those descriptions seems disconnected to individual researcher goals. These tasks often are so overwhelming that researchers usually do them at the end of the research cycle, even though data documentation could provide useful artifacts if applied earlier.

Another major bottleneck to data archiving is that some scientists hoard data—they hesitate giving up their data for others to use. Concerns about being appropriately credited and having their data properly used are sociological problems being addressed by funding agencies, long-term research sites, and some journals. The Ecological Society of America, for example, publishes peer-reviewed data sets, but many scientists feel such publications do not carry the same weight as archival journal publication in terms of recognition and career advancement. Although technology cannot change the sociology of data archiving, applying database technology early in the research cycle could make data documenting considerably easier and turn data archiving into a matter of pushing a button that says, "publish this database."

The final stage (data mining) is still usually accomplished on a person-to-person basis, with some notable exceptions such as weather data, satellite maps, or data sets published for permanent sites such as the Long-Term Ecological Research Network's (LTER; http://lternetwork.edu). Even where data are published electronically, though, few data standards exist in ecology. Two data sets gathered from a single archive might have different data formats or worse, different semantics (even with identical data variable names). Community-maintained common vocabularies, common data formats or components, and data integration tools could help, and several other promising projects also address these problems; see, for example, the Science Environment for Ecological Knowledge Project (http://seek.ecoinformatics.org).

Data documentation and validation are the major bottlenecks to data archiving.

The next two sections of this article describe the Canopy Database Project's efforts at increasing researcher productivity with small database-like tools. We first describe the development and use of a handheld field data acquisition tool, which would find straightforward integration with our database design tool, DataBank. Then we describe an effort to help a lab use data management tools and best practices to carry out within-lab data documentation and archiving. This initial documentation for close collaborators, we postulate, could help automate data validation and render later metadata provision for archives significantly less intimidating. The ideas emanating from this second effort will also be integrated into future versions of DataBank.

A Handheld Data Acquisition Tool

Because ecological field data are typically acquired by hand and fraught with numerous transcription errors, we experimented with digital data collection. Our aim was to determine, first, whether such a tool would be useful to canopy researchers and, second, whether the technology developed for one project could be cost-effectively applied to other projects. Here, we describe our efforts to use a handheld data acquisition tool

to partially automate the collection of field data. We also wanted to integrate the use of databases with the handheld tool.

Ideally, a canopy field project would streamline data handling to integrate all phases of the research process. Study design might include field methodology and technology, which would in turn make data entry amenable to immediate data analysis and sharing. Imagine, for example, field protocols that use instruments to record data as they are collected and provide access for data analysis and real-time communication to a data warehouse. In practice, however, canopy scientists treat research phases as separate tasks, so there are impediments to such sharing or even analysis. One of us (Roman Dial) had an ideal project for applying database technology, in part because he had already used digital devices, but more importantly because he recognized the possibility of improving productivity.

Dial's work quantified the forest canopy's structure, particularly the air-space gaps between trees. Forest canopy data are intrinsically three-dimensional: 3D locations are typically defined using Cartesian coordinates (x, y, z). In canopy science, the x and y values give the planar projection on the ground, with the z value representing the height above ground. In principal, these values should be easy enough to collect, but in practice, locating a canopy element's specific position can be difficult. (A canopy element is a vegetative structure located above ground, such as a tree stem, tree foliage, epiphyte, and so forth.) Instead of Cartesian coordinates, Dial used cylindrical coordinates (r, θ, z), which measure z as the distance above ground of an observer looking at a canopy element (itself located at eye level) and in compass direction θ but a distance r away from the observer.

Given available technology, cylindrical coordinates are easier to collect than Cartesian ones. First, the distance above ground, z, requires a tape measure to be stretched from the ground to the observer. Second, the direction θ, taken from the observer to the canopy element, requires only a magnetic compass (Dial used a digital MapStar model from Laser Technology). Finally, a laser range finder measures the distance r to the canopy element from the observer (Dial used a digital Impulse 200LR model from Laser Technology). The Impulse laser and MapStar compass each report digital measurements in a downloadable form, and a PDA could, if properly configured, automatically receive the data.

Prior to this project, Dial's team used digital laser range finders, but like most ecologists, they recorded data manually, with a pencil and notebook. Three different individuals handled each piece of data: the observer, the recorder, and the data typist. In the canopy, the observer aimed and fired the laser, then called down to the ground-based recorder the observed element identity and its position values r; θ, and z. The recorder wrote the information onto a data sheet. Data entry personnel later typed this data into a computer. Transcribing numbers in the canopy can be awkward: usually only one hand is available to record data while the other uses an instrument to measure state variables or steadies an observer dangling in space. Moreover, recording data by hand and later (perhaps weeks or months later) transcribing them into a digital format is error-prone.

Capturing the data digitally would, we reasoned, eliminate two steps from the process: manual recording and data entry (both the MapStar compass and the Impulse laser range finder produce digital data). Thus, the one person holding the laser and compass could measure, record, and store in a digital database each observation with one push of the button as it was observed. This would increase productivity, decrease error propagation, and eliminate the need for a recorder and a data typist, thereby reducing travel, field costs, and time.

Dial approached Nalini Nadkarni and Judy Cushing (the Canopy Database Project's directors) with his idea of recording compass and range finder data digitally to a PDA. Together, we designed and developed an "electronic data sheet" for the PDA that seamlessly integrated the cylindrical coordinates and canopy element data with off-the-shelf database tools. Dial used the materials and methods we describe next during fieldwork in 2002, when he collected 3D data in two forest canopies: one a tropical rain forest characterized by high heat and humidity, the other a temperate Eucalyptus forest characterized by strong winds (up to 30 mph), and both characterized by great height (Dial routinely sampled to 250 feet above ground). He used the instruments to collect spatial data (defined as the distance and area between canopy elements) and statistical frequencies of canopy elements in both forests. He also used the PDA's database structures to record dates, times, and locations of sensors that measured light, temperature, and relative humidity, as well as the dates, times, and locations of nylon trays positioned to capture the rain of arthropods killed with insecticidal fog.

Research Methodology, Access, and Instrumentation

To collect his data, Dial first positioned himself in the canopy by stretching horizontal ropes between tall trees (termed *emergents*) that extend above the level of neighboring crowns. Often the emergents were so tall that Dial's horizontal ropes were suspended above the level of the forest between them. From these horizontal lines, Dial attached vertical climbing ropes, which served as *transects* (linear sampling units) that reached to the ground, allowing him to sample the full forest height. The vertical transects were located at 5- to 20-meter intervals along the traverses. In essence, Dial's samples consist of a systematic, vertical cross section through a forest.

Palmtop Computer (PDA)

Several PDAs are currently available and most are quickly improving in terms of features and power. Dial used the Palm m105. It was relatively inexpensive (US$150), provided an 8-Mbyte memory, and used replaceable batteries and plastic instead of metal and glass for its construction. The m105 also offered Palm's Grafitti handwriting recognition software. Although Dial never lost data, changing batteries on PDA units in the field sometimes reset those units, which could have meant lost data and programs. A spare unit reset itself three times during two months, even without removing the batteries. Dial had no moisture problems or breakage, despite hundreds of hours in the field and exposure to high winds and humidity. Because the PDA was enclosed in a padded, waterproof case with a clear vinyl window, neither the screen nor the vinyl window fogged, even when used during tropical downpours.

PDA Software

Dial used the Palm PDA's installed software and special-purpose PDA software. Of the Palm's installed software, he used DateBook to plan and record tasks by time and date, NotePad for sketches and other freehand notations, and MemoPad for longer observations. He automatically uploaded the resulting data to his laptop, usually once a day for integration into larger data structures or reports after hot-synching. The advantage of using packaged software was that it eliminated the step of entering hand-recorded data into the computer; the data were recorded electronically and directly on the PDA.

Dial also used AppLaser, a special-purpose application developed by the Canopy Database Project team jointly with Dial, which recorded and uploaded canopy data directly to an MS Access database. AppLaser's requirements include the ability to

- upload digital location data from the MapStar compass and Impulse laser to the Palm along with height above ground, thereby recording cylindrical coordinates (r, θ, z),
- date-stamp data automatically, and
- record hierarchical location data and insert it into an MS Access database.

Figure 3. Ecologist Roman Dial collecting canopy element data in a 250-foot tall Eucalyptus forest in Australia. Dial is suspended by a vertical transect rope as he collects and records data using a waterproof PDA. The PDA is connected to a digital compass and digital laser range finder. These surveying instruments automatically download location measurements into the PDA while Dial uses the PDA's handwriting recognition software to record the located object's identity. (Photo courtesy of Bill Hatcher Photography.)

Because canopy data collection involves canopy element identification and location information, Dial required entry of string text in the application. The composition of canopy elements changes through the canopy, so he also needed a drop-down menu that he could modify with a stylus as he ascended through the canopy (see Figure 3). The data also needed to fit into a common database program and thus be amenable for use by other canopy researchers involved in the Canopy Database Project.

AppLaser Data and Data Entry

AppLaser has five menu screens (denoted here by italics): The main menu lets users select

among four subscreens. Each subscreen corresponds to a data record; the screens together define a hierarchical database in which a *Study_Area* contains one or more *Transects*, each of which contains one or more *Obs_Points*, each of which contains one or more *Obs_Measurements*. Each of the four subscreens has a variable number of fields, some of which are offspring of parent screens. Values of the parental screens are automatically recorded as the user descends the hierarchy. Thus, for example, as the user navigates down to *Obs_Measurement*, the *Study_Area*, *Transect*, and *Obs_Point* values are automatically applied. It is not possible to navigate down without defining the parent record.

The lowermost of the hierarchy, the *Obs_Measurement* screen, includes the two fields automatically filled by the surveying instruments. The field "distance" (to canopy element, *r*) is the measurement sent to the PDA by the Impulse laser and "azimuth" is the compass direction to the element sent by the MapStar compass. The "date" field is a time stamp of date and time information provided by the native PDA operating system. All other non-offspring fields are filled by the observer using Graffiti or via pop-up menus (offspring fields are populated using values from ancestor screens). A user can add, edit, browse, or delete any data record using screen buttons and Graffiti. Offspring fields cannot be altered.

Fundamentally, the data structure records the canopy element's identity and its location information in time and space. The canopy element's position is located globally via the Universal Transverse Mercator (UTM) grid system (in contrast to latitude and longitude). In addition, there are opportunities to record opportunistic observations as comments.

According to Dial, after a day in the field, the PDA feels extremely precious—and it is. It holds an entire day's worth, or more, of hard-won data. Ecologists raised on yellow notebooks and pencils will feel particularly naked. Eventually this feeling passes, but only after hot-synching the PDA to an office, lab, or laptop computer after every data collection session and then copying the resulting databases onto backup media. Hot-synching AppLaser data populates four relational database tables in MS Access: *Study_area*, *Plot_location*, *Observation_point*, and *Observation_measurement*. Each record in an Access table corresponds to a data record that the PDA records. Each hot-synch updates Access tables by appending new records at the end of each database table.

Serendipitous Uses of AppLaser

Dial discovered that AppLaser was robust and flexible enough to handle data other than the spatial data, as originally envisioned. In tandem with spatial data in forest canopies, light data, temperature, relative humidity, and arthropod abundance and diversity are of interest to canopy researchers. Thus, in addition to receiving distance and azimuth signals, he recorded microclimate data and names of arthropod fogging trays by using the "comment" field in *Obs_Point*.

Dial used the common *Study_Area* name for four variable types (space, data logger location, light, and arthropod tray location). However, although each was collected physically along the same vertical rope, each represents a different transect. This means that even though they share the same "bearing," "X," and "UTM" fields, they must have different transect names. Thus, the user might record for transect "T7"— where spatial data were collected at a bearing of 27° and 35 meters from the origin tree (the one whose UTM coordinates define the *Study_Area*'s location)—transects "T7light," "T7temp," and "T7trays." Under *Obs_Point* within transect "T7light," Dial recorded height in "Z field" (where the light measurement was recorded) and the amount of light in the "comment" field. Similarly, in transect "T7temp," he recorded height in the "Z field." The benefit was that these additional observations were automatically recorded into the study's database when hot-synched.

Work Remaining and Lessons Learned

Conceptually, AppLaser proved satisfactory for Dial and his team. A few design errors should be rectified before the software is made available to others:

- *Inclination*. If the "distance" field of *Obs_Measurement* were filled by the laser range finder's slant distance variable, then each observation point would become the origin in a spherical coordinate system. A canopy worker could use this to observe all around, not just in an assumed flat plane with fixed height above ground.
- *Ability to add a new element to the pop-up menu on the fly*. The researcher must currently back out of the *Obs_Measurement* screen to the main menu, then go into a pop up to add a new element.
- *Change of the delete key's position*. When using the location page, research assistants occasionally

tapped Delete when they wanted to tap New, and as a result had to retake the lost data.
- *An index that counts the number of measurements taken at a given Obs_Point and displays this count on the Obs_Measurement screen automatically*. Often a researcher knows enough to take *n* observations per observation point, but won't often know if he or she actually did so. Dial counted measurements using Next and Prev buttons, but this is awkward, slow, and error-prone.

The following new AppLaser features would be advantageous:

- On the *Obs_Point* screen, each line of the "comment" field could be its own field, allowing for multiple entries at a given height that could be downloaded separately to MS Access.
- There should be a separate pop-up menu for the "Found On" field on the *Obs_Measurement* screen. "Found On" would be useful for habitat studies, and it needs its own pop up.
- A Delete key on the main page that completely purges the database could improve speed immensely. The database on the Palm needs periodic cleaning because as it fills, each new observation takes longer to appear, and the extra data slow the application. Right now, the user must press Delete repeatedly to purge the database.
- Better data management on the PDA and better coordination with the MS Access database.
- Improving the quality of the cables connecting the laser finder, compass, and recording device via the RS232 ports would improve the device overall. Although waterproof, attaching or removing them while either the laser or compass is turned on could damage them.

Currently, AppLaser is only programmed to hot-synch on a Windows PC. Dial never experienced problems with transferring the PDA data into MS Access on his laptop PC. Although he usually exported data from MS Access to MS Excel and then into SPSS (a statistics package) or as raw text files for use in Mathematica, we see significant opportunities for improved data validation and management using MS Access database capabilities.

Equipment, contract programming, and testing in the field cost approximately US$50,000 (including Dial's salary). Dial's primary canopy notebook is now a PDA. Although other researchers are less eager to use PDA technology, Dial is confident that canopy science will grow into an extra-arboreal-centered activity and that such tools will make data more productively collected, less error-prone, more easily analyzed, and more readily shared among researchers.

Including an experienced ecologist such as Dial, who acted as a software designer and change agent, has been a valuable asset to the Canopy Database Project. We learned how to integrate field data iteratively into a database during a field session and gained insight into data validation. Dial's use of general data structures to record serendipitous measurements such as microclimate data and names of fogging trays was particularly useful to our DataBank work on a generalized observation data structure. Such generalized structures allow needed flexibility to accommodate changes in field data collection protocols, and let researchers specialize generic software for their own use, thus decreasing costs of future systems.

Within-Lab Metadata Acquisition and Archiving

Unless metadata provision becomes less onerous and more obviously helpful, scientists will continue to balk at archiving their data. To address this issue, we investigated metadata and data-archiving process within one laboratory. We used inexpensive database and spreadsheet tools (MS Access and Excel) to ease the burden of documenting field data sets, which are uploaded to a shared store (an in-lab archive) at key times during research projects. A long-term advantage is that the in-lab metadata will be a first step in archiving the data.

Barbara Bond's lab at Oregon State University conducts research categorized as forest ecophysiology. Bond and her group study how species, community structure, climate, and developmental age affect exchanges of matter and energy between plants and the environment. These interactions occur at many different scales of time and space, ranging from the subcellular level at time scales of a few seconds to the watershed or regional level at time scales of centuries. The data collected are important to nonphysiologists, because researchers investigating fundamental questions of global climate change and biocomplexity rely on physiological information. Thus, timely archiving of ecophysiological data is critical, and, although researchers often ask Bond for data, the documentation required before distributing those data is difficult and time-consuming for those in her lab.

Figure 4. Part of a field installation for a study of environmental controls on ecosystem-scale physiological processes. A student is installing a sensor to continuously monitor the flow of water through the sapwood of a Douglas fir tree. (Photo courtesy of B. Bond.)

Another reason why data documentation is a critical aspect for Bond's lab is that to answer a question in tree physiology, her researchers must piece together several types of information from different sources. Many data sources are used in a typical study, and lots of studies ongoing in the laboratory at any given time, so data sources for concurrent studies typically overlap. A typical small project might use the following information:

- leaf area and biomass information for grasses and other small plants, shrubs, and trees, each with different sampling designs due to differences in scale;
- meteorological data from three measurement stations, each with five instruments of different accuracies, sampling frequencies, and site-specific details;
- measurements of sapflow in trees of different sizes, ages, and species;
- measurements of stomatal conductance performed periodically through growing seasons;
- measurements of species composition and distribution in small watersheds, obtained from vegetation sampling plots; and
- measurements of soil water content using two instruments, one with continuous output and the other sampled manually and periodically.

A complex study could have more than 20 different sets of measurements continued over several years, with modifications in measurement protocol as well as personnel. Some of these data are collected digitally via field instrumentation; others are collected by hand.

In most cases, the data require complex processing before they are useful. For example, biomass information involves combining published allometric equations from other locations with onsite field measurements. Another example comes from sapflow measurements. To measure sapflow, you place thermocouples and heating sensors in trees; the raw data is a stream of temperature differentials between heated and unheated probes (see Figure 4). To interpret this information, you first use algorithms to convert from temperature to sap-flux density. Usually, some information is missing due to faulty sensors; this information is "filled in" statistically. Unless researchers can track changes to the data set, the final mean values for these continuous measurements will show odd abrupt blips as the underlying sample set changes over time. Having a permanent record of this kind of data manipulation is important, but in reality, the details are sometimes lost. Procedures are needed to document the steps of data processing for interpretation months or years later without creating a huge burden for the student or technician doing the initial work.

After filling in missing data, algorithms are devised to "scale up" from the individual tree to the stand level to convert from the amount of water flowing through a tree to the amount of water flowing through a group of trees covering a given ground area. At each step, small but difficult-to-document errors are introduced. In much of the currently published ecophysiological work, these potential errors are seldom acknowledged.

A question this ecophysiology lab faced is how to document data collection and processing, often unique to each data set, without writing prohibitively large volumes of support material. How can researchers be sure that the data they archive are used appropriately? How can they fit data management activity into an already extremely tight schedule? Sharing data about that could require many hours just to explain data idiosyncrasies.

We wanted to explore how to better archive data in the lab and how to capitalize on local sharing to make it easier to later document those data for the external world.

Using an LTER metadata standard (the H.J.

Andrews LTER at www.fsl.orst.edu/lter/data/metadata/guide.cfm?topnav=115), we wrote an MS Access database application to record metadata for Bond's laboratory. Source tables for many of the metadata fields are allowed to grow so that data sets subsequently documented can use those previous descriptions. The database application divides metadata into four kinds:

- *Personnel information*. Information about the people involved in the conception, design, implementation, and documentation of a study and its data tables.
- *Study-level information*. General information about a study such as its title, abstract, purpose, dates, methods, and site characteristics. When a study is not part of a project, the study-level information will also include general information such as use constraints.
- *Entity- and attribute-level information*. Detailed information about individual data tables or files that contain GIS layers or images. This is typically what computer scientists call metadata, but ecologists might call it table-level metadata.

Some existing tools (in particular MetaCat and Morpho[17]) allow metadata documentation and browsing of ecology data sets, but they were not yet available when we needed them. We intend either to phase out our MS Access database application and use MetaCat or to have our application produce data in the Ecological Markup Language, which MetaCat's designers have defined.[9]

We developed an MS Excel spreadsheet program that lets users highlight a table in a worksheet and then reads column headers as variable names from this table. Once a user highlights a table, a new worksheet is created, and the user is queried for study- and table-level metadata. Metadata thus becomes available in the spreadsheet with the data, and the spreadsheet metadata can later be uploaded to the MS Access database. Another spreadsheet program uses these metadata to look for possibly erroneous data.

These applications are currently installed in Bond's lab, where they are being used by graduate students. Working with a laboratory of cooperating researchers has let us experiment with making simple mechanisms for metadata provision available early in the research cycle, yet did not require researchers to drastically change how they deal with their data. We gained important insights from working with Bond's lab on within-lab metadata acquisition and archiving. However valuable database technology might be for documenting a data set for posterity, or even for linking it to collaborators' data sets in an integrative study, scientists will not use that technology unless it increases individual researcher productivity or (as in Bond's lab) provides perceived benefit to a close-knit group of peers. Even in the latter case, data documentation tools should be specialized to the particular science practiced. For example, in Bond's situation, we stage the metadata provision from very simple and informal to more complex and generally applicable. We now believe it is possible to conduct rudimentary data validation using preliminary metadata. We are helping the lab develop lab-specific source tables for research information, keywords, research sites, and instrumentation that are consonant with long-term archiving. In short, this work has helped us better understand how to specialize database tools for related ecological studies.

We have described the need for new ecosystem informatics tools, the ecology research cycle, and two small projects at different stages of that cycle. The first project, a handheld (PDA) data acquisition tool, has numerous benefits and seems sustainable. Next steps would be to connect the tool to a database application that performs validation, visualization, and analysis at further stages of the research cycle, and to build tools that specialize PDA forms for particular studies. The second project, an effort at within-lab metadata acquisition and archiving, shows that metadata provision could be less onerous if accomplished in stages. An obvious next step would be data validation and cleaning at the lab level using those metadata, and transferring metadata from our tool directly to data managers at longer-term data archives, such as H.J. Andrews LTER. Both of these enhancements to the within-lab project are under way.

The two pilot projects have convinced us that current technology can help solve short-term problems, but it can't produce the integrated database systems ecologists need for the future. Furthermore, these "one up" applications are generally not cost-effective for single research studies. The research that will deliver this future

technology will require that ecologists and computer scientists work together—alternating between development that uncovers and defines problems, development that solves those problems in particular contexts, research that generalizes key applications, and further work that tests research prototypes in new contexts.

We believe other scientific disciplines would benefit from using the kinds of database technology we describe but emphasize that the state of technology is currently such that researchers need to learn database design beforehand. Because few scientists want to become database programmers, we believe that end-user database design tools and turn-key applications should be made available at the subdiscipline level. A separate project under Cushing's direction at Evergreen is considering ancillary problems of using database technology to make easier the use of compute-intensive applications.

Finally, although improved researcher productivity is a necessary condition before ecologists will use database tools, it might not be sufficient for widespread adoption of those tools. Moving systems such as those proposed here into the research cycle will inevitably involve some changes in the way ecology is practiced. Although such sociological changes are beyond the Canopy Database Project's scope, our work suggests that both ecologists and computer scientists will play key roles as these rewards are introduced in the scientific arena.

Acknowledgments

We acknowledge significant contributions of Canopy Database Project programmers and research staff Erik Ordway, Mike Ficker, Steve Rentmeester, Abraham Svoboda, Alex Mikitik, Youngmi Kim, Janet Rhoades, Peter Boonekamp, James Tucker, Brook Hatch, and Neil Honomichl. Information Managers Don Henshaw and Gody Spycher at the NSF LTER H.J. Andrew site, as well as its former director Susan Stafford and the national LTER information management team, provided considerable help by providing metadata standards and understanding of ecological information management. We thank our research collaborators (including, but not limited to) Lois Delcambre, Jerry Franklin, Mark Harmon, Hiroaki Ishii, Betsy Lyons, Dave Maier, Robert Mutzelfeldt, David Shaw, Steve Sillett, Akihiro Sumida, and Robert Van Pelt for freely sharing insight and data.

Ben Bloodworth, Jeff Heys, Patrick Boyne, Andrew Lee, Steve Sillett, Jim Spickler, Betsy Young, Emily Bearnhardt, and Matt Dunlap helped Roman Dial with design and field implementation. Starling Consulting programmers and project managers Eugene Ryser, Erica Frandsen, Porsche Everson, Jay Turner, and Bonnie Moonchild did most of the programming on this device and made numerous contributions to the project. Kate George and Georgianne Moore are working with the in-lab metadata acquisition tool at Oregon State and providing helpful metadata, data, and suggestions for improving the software. Travis Brooks ably programmed the MS Access and Excel applications.

NSF grants BIR 9975510, 9630316, and 9300771, INT 9981531, and EIA 131952 and 75066 supported this work. Dial's work was partially supported by the Global Science Society (GF 18-2000-114), and Bond's by the H.J. Andrews Long Term Ecological Research program and the US Department of Energy, through the Western Regional Center of the National Institute for Global Environmental Change under Cooperative Agreement DE-FC03-90ER61010.

References

1. US Nat'l Research Council, *Finding the Forest for the Trees: The Challenge of Combining Diverse Environmental Data (Selected Case Studies)*, Nat'l Academy Press, 1995.
2. US Nat'l Research Council, *Bits of Power: Issues in Global Access to Scientific Data*, Nat'l Academy Press, 1997.
3. M. Moffett, *The High Frontier: Exploring the Tropical Rain Forest Canopy*, Harvard Univ. Press, 1993.
4. M. Lowman and N. Nadkarni, *Forest Canopies*, Academic Press, 1995.
5. N. Nadkarni and G. Parker, "A Profile of Forest Canopy Science and Scientists—Who We Are, What We Want to Know, and Obstacles We Face: Results of an International Survey," *Selbyana*, vol. 15, 1994, pp. 38–50.
6. N. Nadkarni and J. Cushing, *Final Report: Designing the Forest Canopy Researcher's Workbench: Computer Tools for the 21st Century*, Int'l Canopy Network, 1995.
7. D. Maier et al., eds., *Report on a NSF, USGS, NASA June 2000 Workshop on Biodiversity and Ecosystem Informatics*, 2001; http://evergreen.edu/bdei2001, http://bdi.cse.ogi.edu, and www.nsf.gov/cgi-bin/getpub?nst0199.
8. J. Cushing et al., *Summary of VLDB Panel Database Research Issues in Biodiversity and Ecosystem Informatics*, Morgan Kauffman, 2002; www.cs.ust.hk/vldb2002/program-info/panels.html.
9. R. Nottrott, M.B. Jones, and M. Schildhauer, "Using XML-Structured Metadata to Automate Quality Assurance Processing for Ecological Data," *Proc. Third IEEE Computer Society Metadata Conf.*, IEEE CS Press, 1999; http://computer.org/proceedings/meta/1999.
10. W. Michener et al., "Non-Geospatial Metadata for the Ecological Sciences," *Ecological Applications*, vol. 7, 1997, pp. 330–342.
11. G. Spycher et al., "Solving Problems for Validation, Federation, and Migration of Ecological Databases," *EcoInforma*, vol. 11, 1996.
12. W. Michener, J.H. Porter, and S. Stafford, eds., *Data and Information Management in the Ecological Sciences: A Resource Guide*, LTER Network Office, Univ. New Mexico, 1998.
13. W. Michener and J. Brunt, eds., *Ecological Data—Design, Management and Processing*, Blackwell Science, 2001.
14. J. Cushing et al., "Template-Driven End-User Ecological Database Design," *Proc. 6th World Multiconference Systemics, Cybernetics and Informatics*, vol. 7, Int'l Inst. Informatics and Systemics, 2002, pp 361–366; http://216.72.45.230:1081/ProceedingSCI/index98.htm.
15. J. Cushing et al., "Designing Ecological Databases with Components," to be published in a Special Issue of *J. Intelligent Information Systems*, J. Schnase and J. Smith, eds., 2003.

16. J.H. Porter, D.L. Henshaw, and S. Stafford, "Research Metadata in Long-Term Ecological Research (LTER)," *Proc. IEEE Metadata Conf.*, IEEE CS Press, 1997.
17. P. McCartney and M. Jones, "Using XML-encoded Metadata as a Basis for Advanced Information Systems for Ecological Research," *Proc. 6th World Multiconference Systemics, Cybernetics and Informatics*, vol. 7, Int'l Inst. Informatics and Systematics, 2002, pp. 379–384.

Judith Bayard Cushing is a member of the faculty in computer science at The Evergreen College. Her research interests include software engineering and database infrastructures for scientific applications. She received her MA in philosophy from Brown University and a PhD in computer science and engineering from the Oregon Graduate Institute. She is a member of the ACM and the IEEE. Contact her at The Evergreen College, 2700 Evergreen Parkway NW, Olympia, WA 98505-0002; judyc@evergreen.edu.

Barbara Bond is an associate professor in the Department of Forest Science at Oregon State University. Her research interests include studies of physiological processes associated with aging in forest trees and the use of carbon isotope discrimination by ecosystems as a tool for understanding ecosystem function. She has an MS in terrestrial ecology and a PhD in plant physiology and forest science, both from Oregon State University. She belongs to the Ecological Society of America. Contact her at the Dept. of Forest Science, Richardson 313, Oregon State Univ., Corvallis, OR 97331; barbara.bond@orst.edu.

Roman Dial is an associate professor in the Department of Environmental Science at Alaska Pacific University. His research interests include canopy science and theoretical ecology. He holds an MS in mathematics from the University of Alaska Fairbanks and a PhD in biological science from Stanford University. He belongs to the Ecological Society of America. Contact him at the Dept. of Environmental Science, Alaska Pacific Univ., 4101 University Dr., Anchorage, AK 99508; roman@alaskapacific.edu.

Nalini Nadkarni is a member of the faculty in environmental studies at The Evergreen State College. Her research interests include the ecology of tropical and temperate forest canopies and the ecosystem process. She received a PhD in forest ecology from the University of Washington. She belongs to the Ecological Society of America. Contact her at The Evergreen College, 2700 Evergreen Parkway NW, Olympia, WA 98505-0002nadkarnn@evergreen.edu.