Finding the Forest in the Trees

The Challenge of Combining Diverse Environmental Data

Selected Case Studies

Committee for a Pilot Study on
Database Interfaces

U.S. National Committee for CODATA

Commission on Physical Sciences, Mathematics, and Applications
National Research Council

NATIONAL ACADEMY PRESS
Washington, D.C. 1995

4

The H.J. Andrews Experimental Forest Long-Term Ecological Research Site

The H.J. Andrews Experimental Forest was established by the U.S. Forest Service in 1948. Located in the rugged Cascade mountain range of Oregon, the 6,400-ha preserve was covered with virgin forest in the late 1940s. Since then, approximately one-third has been manipulated through logging or research plantations. Old-growth forest stands with trees over 400 years old cover about 40 percent of the area, with mature stands covering another 20 percent. Rapidly flowing mountain streams are the primary type of aquatic system.

The Andrews Forest was designated as a research site under the National Science Foundation's (NSF) Long-Term Ecological Research (LTER) Program in 1980. The Andrews site is one of 18 such sites in the United States. The goals and objectives of the LTER Program are given in Franklin et al. (1990), and the research programs and core data sets of the Andrews site are summarized in McKee et al. (1987) and Michener et al. (1990), respectively.

The LTER Program studies are designed to carry out long-term (decades to two centuries) ecological research on natural ecosystems in the United States. Their basic objectives are to study:

1. Pattern and control of primary production;
2. Spatial and temporal distribution of populations selected to represent trophic structure;
3. Pattern and control of organic matter accumulation in surface layers and sediments;

4. Pattern of inorganic inputs and movement of nutrients through soils, groundwater, and surface waters; and
5. Pattern and frequency of disturbance to the site.

Research at the Andrews site has focused on several areas, including the disturbance regime, vegetation succession, long-term site productivity, and decomposition processes. The commitment to long-term studies is evident in these areas. A good example of this is a log decomposition study, which will determine the effects of log size and quality and of the site environment on the pattern and rate of decomposition and nutrient release. In the largest and longest decomposition experiment, more than 500 logs of four species were placed at six old-growth forest sites. That study is designed to track samples over a 200-year period (Harmon, 1992).

The scientific scope of the committee's case study is limited to the interdisciplinary observational and experimental studies at the LTER Andrews site, although it also reviews the data management and institutional relationships of the Andrews site to the other LTER sites and to NSF.

The research at the Andrews site has certain key similarities to the committee's other case studies. It was intended to sample and study interdisciplinary problems and issues. For example, it includes coordinated studies in air, soil, water, and various forms of biota. The research was well under way, having begun several decades before it was officially designated as an LTER site in 1980. The data collected could be expected to be useful in global change studies and in other types of long-term environmental monitoring efforts. And, finally, a variety of investigators worked on the same general study area.

The committee was primarily interested in the data management activities of the Andrews site. These activities are managed by the Quantitative Science Group under the auspices of Oregon State University and the U.S. Forest Service. There appears to be little distinction between whether an individual works for the university or the Forest Service. This arrangement seems to work well for a number of reasons. There has been a long and close working relationship between the U.S. Forest Service Research Laboratory and the Forest Science Department at Oregon State University's College of Forestry, as well as other departments. The proximity of the two buildings housing the respective scientists also promotes good collaboration. In fact, several university researchers and staff have their offices in the Forest Service Laboratory building. In addition, there has been a history of successful preparation of joint research proposals between university and Forest Service staff. This, plus traditional attributes of working at a university, such as joint appointments and coop-

erating faculty appointments, has helped to maintain an effective working relationship.

VARIABLES MEASURED AND SOURCES OF DATA

Table 4.1 lists and describes all of the data sets collected at the Andrews site.

TABLE 4.1 H.J. Andrews Experimental LTER Site Data Sets

Data Set	Description
Dendrometer measurements in permanent reference stands	Provides an accurate estimate of volume and height for individual trees.
Respiration patterns in logs	Examines the seasonal and successional patterns of respiration losses for four dominant softwood species.
Coarse woody debris density and nutrient content	Describes the external characteristics of coarse woody debris in various decay classes and measures density and nutrient content.
Stream cross-sectional profiles	Monitors changes in channel geometry in response to storms and movement of large organic debris in a range of stream sizes.
Watershed streamflow summaries	Evaluates long-term changes in hydrology associated with various management treatments; provides baseline data for affiliated nutrient, water chemistry, and sediment transport studies; and characterizes the hydrologic regime of old-growth forests at different elevations.
H.J. Andrews watersheds 1, 2, and 3 and miscellaneous suspended sediment samples	Quantifies long-term effects of two intensities of timber harvest on sediment delivery at seasonal and yearly time scales.
Post-logging community structure and biomass accumulation	Patterns plant succession and biomass accumulation following clear-cut logging of an old-growth Douglas fir/western hemlock forest.

TABLE 4.1 Continued

Data Set	Description
Plant biomass dynamics following logging and burning	Documents patterns of plant succession after clear-cut logging and slash burning on two experimental watersheds.
Tagged log inventory	Tags and numbers woody debris and describes the following characteristics: longitudinal position, geomorphic location, log dimensions, decay class, origin, moss cover, root wad, and channel angle.
Population studies of rainbow and cutthroat trout	Assesses fish population and habitat structure in streams (150 to 300 m in length) and basins (greater than 40 km in length).
Watershed 1 and 3 plant succession data, 1962 to 1977	Documents patterns of plant succession after clear-cut logging and slash burning on two experimental watersheds.
Tree permanent plots of the Pacific Northwest	Examines rates of succession and measures mortality and growth in representative forest types in Pacific Northwest.
Stream-upland wood decay experiment	Examines and contrasts the decay of small logs in a stream channel to that on an upland site; examines the movement of small logs in a third-to-fourth-order stream.
Reference stand litterfall study	Determines seasonal and annual rates of litterfall samples at six permanent plots picked to represent a range of habitats and elevations.
Structure and composition of riparian vegetation	Measures biomass of riparian vegetation strata, characterizes phenology of leaf-out and leaf fall, and determines the spatial distribution of foliar biomass, and timing and amount of annual foliar inputs into steams.

continues

TABLE 4.1 Continued

Data Set	Description
Rainwater samples: long-term precipitation chemistry patterns	Precipitation chemistry sampled at a low-elevation site and analyzed for pH, alkalinity, conductivity, total P, ortho-P, total N, NO_3-N, suspended sediment, Si, Na, K, Ca, Mg, SO_4-S, and Cl.
Watershed grab samples: long-term stream chemistry patterns	Describes long-term patterns of nutrient output from: a first-order, old-growth watershed; a first-order watershed after clear-cutting; a second-order old-growth watershed; a second order watershed logged and burned in 1966; and a third-order old-growth watershed. Provides baseline environmental monitoring data for studying nutrient availability for stream organisms, and recovery patterns of disturbed watersheds.
Nicotinamide adenine dinucleotide phosphate (NADP) precipitation chemistry	Measures precipitation samples collected weekly for pH and conductivity on site. Samples are then mailed to a Central Chemical Laboratory and analyzed for Ca, Mg, K, Na, NH_4, NO_3, SO_4, PO_4, pH, and conductivity.
Watershed proportional samples: long-term stream chemistry patterns	Stream chemistry sampled to characterize the timing and amount of elemental losses in undisturbed conditions, and to determine the effects of logging on rates of nutrient release.
Primary meteorological station at headquarters	Provides climatic summaries and documentation for the primary meteorological station at H.J. Andrews, 1972 to present.
Climate station at watershed 2	Continuously records precipitation, relative humidity, and air temperature.

TABLE 4.1 Continued

Data Set	Description
High-elevation meteorological station	Takes measurements of air and soil temperature, soil moisture equivalency in both clear-cut and shelterwood; and solar radiation, precipitation, and wind speed and duration in the clear-cut.
Rain gauge network	Provides baseline information on variation in precipitation across a wide range of site conditions.
Air, soil, and stream temperature in various habitats in and around the Andrews Forest	Continuously monitors air, soil, and stream temperature at selected habitats.
Snow survey	Provides a baseline for characterizing variation in snow depth, moisture, and duration in the western Cascades for hydrologic modeling and to distinguish the differences in the microclimates of dominant plant communities.
Plant component biomass equations and data for the Pacific Northwest	Contains data on biomass, leaf area, and sometimes other measurements of plants collected in the Pacific Northwest.

Source: Michener et al. (1990).

DATA MANAGEMENT AND INTERFACING

The LTER concept is to have a network of intensively studied sites around the United States in various ecosystems, all measuring similar parameters and studying similar ecosystem processes. Long-term monitoring of selected environmental parameters is also a major objective (see Institute of Ecology, 1981). The LTER Program, while meeting some of these goals, has developed more along a principal investigator driven agenda with all the diversity of research that implies. A major reason is that NSF has not posed any specific research questions and only very broad goals for the various LTER sites. Therefore, the individual investigators have considerable autonomy in setting their research agendas. Consequently, the Andrews LTER as well as the overall site program is a collection of individual-investigator projects tied together by a series of

conceptual models. The committee found this to be an important factor in reviewing the data management and integration activities there.

At the heart of the data system for the Andrews site is the Forest Science Data Bank (FSDB). This database began to be developed during the International Biological Program, before the Andrews Forest was designated as an LTER site. It has benefited from the direct participation of many scientists interested in conducting regional research on the structure and function of the forest and stream ecosystems and their response to natural disturbances, land use, and climate change. Currently, 50 scientists from several institutions participate in this effort. FSDB houses 2,400 data sets from over 350 studies (databases) and adds data from about 20 new studies a year. The data are organized in 11 categories, such as hydrology and vegetation management. The total volume of the ground observation data is less than 300 megabytes, with approximately 200 gigabytes of remotely sensed data. Over $100,000, representing a significant fraction of the program's total research budget, is devoted to information management support for FSDB.

FSDB resides on a local-area network server. Local users have on-line access to the server and a set of coupled central catalogs. These catalogs contain information on the nature of the studies, their purpose and goals, their data collection activities and their periods, parameter lists, location information, experiment design, and many other relevant factors. The coupled catalogues allow search and cross-referencing for the purpose of locating the data sets that may be of potential interest to a researcher. Actual data and metadata (e.g., definition of a variable, minimum and maximum values) are stored in separate subdirectories for each study. New features built into the system allow automated export of data into the analysis systems, such as Geographic Information System or statistical analysis tools, for further processing.

The management of data in FSDB has certain characteristics that are fairly typical of ecological research. For example, data sets tend to be small and highly diverse, there is a tendency to keep data sets at the individual-investigator level, and the methodologies used in obtaining and managing the data are diverse and not necessarily standardized. As the LTER Program has progressed, the value of these disparate data sets has increased, not only for the originating principal investigators, but also for the co-investigators and other scientists, who have begun to integrate multiple data sets.

These factors have tended to help the development of FSDB. The managers of the data system stressed that the biggest incentive for scientists to use the system is improved access to one's own data, as well as better access to other researchers' data, both at the Andrews site and across all of the LTER Program's sites. Because of the diverse nature of

the data sets in general and the long-term nature of the research, the data must be well documented to ensure not only that principal investigators can use the data, but also that future researchers can understand how the data were taken.

The existence of two complementary demands—the long-term nature of the data collection and the diverse nature of the data sets and collection methodologies—has led to the development of a sophisticated metadata management system. The committee was impressed by the time and effort spent by the Andrews LTER project in this area.

This improved access to the data by the researchers associated with the Andrews site, however, has also presented a problem of ownership to the data managers. While individual principal investigators have seen the advantages to obtaining their colleagues' data sets, they also have perceived the danger of unauthorized access to their data. Therefore, the managers of the database have built in a safeguard that allows a principal investigator to veto any use of his or her data. The amount of data that the data managers can actually release on their own authority is quite limited. Nevertheless, such restrictions have been replaced by federal regulations that require all data collected with federal money to be made publicly available no later than 2 years after collection. In reality, this issue has not proved to be much of a problem. Within the initial 2-year time frame, the only people who generally would know of and want the data set were researchers associated with the project. Therefore, the rules governing data distribution remained in effect without difficulties. During the few times when outside groups asked for data, the requests were accommodated.

The data collected in the early stages of the LTER Program, including the Andrews site, are more difficult to obtain along with adequate metadata. This situation has improved, not only because of the reasons given above, but also because an increasing reliance on mathematical models has encouraged the use and interpretation of a variety of data sets. The data management team as well as the principal investigators now try to anticipate data management issues at the beginning of each individual study project, including the incorporation of metadata support.

The committee supports use of the Andrews site approach that emphasizes the creation and electronic distribution of metadata catalogs as an appropriate first step toward better integration of the other LTER sites. This step in isolation, however, does not ensure evolution toward a fully accessible and optimal data system.

With regard to institutional issues, there appear to have been few problems between U.S. Forest Service and Oregon State University personnel in collaborating on this project. NSF has been very supportive

and, in general, is gently pushing the other LTER sites to greater coordination, including intercalibration of data collection and study techniques. The local institutional aspects thus appear positive and headed in the right direction. The same features that encouraged good working relationships, as described earlier in this chapter, are responsible for helping to develop a good data management system as well.

With regard to NSF's overall management of LTER sites, interaction among sites is hoped for, but certain impediments get in the way of significant site interaction. First, each site needs to compete for funding every 6 years and, in a sense, is always in competition with other existing or potential sites. Second, until recently, there was no mechanism for funding to cooperate across sites. This, however, appears to be changing because of a realization that there is now a large body of data from different ecosystems in the United States and some means needs to be developed to provide access to these data sets.

NSF is now trying to encourage data sharing more actively among LTER sites. For example, the agency has funded an LTER data management center at the University of Washington to facilitate exchange among the principal investigators at the different sites. That consists primarily of a rapid means of e-mail communication, directories of addresses of other investigators, and the generation of a data index catalog that anyone can use to see what kinds of data sets are available and where they can be obtained.

NSF also is encouraging the development and interchange of data through funding allocations and equipment grants. For example, there is a small amount of money available to support intersite data management. In addition, NSF has facilitated the development of local-area networks, other wide-area networks, and high-capacity data storage and has provided funding for GIS equipment to help develop data management capabilities. No further technical standards beyond the Minimum Standard Installation for the LTER internetwork effort are anticipated for the near future.

LESSONS LEARNED

The Forest Science Data Bank is an excellent example of a scientific information management system that has been created and has gone through several evolutionary phases within an academic environment. During the early phases of development, the system evolution was dominated by the desires of individual researchers without major attempts at coordination, integration of functions, or identification and definition of high-level system requirements. The lack of an architecture that usually results from such an approach, along with the desire of scientists to minimize data entry and file storage costs, resulted in an unstructured system

and led to difficulties in system maintenance and enhancement process. System upgrades proved to be especially costly and cumbersome, not because of the cost of the new hardware and software, or inconsistent cooperation between the researchers, but because local optimizations had led to structural flaws, such as absent or incomplete metadata, or lost and/or incomplete files. These and other similar deficiencies were difficult to detect, and when detected were difficult to correct.

Centralized information management support by a group of competent individuals grasping both science and data management issues has been the start of a new and successful phase in the evolution of FSDB. The activities of this phase have brought discipline to the collection and organization of the data and metadata and have improved users' access through relational catalogs, which can be searched and cross-referenced. The cost of these activities, however, is not trivial, running at about 20 percent of the total research budget.

Even though FSDB has made positive steps, it should not be considered a modern state-of-the-practice scientific data system. The system lacks a modern users' interface, has limited access capability, is made up of a large number of small data sets, and will probably continue to be costly to maintain and upgrade. Because of the small number of principal investigators (the primary users), these shortcomings have not posed a serious operational problem so far. The situation, however, could become an issue when more widespread access by other LTER sites is required. On the positive side, and as far as collection and organization of metadata are concerned, FSDB should be considered an excellent model. Considerable effort has been and continues to be devoted to the standard format and automated data entry procedures for metadata. These steps have led to less time-consuming efforts by the researchers and a better organized set of very useful metadata.

REFERENCES

Franklin, J.F., C.S. Bledsoe, and J.T. Callahan. 1990. Contributions of the Long-term Ecological Research Program. *Bioscience* 40(7): 509-523.

Harmon, M.E. 1992. *Long-term Experiments on Log Decomposition at the H.J. Andrews Experimental Forest*. USDA Forest Service General Tech. Rep. PNW-GTR 280, Portland, Ore.

Institute of Ecology. 1981. *Experimental Ecological Reserves: Final Report on a National Network*. The Institute of Ecology, Indianapolis, Ind.

McKee, A., C.M. Stonedahl, J. Franklin, and F. Swanson. 1987. *Research Publications of the H.J. Andrews Experimental Forest, Cascade Range, Oregon, 1948 to 1986*. USDA, Forest Service, Pacific Northwest Research Station, General Tech. Rep. PNW-201, Portland, Ore.

Michener, W.K., A.B. Miller, and R. Nottrott. 1990. *Long-term Ecological Research Network Core Data Set Catalog*. Belle W. Baruch Institute for Marine Biology and Coastal Research, University of South Carolina, Columbia, S.C.