Understanding Environmental Complexity through a Distributed Knowledge Network

SANDY J. ANDELMAN, CHRISTY M. BOWLES, MICHAEL R. WILLIG, AND ROBERT B. WAIDE

Understanding environmental complexity and other dimensions of ecological systems necessitates a holistic approach that can be achieved only by identifying, retrieving, and synthesizing diverse data from distributed sources; by collaborating with other scientists from a broad range of disciplines; and by investigating many different systems. Knowledge Network for Biocomplexity (KNB) is developing new software tools to advance ecological understanding through discovery, access, retrieval, and management of distributed and heterogeneous ecological and environmental data. To address the need for cultural change in ecologists and other environmental scientists and to promote collaborative and synthetic approaches, KNB and the National Center for Ecological Analysis and Synthesis are training a cadre of young investigators in techniques for the management and analysis of ecological data, with emphasis on multiscale integration and synthesis.

Keywords: biocomplexity, education, computers in biology, ecology, environmental policy

Traditionally, ecological understanding has advanced through empirical observations or manipulative experiments conducted at relatively small scales by one or a few investigators (Kareiva and Anderson 1988, Brown and Roughgarden 1990). Synthesis of existing ecological knowledge commonly has involved comprehensive reviews of theory or compilations of data within a single subdiscipline (e.g., population biology, community ecology, ecosystem ecology). In contrast, current priorities for ecological research, such as the need to understand biocomplexity (Michener et al. 2001) and other environmental properties at broad spatial and temporal scales, have created a demand for multiscale, interdisciplinary environmental synthesis involving the integration of ecological information and ideas across a range of spatial, temporal, and organizational scales. Despite this need, undergraduate and graduate education primarily prepares students for intradisciplinary rather than interdisciplinary research, and the reward systems at most universities are designed to hire and promote individuals mainly on the basis of records of scholarly achievement that reflect research at the center of disciplinary domains, conducted by one investigator or by a few collaborators (e.g., Nature 2003).

To facilitate the process of environmental synthesis, the National Center for Ecological Analysis and Synthesis (NCEAS), the Long Term Ecological Research (LTER) Network office at the University of New Mexico, the San Diego Supercomputer Center, and Texas Tech University are collaborating to create a Knowledge Network for Biocomplexity (KNB; *http://knb.ecoinformatics.org*). KNB is developing software tools to advance ecological understanding through remote discovery, access, retrieval, and management of ecological and environmental data. To facilitate evolution in the culture and conduct of the environmental sciences, we are training a cadre of investigators in cutting-edge techniques for the management and analysis of ecological information, with particular emphasis on multiscale integration, synthesis, and analysis. Our approach involves a series of coordinated and distributed graduate seminars conducted simultaneously at multiple universities throughout the United States (figure 1).

The distributed course

We adopted and modified a novel distributed-seminar model that was used successfully in two previous NCEAS working groups (Savage 1998, Kareiva 2002). Here we describe the

Sandy J. Andelman (e-mail: andelman@nceas.ucsb.edu) is the deputy director, and Christy M. Bowles is a research assistant, at the National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara, CA 93101. Michael R. Willig is a professor of biology in the Ecology Program, Department of Biological Sciences, and in the Museum of Texas Tech University, Lubbock, TX 79409. Robert B. Waide is executive director of the Long Term Ecological Research Network office, University of New Mexico, Albuquerque, NM 87106.

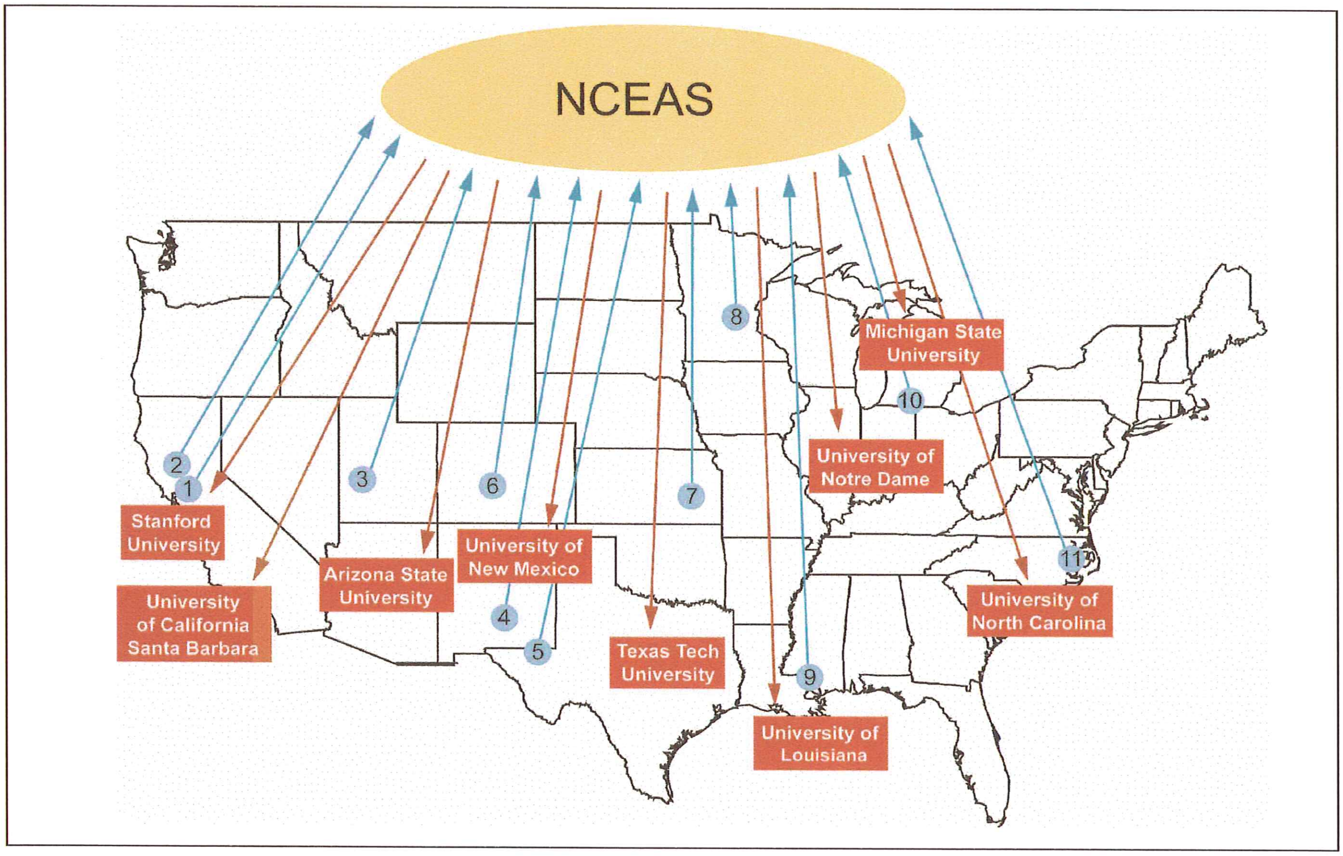

Figure 1. Map showing the geographic distribution of participating universities (red boxes) and of sites that provided data (blue circles). Data for individual sites were managed locally by students and, together with metadata, were uploaded to the National Center for Ecological Analysis and Synthesis (NCEAS). The overall project and database were managed from NCEAS. Eleven sites contributed data: (1) Hastings Natural History Reservation, (2) McLaughlin Natural Reserve, (3) Great Basin, (4) Sevilleta Long Term Ecological Research (LTER) site, (5) Jornada Basin LTER site, (6) Shortgrass Steppe LTER site, (7) Konza Prairie LTER site, (8) Cedar Creek LTER site, (9) Pearl River, (10) Kellogg Biological Station LTER site, and (11) North Carolina grasslands.

overall structure and mechanics of the distributed course, which provides an innovative model for research training by engaging students in interdisciplinary, collaborative, and synthetic research. In this approach, integrated graduate seminars are conducted simultaneously at several universities throughout the country. A project leader (typically a faculty member) coordinates the overall seminar, with assistance from a graduate or postdoctoral student data manager. A comprehensive syllabus, with associated PowerPoint presentations, provides a common conceptual framework for the course, but faculty leaders have flexibility to adapt, add, or delete materials depending on their interests and needs. Individuals at different universities communicate through the Web, e-mail, and conference calls (figure 2).

The overarching theme of our course was the importance of spatial and temporal scale in understanding the relationship between aspects of biodiversity and ecosystem function. Consequently, we structured the course to engage students in multiscale analyses of these relationships. Each local seminar focused on the relationship between biodiversity and ecosystem function in a form that was specific to a particular geographic area. Collectively, the entire group focused on larger-scale, cross-site themes.

Course content and research activities were organized into three interconnected and overlapping phases. The first phase concentrated on understanding the relationship between biodiversity and ecosystem function within a single site. This phase provided students with direct experience in data and metadata management, including the fundamentals of structured metadata, software tools for creating structured metadata, and social issues related to data sharing. The second phase focused on cross-site synthesis and collaboration, including the data integration process, associated problems of reconciling scale, and quantitative methods for analysis of large data sets. The third phase involved additional cross-site synthesis, development of follow-up activities, and dialogue between students and software developers. The course syllabus is available on the KNB project Web page (*http://knb.ecoinformatics.org/education/*).

To date, 14 faculty members, 83 graduate and 2 undergraduate students, and 6 postdoctoral scientists have participated in the KNB distributed course. During the first year

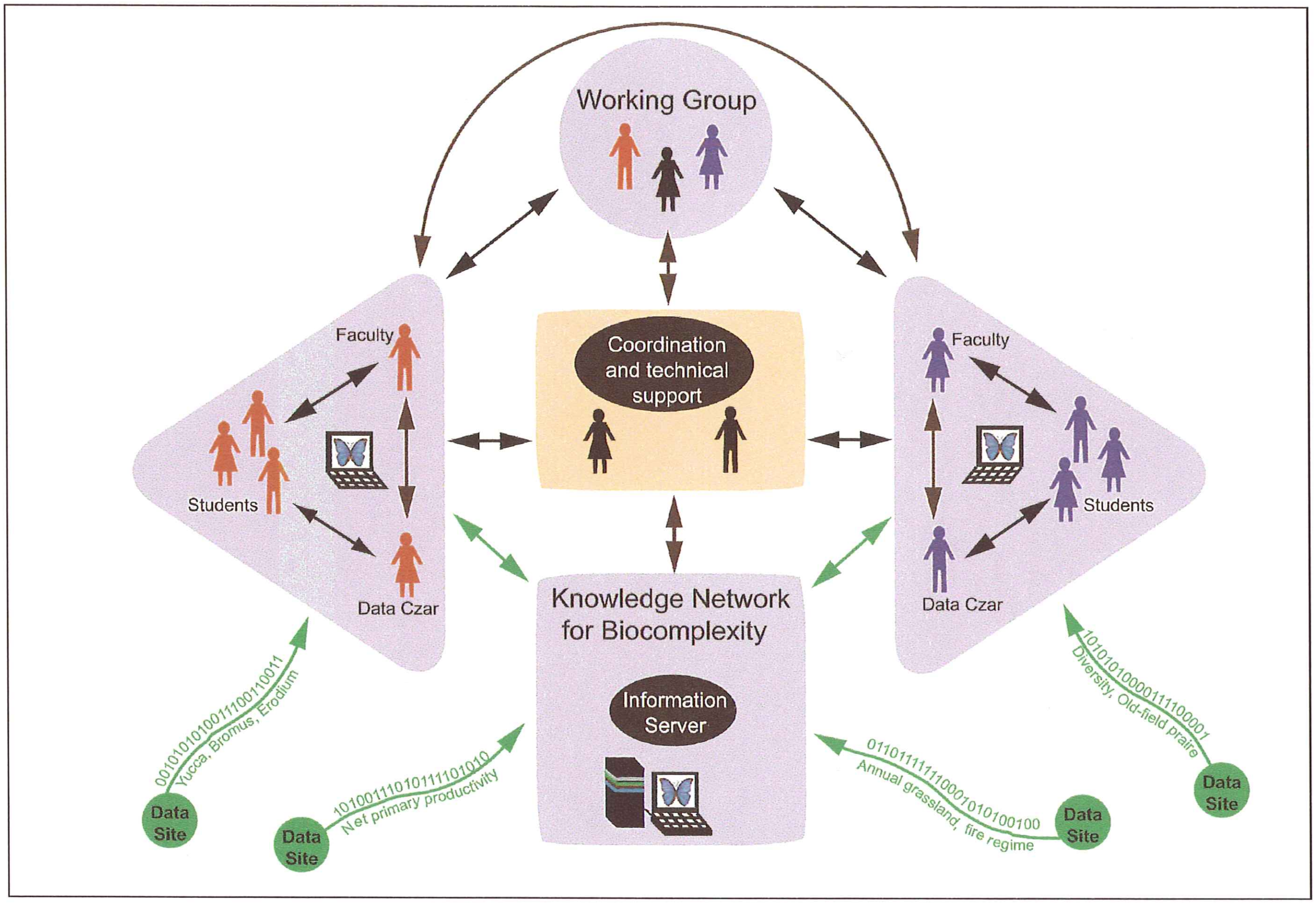

Figure 2. The organizational structure for the distributed course.

(2001), we offered seminars at three universities (Texas Tech University, University of California–Santa Barbara, and University of New Mexico), involving 4 faculty, 17 graduate students, and 2 undergraduates. In 2002, 83 participants, including 66 graduate students, 6 postdoctoral researchers, and 10 faculty from six universities (Arizona State University, Michigan State University, University of Notre Dame, Stanford University, University of Louisiana–Lafayette, and University of North Carolina–Chapel Hill), participated in the course (figure 1). To a large extent, the course was student directed, with faculty serving as consultants or facilitators. The postdocs also served as consultants and helped with overall coordination, data integration, and analysis.

Data-sharing philosophy and data-management responsibilities. The course and associated research activities were predicated on a philosophy that individual scientists and institutions have a responsibility to make the data from individual studies available to the scientific community. To promote data sharing, ensure proper attribution, and minimize the likelihood of misinterpretation or misuse of data, we negotiated data-use agreements for each data set (box 1). In some cases, these agreements were developed quickly, while in others, negotiations required weeks. The details of data-use agreements differed among investigators and among sites. Specific data-use agreements for each data set were incorporated into the metadata, and at the outset all course participants agreed to abide by the terms of these agreements. As in any large-scale collaboration, the diversity of data-use requirements complicated the research process. For example, as research questions evolved and were refined, in some cases

Box 1. Variation encountered in data-use policies required by individual investigators and sites

- Use of the data is unrestricted.
- The data owner and funding source must be acknowledged in publications that use the data.
- Use is restricted to projects associated with the distributed course.
- Data may be used only after the research topic is approved by the data owner.
- Any papers resulting from use of the data must be reviewed by the data owner before submission.
- The data owner must be an author on any publications that use the data.

it was necessary to contact the data owners and renegotiate use agreements. In other cases, several levels of manuscript review were required before publication to ensure compliance with these agreements.

While focusing on data from local sites analyzed within local seminars, participants were aware that data and metadata from their analyses would form the backbone of cross-site, multiscale syntheses in the second phase of collaboration. To facilitate this integration process, one graduate student from each institution served as local "data czar." This ensured intimate, first-hand knowledge of the data and metadata, a critical prerequisite to effective synthesis. To facilitate cross-site data integration and to ensure consistency in data management and analysis, an additional graduate student, based at NCEAS with the project leader, served as the overall information manager, with responsibility for maintaining overall data quality (including control of standards), coordinating data integration, and implementing overall data management policies. This role was critical to the overall course success. We provided the students with course materials related to the foundations of structured metadata and with access to the software tools, primarily through Web-based materials. However, the data czars from each university also came to NCEAS during the first week of the seminars to learn the basics of managing metadata and using the software tools. These data czars served as local data managers and as local technical consultants for students and faculty. They also returned to NCEAS midway through the course to refine research questions and to resolve methodological problems related to multisite data integration.

Knowledge Network for Biocomplexity software tools. KNB software development has focused primarily on tools for data access and for information and knowledge management. The core of KNB is Ecological Metadata Language (EML) (Michener 2000; *http://knb.ecoinformatics.org/software/eml/*), a method for formalizing and standardizing the set of concepts that are essential for describing ecological data (e.g., variable names and definitions, units of measurement, time and location of data collection, investigator identity). Spearheaded by NCEAS and the LTER Network, EML developed from an open-source, nationwide, community-based effort involving ecologists, information managers, and software developers. Morpho is a data and metadata management tool, being developed as part of KNB, that can work on any computer platform. With Morpho, a user can create and edit tables for data and metadata, manage data and metadata on a local computer, and search and query ecological data, either locally or through the Internet. All EML terms (currently there are about 2000) for documenting ecological data are available through Morpho (*http://knb.ecoinformatics.org/software/morpho/*). A third element in KNB is the Metacat Server, a flexible database based on XML (Extensible Markup Language) that stores a wide variety of metadata documents and makes them accessible through the Internet. Students used and evaluated these tools as part of the course. In particular, because of its foundational importance to the system, they were encouraged to use EML as much as possible to clearly document the data and to facilitate data integration.

Integration of multiscale, multisite data. In addition to the site-specific research, students from different universities designed a collaborative research project with an expanded geographic extent, informed by integration and analysis of data from multiple sites. The data integration process was the most time-intensive part of the exercise. Although some of the data integration and preliminary analysis was done remotely, using e-mail and the Internet, face-to-face contact was essential for producing meaningful results. Throughout the course and in the working group meetings, students documented the process of data integration and analysis by using scripted languages such as MATLAB or SAS and by creating structured metadata. Scripted languages provide an explicit and documented sequence of instructions for data assimilation analysis, which are carried out on the computer, so the script serves simultaneously as a series of software commands and as metadata documenting the integration and analytical process.

The grand synthesis working group. At the end of the course, a larger group of students and faculty participated in a working group at NCEAS to conduct comprehensive analyses and syntheses, to interact with the software developers, and to provide feedback on their experiences in the course. During an initial plenary meeting, each data czar provided an overview of local data and metadata and a synopsis of the research within his or her site. This was followed by further discussions in plenary sessions and subgroups. Research questions for cross-site analyses initially were identified through e-mail discussions and conference calls. However, these questions underwent considerable refinement during face-to-face discussions. Most working group activities took place in subgroups of two to five students, with each subgroup addressing components of cross-site analyses. Most students chose to work on a single project or subproject, chosen on the basis of their individual interests and analytical expertise, but some students worked on several projects. Each subgroup identified or elected a leader who coordinated group communication and activities. In most cases, this individual also agreed to serve as the lead author on a manuscript resulting from the subgroup's analyses.

During a final plenary session, each subgroup leader summarized the methods and results of analyses pertaining to that subgroup's particular cross-site research question and solicited input and critique from other participants. In addition, each subgroup submitted written text to (a) document the methods, analytical approaches, and results of cross-site research and (b) define a trajectory of future activities, including time lines, hallmarks, and individual responsibilities for particular tasks.

Lessons learned

Throughout the process, participants provided feedback on the course content and the research experience. They also provided feedback to the developers on the design and usability of software tools, which in turn evolved in response to user input.

Perhaps the most valuable aspect of the distributed course was the opportunity afforded to graduate students and postdocs to develop collaborative and synthetic research projects. Before taking the course, 89 percent of the student participants had no previous experience with the organizational and social challenges engendered by multiscale, interdisciplinary, synthetic research. Although 93 percent of the students said the collaborative interactions were intellectually stimulating, they also learned that this stimulation came at the cost of greater investment in communication, coordination, negotiation, and documentation, which requires time and can be frustrating.

Students initially underestimated the amount of effort required to obtain data and adequate metadata from other researchers. Even when investigators were willing in principle to provide access to data, these data exhibited a variety of formats, reflected different spatial and temporal sampling designs, and included different suites of environmental data (box 2). In addition, the original metadata provided by sites were highly variable in extent, depth, quality, and associated types of uncertainty (box 2; Regan et al. 2002). Consequently, to interpret the data, students had to communicate directly with data owners by e-mail or phone. Uncertainty in the metadata was manifest in several ways. In some cases, descriptions were incomplete (e.g., only abbreviations for species were provided but not species names), or the language used to describe data or data collection procedures was vague, ambiguous, or context dependent (box 3).

Box 2. Types of heterogeneity encountered during the course and illustrative examples

Data heterogeneity

- File type: ASCII, Excel, Access
- Graphic image: map, photograph, sonogram
- Videotape
- Audio recording
- Museum specimen record

Measurement heterogeneity

- Temporal scale: days, months, years, decades
- Spatial scale: 0.25 square meter (m^2), 1 m^2, 100 m^2, 10^{10} m^2
- Choice of metrics: number of individuals, biomass per species in grams dry weight, total biomass for all species in grams dry weight
- Levels of precision: number of significant digits or rounding error
- Measurement error: variable diameter at breast height
- Sampling biases (related to methodological variation): temporal frequency of sampling, spatial distribution of plots
- Linguistic uncertainty: site, plot, study area, transect
- Taxonomic convention: different names for the same underlying species at different sites because of synonymies among different taxonomic conventions

Other types of heterogeneity

- Storage media: paper, diskette, CD
- Operating system differences: Windows, Mac OS, Linux, Unix
- Software differences: Excel, Access, filePro, SAS, MATLAB
- Format and organizational differences: transposing rows to columns

When creating metadata, considerable effort is needed to reduce linguistic uncertainty, which is widespread in ecology (Regan et al. 2002). What quantity of metadata is sufficient to enable data that were initially collected for a single, relatively narrow purpose to be understood and used appropriately for a variety of other purposes? This question does not have a simple answer, and researchers will need to analyze and use ecological data sets many times, and for diverse applications, before generalizations emerge. In this study, one data set ultimately could not be used in the cross-site integration because the metadata were insufficient to understand the method used to estimate biomass. In addition, variability in data-use restrictions added complexity (e.g., additional layers of review, time lags, additional authors) to the process of producing publishable manuscripts, beyond what students might encounter when attempting to publish articles resulting from analysis of their own data.

From the outset, the course was intended as a mechanism for involving a cross section of the ecological community in the design and evolution of KNB software tools, particularly tools for creating structured metadata. Students were extraordinarily generous with their time in testing the tools and providing feedback to developers. Nonetheless, because the tools were in the early stages of development, they slowed rather than accelerated the process of creating and managing metadata.

Although the course was designed to be executed over a quarter or a semester (i.e., 10 to 15 weeks), we found that the development of productive collaborations, particularly in the context of long-distance interactions, presented numerous challenges that required a year or more before such research attained fruition. Communicating the necessity of this time commitment to students early in the course will minimize frustration.

Quantitative skills facilitate data synthesis, both in the context of this course and in the context of other synthetic activities (e.g., NCEAS working groups). Students in the KNB course were generally unfamiliar with multivariate statistics and with the range of models for regression and analysis of covariance. Ninety-three percent of students did not have skills in the scripted programming languages (e.g., SAS or MATLAB) that are needed for the integration of large data sets.

Box 3. Linguistic uncertainty in ecological metadata

Linguistic uncertainty is the uncertainty produced by statements in natural language (Walley and de Cooman 2001, Regan et al. 2002). In this example, the word *site* was used in different ways by different investigators. For example, in one study, *site* might refer to the overall study site, whereas in another, it might refer to a specific sampling location within the overall area of investigation. Levels of spatial scale ranged across at least 10 orders of magnitude, and there was no conventional terminology for describing spatial hierarchies in sampling designs. To address this linguistic uncertainty, students created a standardized vocabulary to describe levels of spatial scale, presented below.

Definition of site	Spatial scale
Reserve or study site	10^8–10^{10} square meters
Community	10^5 square meters
Field	100 square meters
Plot	1 square meter

As a result of their experience in the course, and motivated by the need to manipulate large amounts of heterogeneous data, one group of graduate students participated in a SAS tutorial, because the statistical tools with which they were familiar were inadequate for the kinds of analyses and the level of integration needed for using multisite data sets.

Technological advances such as password-protected, collaborative Web sites; e-mail; conference calls; and software tools for creating structured metadata and remote data access facilitate synthesis of data from multiple sites and collaboration among individuals at distant institutions. However, our experience suggests that these approaches are particularly valuable as an adjunct to, rather than as a surrogate for, face-to-face collaborations. The importance of face-to-face interactions must not be underestimated. In the first year, particularly early in the course, students generally focused on interactions with peers and faculty at their own institutions, with infrequent and sporadic e-mail discussion and collaboration in the development of research questions and approaches by students from different institutions. In the second year, as a result of adding a short, intensive workshop involving the data czar from each university, students did much more interactive work with other universities.

Future directions

The distributed course model we used was initially designed by Peter Kareiva for an NCEAS working group that examined the role of science in developing habitat conservation plans under the Endangered Species Act (8 universities, 10 faculty, 108 graduate students; Savage 1998, Kareiva et al. 1999). Later, it was used in a working group (19 universities, 33 faculty, 263 graduate students) led by Peter Kareiva and Dee Boersma to evaluate the use of science in developing recovery plans for endangered species (Boersma et al. 2001). Thus, more than 450 students from 36 universities have been involved in this educational model. Clearly, the model is flexible and adaptable to a wide range of topics.

Internationalization. To expand the geographic scope of the distributed seminar model and to diversify the cross section of the ecological community participating in biocomplexity research, education, and software development, the next cohort of the distributed course will involve collaboration among graduate students, faculty, and agency scientists from the United States and southern Africa. A similar model is being considered for training ecologists in data and metadata management in conjunction with a global network of field stations for monitoring biodiversity, implemented by the TEAM (Tropical Ecosystem Assessment and Monitoring) program of Conservation International. The model also might be useful for the National Science Foundation's proposed National Ecological Observatory Network initiative. Ultimately, the information needed to address broadscale questions that are essential to solving a host of pressing environmental problems facing society in the 21st century will need to be derived from databases with a global domain.

Evolving culture. Early in their careers, scientists need to appreciate the value of data and metadata, not only as they pertain to the specific questions for which they were gathered, but also as a long-term community resource. The culture of data ownership must evolve from a private to a collective paradigm. Funding agencies and foundations that support environmental research should provide incentives and financial support for adoption of these standards.

Analytical tools. Beyond the ability to discover and retrieve data, the greatest limitation to synthesis that the students faced was related to data concatenation, manipulation, and analysis. The increasing need for synthesis and analysis of large, heterogeneous data sets requires future analysts to be skilled in the tools and principles of relational database management, including data manipulation and integration.

The distributed course model has many benefits as a pedagogical and intellectual instrument. Its basic framework can easily be modified to accommodate exploration of a variety of research topics within a range of settings (e.g., academia, natural resource agencies, nongovernmental organizations). As is the case with any large collaborative enterprise, organization, coordination, and communication involve substantial investments of time from both faculty and students, well beyond what would typically be involved in a traditional graduate seminar.

Acknowledgments

The distributed course and the research for this article were supported by the National Science Foundation (NSF), grant DEB 99-80154; by the Andrew W. Mellon Foundation; and by the National Center for Ecological Analysis and Synthesis, a center funded by NSF grant DEB 00-72909, the University of California system, and the University of California–Santa Barbara campus. We thank the student, postdoctoral, and faculty participants who made the course a success; the investigators who generously provided data; and the software developers, particularly Daniel Higgins, Matthew Jones, and Mark Schildhauer, for assistance with tools. Eric Fegraus assisted with coordination and data management, and Richard Stevens provided useful comments on the manuscript.

References cited

Boersma PD, Kareiva P, Fagan WF, Clark JS, Hoekstra J. 2001. How good are endangered species recovery plans? BioScience 51: 643–649.

Brown JH, Roughgarden J. 1990. Ecology for a changing earth. Bulletin of the Ecological Society of America 71: 173–188.

Kareiva PM. 2002. Applying ecological science to recovery planning. Ecological Applications 12: 629.

Kareiva P, Anderson M. 1988. Spatial aspects of species interactions: The wedding of models and experiments. Pages 35–50 in Hastings A, ed. Community Ecology. New York: Springer-Verlag.

Kareiva PM, et al. 1999. Using Science in Habitat Conservation Plans. Santa Barbara (CA): National Center for Ecological Analysis and Synthesis; Washington (DC): American Institute of Biological Sciences.

Michener WK. 2000. Transforming data into information and knowledge. Pages 142–161 in Michener WK, Brunt JW, eds. Ecological Data: Design, Management and Processing. Malden (MA): Blackwell Science.

Michener WK, Baerwald TJ, Firth P, Palmer MA, Rosenberg JL, Sandlin EA, Zimmerman H. 2001. Defining and unraveling biocomplexity. BioScience 51: 1018–1023.

Nature. 2003. Who'd want to work in a team? Editorial. Nature 424: 1.

Regan HM, Colyvan M, Burgman MA. 2002. A taxonomy and treatment of uncertainty for ecology and conservation biology. Ecological Applications 12: 618–628.

Savage LT. 1998. Innovative national graduate student seminar analyzes habitat conservation plans. Integrative Biology 1: 45–48.

Walley P, de Cooman G. 2001. A behavioural model for linguistic uncertainty. Information Sciences 134: 1–37.