The Resource Discovery Initiative for Field Stations: Enhancing Data Management at North American Biological Field Stations

JAMES W. BRUNT AND WILLIAM K. MICHENER

Biological field stations in North America have significant potential for addressing the most pressing environmental challenges facing science and society. Many of these field stations are now actively engaged in research networks and developing environmental observatory networks. The Resource Discovery Initiative for Field Stations (RDIFS) represents a research coordination network developed to enhance data management capacity and better position field stations for the critical role they are to play in addressing environmental challenges. The RDIFS developed information resources and training programs to facilitate storage, discovery, and access to data and information that are collectively held at North American biological field stations. In this article, we highlight the capabilities and needs of biological field stations, identify specific data management challenges faced by field stations, describe the products of the RDIFS effort, and provide insight into the future of data management at field stations, especially in relation to participation in environmental observatory networks.

Keywords: data management, field stations, research coordination network, Organization of Biological Field Stations, databases

A central challenge in the 21st century is to improve our understanding of the natural world so that biodiversity, natural resources, and quality of life can be sustained. Such a science challenge is associated with an array of informatics challenges. For instance, understanding patterns of biodiversity and their underlying ecological and evolutionary mechanisms requires vast quantities of extraordinarily diverse, complex information from many scientific and sociological disciplines. Scientists continue to amass data and information about the natural world, but they often fail to adequately document the data (i.e., the metadata are incomplete), which would enable their reuse or use by others; promote discovery and acquisition of their data; and support preservation of data and information beyond the duration of their study.

The Ecological Society of America (ESA) and the National Research Council (1991), as well as the ecological informatics literature (e.g., Michener et al. 1997, Michener and Brunt 2000, Cook et al. 2001), have emphasized the importance of increasing access to biological and environmental data. The report of the ESA Committee on the Future of Long-term Ecological Data (FLED), for example, documented the importance of long-term data sets; the causes for their loss; and the critical need to develop mechanisms to promote their preservation, maintenance, and use (ESA 1996). In response to the FLED report, and with concomitant improvements in electronic communication and data storage, the ESA established Ecological Archives to publish peer-reviewed data papers and digital appendices (see *http://esapubs.org/archive/*). Ecological Archives provides an important outlet for storing and documenting peer-reviewed data sets that are deemed extremely valuable by the scientific community.

In addition, workshops funded by the National Science Foundation (NSF) and hundreds of workshops conducted at the National Center for Ecological Analysis and Synthesis (NCEAS) have highlighted the current challenges associated with understanding the complexity of biological systems. For instance, representatives of a broad spectrum of subdisciplines participated in a Frontiers in Ecology workshop held at the NSF in December 1999, at which they identified several critical informatics-related hurdles that hinder progress in dealing with the major environmental issues that society faces. In particular, they emphasized the need for "ecologists

BioScience 59: 482–487. ISSN 0006-3568, electronic ISSN 1525-3244. doi:10.1025/bio.2009.59.6.6

trained to manage large databases, who can organize the storehouse of past ecological data, mine it for new results, and make it accessible to others" (Thompson et al. 2001).

Despite greater attention to ecological informatics and proactive measures such as Ecological Archives, significant 21st-century challenges persist in discovering, accessing, and acquiring environmental information. For example, thousands of environmental data sets routinely collected at the field stations that make up the Organization of Biological Field Stations (OBFS) have not been systematically archived in electronic media, nor have they been readily discoverable or accessible for analysis and synthesis. Thus, these data accumulated over the past century were unavailable for solving critical environmental problems and advancing basic biological science (Stanford and McKee 1999). Making this wealth of ecological data useful again provided the focal point for the collaborators on the Resource Discovery Initiative for Field Stations (RDIFS).

Field station capabilities and needs

Field station data are derived from all scales of biological organization, are highly heterogeneous in content and format, and span enormous scales of space and time. For instance, an unpublished 1999 OBFS survey of member field stations (summarized in Swain et al. 1999) found that 60% had active research programs on endangered species, 43% on habitat loss and fragmentation, 38% on fire processes, and 60% on exotic species. To better understand the nature and pace of environmental change, 62% of OBFS member stations conduct research on water quality, 24% on air quality, and 33% on global change. Many field stations have established valuable partnerships with agencies and nongovernmental organizations. The same survey found that, at the federal level, 31% work with the US Department of Agriculture Forest Service, 14% with the US Fish and Wildlife Service, and 17% with the National Park Service. At the state level, 38% work with state fish and game agencies and 41% with state environmental protection agencies. In addition, 38% of the field stations work with their local government. Most field stations collaborate with local environmental groups (71%) and citizen organizations (60%). More than 70% of surveyed field stations offer K–12 programs, thereby providing an important link between research and K–12 education.

Scientists from the Long Term Ecological Research Network (LTER), the OBFS, the NCEAS, and the San Diego Supercomputer Center envisioned RDIFS. The research coordination network was designed to address the need for better coordination of approaches to developing data and information management capacity among field stations, as documented in a 1998 NCEAS workshop (Stanford and McKee 1999).

In planning for RDIFS, 160 OBFS member field stations were surveyed by mail in April 2001 about their information management capabilities and needs; 80 field stations (50%) responded to the survey. Respondents reflected the breadth of North American field stations, which are found from Alaska to Costa Rica; cover freshwater, coastal, and terrestrial systems; are small, medium, and large; and have missions ranging from research to education, although most stations have a mixture of both.

Four primary issues were identified as impeding the storage, discovery, and access of field station data: (1) insufficient network connectivity, (2) obsolete equipment and software, (3) inadequate data management and systems administration support, and (4) lack of training. Furthermore, field data and metadata, site bibliographies, records of ongoing site research projects, and other valuable resource information frequently were not maintained at all, or were inconsistently maintained on paper or in myriad text or word processing files. Importantly, 70 of the 80 stations reported that they had a critical need for informatics training in numerous areas, including database management systems and implementation approaches, geographic information systems (GIS), metadata management tools and implementation, data quality assurance and quality control (QA/QC) methods, networking and site computing environments (including wireless), and programming database access from the Web.

RDIFS activities

Following the survey of field station needs, two principal RDIFS activities were designed to promote the storage, discovery, and access of data and information resources at North American biological field stations.

First, informatics research arising from intensive, product-oriented workshops supported the design and development of five databases:

- A North American field station data registry and repository, to promote discovery and access of data and metadata
- A controlled vocabulary or thesaurus, to provide more consistent mechanisms for documenting and discovering field station data
- A site characteristics database, to provide standardized descriptions of field stations and their associated habitats and ecosystems
- A bibliography of North American field station publications, to provide access to research results and to document the scientific productivity at field stations
- A reference database for standard methods, including data QA/QC, to promote best practices and facilitate standardization of methodologies across studies

Second, a series of training workshops between 2002 and 2006 exposed participants to fundamentals of ecological informatics and GIS, as well as to more specialized topics such as biodiversity databases and environmental sensor networking, which varied from year to year depending on the

participants' needs. In addition, a portion of each workshop was devoted to populating the data registry and bibliography databases. We discuss the RDIFS databases and training workshops below.

The North American field station data registry and repository. The North American field station data registry and repository was developed to meet the data discovery and access needs of field biologists and other environmental scientists working at field stations. The data registry is based on and integrated with a more comprehensive data and metadata management system developed by the Knowledge Network for Biocomplexity (KNB) project (Jones et al. 2001); it includes a subset of metadata descriptors originally defined by Michener and colleagues (1997). The data registry descriptor fields include title; contact information; abstract and keywords; temporal, spatial, and taxonomic coverage of the data set; a brief description of the data collection methods; and distribution information. These descriptors were chosen because they are easy to understand and can be easily filled in by scientists and students on a simple, Web-based form with multiple pull-down menus (figure 1). Moreover, the descriptors capture the salient information needed to identify data that satisfy thematic, temporal, spatial, and taxonomic search criteria. The data registry is housed and maintained by the NCEAS at the University of California in Santa Barbara and replicated at the LTER Network Office at the University of New Mexico.

Individuals from the field station community who attended RDIFS training sessions were strongly encouraged to bring supporting materials that would enable 10 or more data sets from their station to be included in the data registry. More than 11,000 data sets are now associated with the combined OBFS and KNB data registries, and approximately 5200 data sets are registered for North American field stations and natural reserves. Recently, the LTER Network Office linked the OBFS and KNB data registries with the National Biological Information Infrastructure (NBII) data clearinghouse and the Oak Ridge National Laboratory Distributed Active Archive Center, providing an even more effective mechanism for promoting data discovery by scientists throughout the world.

Thesaurus for field biology. Using common terminology facilitates data discovery and effective communication among scientists. In particular, data description and discovery are most efficient when scientific concepts are associated with unique terms. In lieu of direct one-on-one communication between the data producer and the potential data user (i.e., searcher), a thesaurus, or controlled vocabulary, can provide a constrained list of terms for optimal use in indexing and searching the information in a database (Batty 1998). For instance, the Global Change Master Directory (GCMD) includes a hierarchically based thesaurus from which keywords are extracted and used to describe the directory's data sets (Olsen et al. 2007; see *http://gcmd.gsfc.nasa.gov/valids/*). At the inception of the RDIFS, existing thesauri like the GCMD, which is most relevant for the earth sciences, were missing many of the keywords that biologists typically use to describe their data.

Since 2002, three events in particular have facilitated data discovery. First, one of the RDIFS-affiliated scientists was added to the GCMD advisory team, which led to improvements in the GCMD thesaurus and to its adoption for use in the OBFS data registry. Second, a comprehensive biological thesaurus developed for the NBII has been broadly adopted. Third, RDIFS investigators engaged with a broader community of informatics experts from the LTER community and the Science Environment for Ecological Knowledge project (Michener et al. 2007) to develop a more effective controlled vocabulary for field-oriented biological studies. Analyses were performed using data registry keywords, metadata text, and bibliographic entries. For example, in analyses of metadata documents in the LTER Data Catalog, more than half (1616 of 3206) of the key terms were used only once, and only 104 of the terms were used at 5 or more of 26 LTER sites (Porter 2006). The situation is similar for other lists of words (table 1). Scientists affiliated with the RDIFS continue to

Figure 1. The Organization of Biological Field Stations data registry.

Table 1. Evaluations of selected keywords and terms from the Long Term Ecological Research (LTER) and Organization of Biological Field Stations communities.

Source	Total number of terms	Number of terms used at five or more sites	Ten most frequent concise scientific terms (number of uses)
Ecological Metadata Language (EML) keywords from data registry	3206	104	Temperature (701), pH (406), light (300), phosphorus (299), conductivity (299), nitrate (287), nutrients (252), alkalinity (220), salinity (219), nitrite (171)
EML attributes from data registry	6318	436	Water (1568), species (1218), temperature (1106), plant (685), soil (614), cover (543), air (479), biomass (416), salinity (361), carbon (328)
Bibliography titles from LTER scientific bibliography database	13,538	1855	Forest (2050), soil (1362), nitrogen (1002), ecosystem (947), lake (819), prairie (780), water (777), carbon (757), plant (700), species (673)

Note: Terms are single words only. Sites include participating field stations, marine labs, and LTER sites. Terms include all scientific and other words found in the source. Scientific terms were distinguished qualitatively as concise by the authors on the basis of the potential for ambiguous interpretation.
Source: Porter (2006).

collaborate with LTER information managers to improve the searchability of ecological and environmental data. Such efforts will contribute to the ultimate development of domain-specific ontologies that represent a flexible and powerful mechanism to capture the structure, content, semantic subtleties, and relationships among data variables (Madin et al. 2007).

Site characteristics database for North American field stations. Greater interest in broadscale ecosystem studies and similar types of comparative and synthetic research has created the need for other resource discovery tools. In particular, scientists need to be able to locate field stations, natural reserves, and research sites that meet specific criteria, such as the habitats a site encompasses, climatic characteristics, and infrastructure capacity. We implemented and populated a site characteristics database for North American field stations that satisfies this objective (*www.obfs.org/stations/*). The site characteristics database is an extension of a similar database associated with the LTER Network Information System (Baker et al. 2000). The site characteristics database contains modules that include descriptors for contact information; location; climate; hydrology; facilities; research and education programs; data management support; and classifications for soils, geology, and disturbance.

Bibliography of North American field station publications. Another effective way to promote resource discovery at North American field stations is to have a centralized bibliography. Until the RDIFS, no such centralized resource existed for the field station community, and with relatively few exceptions, it was very costly and time-consuming to identify publications produced at specific field stations.

We have developed a bibliography of North American field station publications, which can be continually updated. The bibliography was initially implemented using off-the-shelf bibliographic software (i.e., EndNote) and an unsupported software package that allowed EndNote libraries to be posted on the Internet with a user-friendly, searchable interface. In 2004, the North American field station publications were included in the development and expansion of an LTER All-Site Bibliography (Brunt 2005) using open-source approaches and conforming to bibliographic standards.

The existence and growth of a bibliography of North American field station publications achieves several important objectives. First, biological field stations have a reliable mechanism for tracking and maintaining records of publications resulting from research by students and scientists. Second, the bibliography serves as a resource discovery tool, enabling students and scientists to more readily identify research that has been completed and published at a station. Students can better plan graduate research projects, and scientists are better prepared to integrate information within and across sites and to ask similar scientific questions in a wide variety of habitats (see *http://search.lternet.edu/biblio/*).

Database of standard methods. Data quality is enhanced, and analysis and interpretation of field biology studies are often facilitated, when standard methods are employed and when significant attention is devoted to QA/QC. For instance, very different results are often obtained when different field and laboratory methods are used to measure identical phenomena. Unless a particular method is intercalibrated with other methods, it may be impossible to relate data from one study to that from others. Consequently, many studies of a particular type rely on a limited number of methods, sampling gear (e.g., types and mesh sizes of plankton nets), and instrumentation. When standard methods exist and have been well documented in the literature, their use can benefit field biologists in numerous ways. For example, comparison with other studies can be facilitated, documentation of the methods for publications and metadata is easier, and costs are often lower.

The database we developed of standard methods for field and laboratory studies and QA/QC methods contains more than 1300 citations of methods volumes and papers (*http://search.lternet.edu/methods/*). The database, which is a subcomponent of the publication database, is accessible through the OBFS Web site. It includes annotated references to QA/QC

methods that are relevant to the data collected at field stations, as well as pointers to standard field and laboratory methods. The database can be easily maintained, revised, and supplemented through Web-based interfaces.

Training workshops. By 2006, more than 140 field station personnel from more than a third (> 60) of the OBFS-affiliated field stations representing the United States, Canada, the Bahamas, Costa Rica, Puerto Rico, and French Polynesia had participated in the training sessions. The RDIFS supported annual two-week training sessions in ecological informatics and GIS. The sessions were scheduled during late fall through early spring, a flexible period for most biological field station personnel. In addition, two shorter workshops focused specifically on designing and implementing environmental sensing networks in the field were held midway through the RDIFS project.

Figure 2. John Porter, of the University of Virginia, instructs participants in the Resource Discovery Initiative for Field Stations in the ecoinformatics training laboratory at the University of New Mexico. Photograph: McOwiti O. Thomas.

Intensive hands-on training in ecological informatics targeted a modular series of topics that were varied annually. Training modules included database management systems; Web site design and page-authoring fundamentals; advanced Web programming; metadata management; and hardware, software, and communications. The GIS course focused on managing and analyzing spatial data using commercial software, as well as on collecting and processing data collected from global positioning systems. Training activities were held at the Sevilleta Biological Field Station in New Mexico, the La Selva Biological Station in Costa Rica, and in the state-of-the-art informatics training lab at the University of New Mexico. A steering committee chose the mix of informatics modules that were taught in any given year. Students were surveyed before the workshops to determine their levels of expertise and their objectives for the training, and again after the workshops to assess how well the workshops had met their expectations. The posttraining surveys indicated that 98% of the participants felt that the training met their expectations. Similar percentages of respondents said that the skills they learned would be useful in their jobs, and that they would recommend the training to others. Both evaluation activities proved useful in customizing the training workshops to meet student needs.

Conclusions

Biological field stations have long been recognized for their role in integrating long-term research, education, and outreach (Eisner 1982, Wilson 1982, ESA 1996). Their potential value as components of integrated national research and monitoring networks has also been highlighted (Lohr and Stanford 1995). Today, thanks in part to RDIFS activities, field stations form the backbone for environmental observing networks such as the National Ecological Observatory Network, whose continental-scale research agendas cannot be addressed without an effective framework for managing, storing, discovering, and communicating relevant data.

The RDIFS Research Coordination Network engaged scientists from many disciplines and institutions to develop the databases needed by North American biological field stations. Using the site characteristics database, it is now possible to identify field stations in particular locations or with particular ecosystem types, or stations that meet specific infrastructure needs, such as being able to house a class of a certain size. Thanks to the data registry, associated controlled vocabularies, and the bibliography of North American field station publications, scientists can also ascertain what data have been collected and what publications have resulted from research at a particular field station. The database of standard methods is an invaluable resource that facilitates identification of best practices for field sampling, laboratory analyses, and QA/QC.

RDIFS database development and training activities were designed to benefit biological field stations in North America and their associated scientists and students by promoting best data management practices and enhancing workforce expertise and diversity. Cyberinfrastructure and environmental sensing technologies are changing rapidly. These changes present a real challenge for biological field stations, ensuring a continuing need for funding programs that support acquisition of these new technologies and provide

workforce training in advanced technologies. We envision that such an investment in human and technological infrastructure will ultimately determine the success of emerging environmental observing networks.

Acknowledgments

This work was directly supported by a National Science Foundation research coordination grant (DBI-0129792), an LTER Network Office Cooperative Agreement (DEB-0236154), the Science Environment for Ecological Knowledge Large Information Technology Research project (DEB-0225665), and the National Center for Ecological Analysis and Synthesis (EF-0553768). This work would not have been possible without the tremendous support of the many individuals who contributed their time and effort to database development, participated in field station surveys, and taught and attended the training workshops.

References cited

Baker KS, Benson BJ, Henshaw DL, Blodgett D, Porter JH, Stafford SG. 2000. Evolution of a multisite network information system: The LTER information management paradigm. BioScience 50: 963–978.

Batty D. 1998. WWW—wealth, weariness or waste: Controlled vocabulary and thesauri in support of online information access. D-Lib Magazine (November). (23 March 2009; *www.dlib.org/dlib/november98/11batty.html*)

Brunt JW. 2005. The LTER Network All-Site Bibliography. LTER Databits (spring). (23 March 2009; *http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/05spring/#4fa*)

Cook RB, Olson RJ, Kanciruk P, Hook LA. 2001. Best practices for preparing ecological data sets to share and archive. Bulletin of the Ecological Society of America 82: 138–141.

Eisner TE. 1982. For the love of nature: Exploration and discovery at biological field stations. BioScience 32: 321–326.

[ESA] Ecological Society of America. 1996. Final Report of the Ecological Society of America Committee on the Future of Long-Term Data (FLED), vol 1: Text of the Report. ESA.

Jones MB, Berkley C, Bojilova J, Schildhauer M. 2001. A distributed XML database system for managing scientific metadata. IEEE Internet Computing 5: 59–68.

Lohr SA, Stanford J. 1995. A New Horizon for Biological Field Stations and Marine Laboratories. Elsevier. doi:10.1016/0169-5347(96)20032-1

Madin J, Bowers S, Schildhauer M, Krivov S, Pennington D, Villa F. 2007. An ontology for describing and synthesizing ecological observation data. Ecological Informatics 2: 279–296.

Michener WK, Brunt JW. 2000. Ecological Data: Design, Management and Processing. Blackwell Science.

Michener WK, Brunt JW, Helly J, Kirchner TB, Stafford SG. 1997. Non-geospatial metadata for the ecological sciences. Ecological Applications 7: 330–342.

Michener WK, Beach JH, Jones MB, Ludäscher B, Pennington DD, Pereira RS, Rajasekar A, Schildhauer M. 2007. A knowledge environment for the biodiversity and ecological sciences. Journal of Intelligent Information Systems 29: 111–126.

National Research Council. 1991. Solving the global change puzzle: A U.S. strategy for managing data and information. National Academy Press.

Olsen LM, et al. 2007. NASA/Global Change Master Directory (GCMD) Earth Science Keywords. Version 6.0.0.0.0. (25 March 2009; *http://gcmd.nasa.gov/Resources/valids/archives/keyword_list.html*)

Porter J. 2006. Improving data queries through use of a controlled vocabulary. LTER Databits (spring). (23 March 2009; *http://intranet.lternet.edu/archives/documents/Newsletters/DataBits/06spring/#4fa*)

Stanford J, McKee A. 1999. Field Station 2000 Initiative: Results of a Workshop Held May 17–22, 1998 at the National Center for Ecological Analysis and Synthesis, Santa Barbara, California. Organization of Biological Field Stations. Publication no. 2.

Swain HM, Pickert RL, Stromberg M. 1999. 1999 Archbold Survey Results. (16 April 2009; *www.obfs.org/modules.php?name=UpDownload&req=viewdownload&cid=1&meid=99*)

Thompson JN, et al. 2001. Frontiers of ecology. BioScience 51: 15–24.

Wilson EO. 1982. The importance of biological field stations. BioScience 32: 320.

James W. Brunt (e-mail: jbrunt@LTERnet.edu) is the associate director for information management, and William K. Michener (e-mail: wmichene@LTERnet.edu) is the associate director for development and outreach, at the Long-term Ecological Research Network Office, Department of Biology, University of New Mexico, Albuquerque.