Ecological Informatics: a Long-Term Ecological Research Perspective

William K. MICHENER & James W. BRUNT
LTER Network Office
University of New Mexico
Department of Biology
Albuquerque, NM, 87131-1091, United States

Kristin L. VANDERBILT
Sevilleta LTER Program
Department of Biology
University of New Mexico
Albuquerque, NM 87131-1091, United States

ABSTRACT

Scientists within the Long-Term Ecological Research (LTER) Network have provided leadership in ecological informatics since the inception of LTER in 1980. The success of LTER, where research projects span wide temporal and spatial scales, depends on the quality and longevity of the data collected. Scientists have devised data collection, data entry, data access, QA/QC and archiving strategies for ensuring that high quality data are appropriately managed to meet the needs of a broad user base for decades to come. The LTER cross-site Network Information System (NIS) is being developed to foster data sharing and collaboration among sites.

Keywords: LTER, ecological informatics, information management, metadata

1. INTRODUCTION

Ecological informatics is a *broad interdisciplinary science that incorporates conceptual and practical tools for the understanding, generation, processing, and dissemination of ecological data and information.* It encompasses activities that are central to the management of ecological data and information including: (1) the project or experimental design phase; (2) data design; (3) data acquisition and data management; (4) quality assurance and quality control; (5) metadata implementation; (6) data archival; (7) data access and dissemination; and (8) facilitation of data analysis [1].

Ecological informatics plays a prominent role in the United States' Long-Term Ecological Research (LTER) Program. The LTER program was conceived to elucidate multi-decadal population to ecosystem-scale phenomena [2]. Successfully addressing these phenomena requires multidisciplinary, broad-scale, and long-term approaches and perspectives, as well as the availability of data from many hierarchical scales of aggregation (e.g., population, community, and ecosystem).

LTER is supported by the National Science Foundation and includes 24 sites (ranging from urban watershed to tropical rainforest to marine and dry valley sites in Antarctica) plus a Network Office that coordinates intersite communication and other activities. During the early development of LTER in the 1980's, attention to management of LTER data and information was almost entirely based on initiative and perceived needs at individual sites. As the LTER Program matured, the number of LTER sites expanded, the size and complexity of site-specific databases increased, and synthesis and integration increased in importance. Consequently, there has been greater emphasis on coordinating site activities and developing a network information system.

Objectives of this paper are to examine ecological informatics in the LTER Network emphasizing site activities, personnel, policies, and the development of a network information system (NIS) that provides some coordination for the 24 sites. In addition, we outline some of the future plans for LTER ecological informatics**.**

2. CURRENT STATE OF ECOLOGICAL INFORMATICS IN THE UNITED STATES LTER PROGRAM

Site

In this section, we describe data acquisition, data entry, metadata, data archival, and data access and dissemination. Next, we discuss the types of personnel that are associated with information management, costs of informatics activities as a percentage of project budgets, and LTER data policies. Lessons learned over the past two decades of LTER Program development will be emphasized.

Data design and data management*:* Experiences from the LTER Program indicate that the experimental design through the analysis phase benefit

from close collaboration among scientists, statisticians, and information management personnel throughout the duration of the project. In designing databases and the structure of data sets, we conceptualize and implement a logical structure within and among data sets that can facilitate data acquisition, entry, storage, retrieval and manipulation. When creating a data set, it is useful to follow the six recommendations that are presented below [3]. These recommendations reflect experiences from managing data at individual LTER sites, as well as several of the large data archives at Oak Ridge National Laboratory. In following these guidelines, the data will be better organized, more usable, and persistent for longer periods.

1) Assign descriptive file names -- File names should be unique and reflect the file contents.

2) Use consistent and stable file formats -- ASCII file formats, or other generic formats, should be used rather than proprietary formats that may become obsolete in the future. Data should be consistently formatted, ensuring that the number and order of columns is the same throughout the data file. Within the ASCII file, fields may be delimited by commas, tabs, pipes (|), spaces, or semicolons, preferably in that order.

3) Define the parameters -- Use commonly accepted parameter names that describe what the parameter is, and denote the parameter name consistently throughout the data file.

4) Use consistent data organization -- Files should contain sets of similar measurements taken for one study, using the same methods and instruments. One large data file spanning sites and time is easier to manage than several smaller files defined by, for example, month or site.

5) Assign descriptive data set titles -- Data set titles should ideally describe the type of data, time period, location, and instruments used. Titles should be restricted to 80 characters, and should be similar to names of data files. The title "Net Nitrogen Mineralization in Grasslands at the Sevilleta LTER, New Mexico, 1999-2001" is, for example, preferable to the title "N Mineralization."

6) Provide documentation (metadata) -- Record why the data were collected, when they were collected and how they were collected. Describe the structure of the data file, and note any changes that have been made to the data.

Data acquisition and management: Acquisition of high-quality data depends on a number of factors. Data quality is inextricably linked to the knowledge and skill levels of the personnel that collect the data. Thus, the time invested in training field and laboratory personnel can pay handsome dividends in data quality. Instrument accuracy (proximity of measurement to "truth") and precision (variation of measurement within a sampling distribution) also influence data quality. Accuracy and precision are a function of the quality of the instrument, instrument maintenance and operation, and independent verification of results.

The way in which data are acquired also affects data quality by influencing the amount of human error introduced into measurements. Properly designed data sheets are inexpensive, easy-to-use, and provide a long-term hard copy of data, but may be less efficient in the field than other methods. Tape recorder data collection eliminates initial transcription errors, but problems including battery and tape maintenance, wind noise, and dust and rain in equipment may make this method undesirable. Field entry into hand-hand computers reduces data entry errors because data are entered only once, but problems include battery life and heat, cold, dust and rain that may damage computers.

For large data collection efforts like LTER, there are often many benefits associated with entering data into a commercial database management system (e.g., facilitation of data entry, sorting, security, etc.). Some of the more common DBMS types include: (1) file-system based DBMSs, which utilize files and directories to organize information; (2) relational DBMSs, which store data in tables that can be linked by key fields; (3) object-oriented DBMSs, which store data in objects that include methods for accessing and manipulating the data; and (4) hybrid DBMSs, which use a combination of relational and object-oriented schema [4].

Quality Assurance and Quality Control (QA/QC): QA/QC refers to strategies that are designed to prevent the introduction of errors, or data contamination, into a data set. Specifically, quality control mechanisms are applied during the data acquisition process to help identify data entry errors or malfunctioning instrumentation. Quality assurance mechanisms are applied after the data have been entered into a computer to identify potential outliers.
Application of quality control measures during data acquisition and data entry can greatly reduce data contamination. Simply double-checking data sheets as they are completed to confirm that all fields have been entered and that codes and measurements were entered correctly can greatly reduce errors. Enforcing standards for formats, codes and measurement units helps ensure that data are entered consistently. Illegal data filters in data entry programs that flag data not meeting variable constraints (e.g., a legal range of values) permit data entry personnel to correct typing errors as they occur or to document data points that may be incorrect due to measurement or instrumental error. Double keying of data by independent data entry technicians followed by

computer verification is an ideal way to prevent data contamination.

Quality assurance measures include checking for unreasonable patterns in data, performing and reviewing statistical summaries, and assessing overall data quality. There are numerous graphical methods and statistical tests for detecting unusually extreme values of a variable (i.e. "outliers") [5]. Outliers may or may not represent data contamination and an explanation should be sought for extreme values. Statistical summaries of data can be compared to summaries from previous years to determine if central tendency or variability within the data has changed markedly. Finally, data validation through review by qualified scientists also increases confidence in data quality.

Data documentation (metadata): Metadata are defined as "data about data" or, more appropriately, "higher level information that describe the content, quality, structure, and accessibility of a specific data set" [6]. Comprehensive metadata are critical for slowing "information entropy" (Figure 1) which is defined as the normal degradation in information content associated with data and metadata over time [6]. For example, specific details are generally "lost" first, followed by more general details. Accidents, as well as retirement, career change, or death of key personnel can accelerate the rate of information loss.

Figure 1. "Information entropy," the loss of information content of data and metadata over time [6]

Numerous metadata standards have been developed or are under development. For example, the Dublin Core, Global Change Master Directory, and others focus on providing a limited number of descriptors that primarily support data discovery. The International Standards Organization (ISO) is currently preparing a more comprehensive metadata standard for release to the international community. The ISO standard will be appropriate for many types of biological data, particularly those that have a large geospatial component.

Each site within the LTER Network is responsible for its own metadata management system, which has led to a high level of heterogeneity in site metadata content, format and storage. This heterogeneity, which ranges from ASCII text residing in a flat file to more comprehensive DBMSs, makes the development of software tools for cross-site data sharing extremely difficult. To enhance opportunities for data discovery and synthesis among LTER sites, the LTER Network has recently adopted the Ecological Metadata Language (EML) as its metadata standard. EML is based on earlier efforts to describe a standard for non-spatial ecological metadata [6]. EML is a modular and extensible means of documenting ecological data through a series of XML document types (http://knb.ecoinformatics.org). Each EML module describes one logical part of the comprehensive metadata that should be included with all ecological data sets. A user-friendly EML management tool, Morpho, has been developed which permits users to enter, edit, query, and retrieve EML documents (http://knb.ecoinformatics.org). EML documents generated by LTER sites may be stored in a centralized Metacat (http://knb.ecoinformatics.org), an XML database, to enhance data searching capabilities. Implementation of EML by the LTER Network will facilitate cross-site data synthesis as tools for sharing, integrating, and analyzing data are developed that can work together seamlessly since they rely on standard EML input.

Data archives: A data archive is a collection of data sets, usually electronic, stored in such a way that a variety of users can locate, acquire, understand and use the data [7]. The goal of ecological archives is to foster broader ecological objectives, such as regional and multidisciplinary analyses, through data sharing. A data archive must not only preserve the data, but also provide complete metadata to guide the use of the data, offer search mechanisms to allow archive users to readily identify data sets of interest, and provide a means of delivering the data to the user.

Several formal archives exist that house data related to a particular research theme. Examples of such archives include NASA's DAACs (Distributed Active Archive Centers) where multidisciplinary data of interest to global change researchers and policy-makers are stored. A web-based search engine is often used for browsing and viewing the data before ordering it via an electronic interface. Data may be made available on a variety of media.

LTER project databases provide some of the functionality of archives. For instance, the Sevilleta LTER stores data locally and makes data accessible via the Internet. To prevent data loss, incremental tape backups are made

daily of all Sevilleta data; that is, any changes to data sets are recorded on backup tapes each night. Monthly, the Sevilleta database is completely backed up on tape. A complete set of a month's backup tapes is stored off-site. Sets of tapes are recycled every three months. Some GIS and remote sensing data sets are stored on CD, due to their size.

Data access and dissemination: Almost all LTER data are easily accessed through the LTER web site (http://www.lternet.edu/data/). More than 2,000 data sets can be accessed from this central location. Although most data sets are managed at individual sites using a relational database management system, publicly available data are typically in comma-delimited ASCII text format. Data and metadata may be accessed independently or as a bundled package, depending on the practices at individual LTER sites. Other data, including climate and hydrological data, may be acquitted from a centralized server that harvests records from multiple sites. These data may be acquired in a variety of formats (e.g., daily values to monthly averages). In addition, value-added programs (e.g., graphical analyses) are available to facilitate interpretation.

Data are also available through individual sites. For example, data may be acquired directly from the Sevilleta's web site (http://sevilleta.unm.edu/). Anyone downloading Sevilleta's data is asked to email the Sevilleta information manager with details of how the data were used; such policies encouraging attribution of data to the site and data set authors are commonplace in the LTER Network.

Data discovery and dissemination are facilitated through an online data catalog at the LTER web site. Many LTER data sets are also cataloged in the Global Change Master Directory and NASA DAACs. Furthermore, most sites have begun to invest heavily in developing data discovery, dynamic querying, and analysis tool s that facilitate research.

People and the Cost of Information Management

Most LTER sites employ a full-time information manager who is responsible for the design and implementation of the information management system. This individual frequently has interest and training in both the ecological and computer sciences (e.g., programming, DBMS). Other personnel requirements vary considerably from one site to another and partially depend upon the degree of institutional support available for system administration and other ancillary activities. Consequently, other personnel may include a system administrator, programmer, data entry specialist, and GIS specialists.

The cost of managing data at individual LTER sites is estimated to range from 10 to 20% of the annual budget, although this figure may actually range up to 40% or more if all technician time devoted to data management and GIS activities are included. The lower figure of 10% may best represent newer sites, where there are less data to manage, or to established sites, where considerable data management support may also come from other sources. Budget items primarily include salaries, communication costs (e.g., internet service providers), supplies, and minor equipment. Major pieces of equipment (e.g., servers, RAID units) are generally not included in annual budgets, but are often acquired through specific equipment grants and other sources.

Data Policies

LTER sites are dedicated to having all long-term data sets and key short-term data on-line and available to the scientific community and general public in a timely fashion (see http://www.lternet.edu/data/netpolicy.html). It is also generally recognized that researchers have an obligation to publish LTER data and that LTER investigators must have a reasonable opportunity for first use of data they collected. The LTER data access policy attempts to address these competing demands.

In return for providing access to LTER data, many LTER sites require that the LTER program and the individual investigators receive credit for their efforts. For example the Konza Prairie LTER site (KNZ) includes the following statement in their access policy:

> "Finally we ask all publications, reports and proposals who use any data from KNZ acknowledge/cite the KNZ program using the following statement: "Data for XXX was supported by the NSF Long Term Ecological Research Program at Konza Prairie Biological Station"; where XXX is the list of data sets used in the publications, reports or proposals."

The LTER data access policy has resulted in many significant changes, including an increase in the number of cross-site and multi-site analysis and synthesis activities, better data and science via enhanced attention to the quality of data and metadata, greater reliance on web-based databases to facilitate data accessibility, and growing recognition that data and metadata are valuable products of the research enterprise (in addition to publications).

3. THE LTER NIS INFORMATION SYSTEM

The Long-Term Ecological Research (LTER) Network Information System (NIS) is a cooperative, federated database system (Baker et al. 2000) supporting more than 1200 scientists and students that are investigating complex ecological phenomena at the 24 sites. The LTER NIS was established with the overall goal of implementing a transparent shared virtual environment

that is high-performance, distributed, and secure. The LTER NIS builds and expands upon successful grass-roots efforts that are initiated at one or more of the 24 LTER research sites. Currently, the LTER NIS consists of operational modules that are located at the LTER Network Office at the University of New Mexico as well as prototype modules that are being developed at individual sites. These modules are designed to facilitate intersite communication and technology transfer, and support research. Modules include a personnel directory, electronic mailing lists, bibliography, data catalog, site description directory, climate database, and aboveground primary productivity database [8].

Modularity and extensibility have been critical to the success of the LTER NIS. Modules may be developed at individual or cooperating sites, and often undergo extensive revision as they are implemented across the network. The "LTER cycle" is one whereby new software and information management approaches are experimented with at a few sites. In some cases, the experiment may be perceived as a failure and the software or information management approaches are abandoned. Conversely, the experiment may be viewed as successful and the software or information management approaches spread to other sites where they continue to be used and evaluated. The process is never static; new software and approaches continue to be evaluated and only those approaches that are robust and successfully scale to the full network are broadly adopted (John Porter, pers. communication;[8]).

Several important lessons have been learned from LTER NIS efforts. First, meeting standardized goals with a variety of site-specific solutions has built strength into the LTER Network. Second, leaving data at sites where they can best be managed, while making them accessible via a common interface represents a viable solution. Third, without an adequate information system and the requisite personnel, the significant effort required to manage large and complex data sets can present a substantial barrier to intersite research and synthesis. Finally, the successful evolution of the LTER NIS can be related to four key design elements: (1) partnerships and a central focus that are well defined; (2) explicitly stated individual site responsibilities; (3) modularity in software design; and (4) development of independent prototypes that undergo rigorous testing for scalability and sustainability prior to broad acceptance [8].

4. FUTURE OF ECOLOGICAL INFORMATICS IN LTER

Future efforts in LTER ecological informatics will focus on forming and expanding partnerships to facilitate data discovery, methods standardization, data sharing and integration, and data archival. There are currently many barriers to these activities. Discovery of ecological data remains a challenging endeavor, particularly for data that are not "registered" in a data directory and have not been cited in scientific peer-reviewed publications. Numerous opportunities will be explored for directly linking LTER data to national and international data cataloging efforts.

Non-standardized methods can inhibit data integration; conversely, adoption of standard methods supports intersite comparisons, facilitates documentation of the methods for publications and metadata, and can reduce project costs [9]. Collection of data in a consistent manner is vital. LTER scientists have published books like Standard Soil Methods for Long-Term Ecological Research [10] in an effort to encourage consistent data collection among sites. Similar efforts to standardize net primary production and other data types are underway.

Other barriers to data sharing and integration include inadequate and incomplete metadata [6], as well as concerns over publication rights, credit and other "rewards" for database development [11]. Many valuable ecological data sets are lost because researchers have few incentives to preserve their data once results have been published. An exception is the Ecological Society of America's Ecological Archives, which publishes peer-reviewed and cited data papers that consist of data sets and associated metadata deemed to be of significant interest to ecologists, thereby rewarding scientists for investing time in data archival. It is anticipated that significant future effort will be devoted to publishing LTER data in outlets like Ecological Archives.

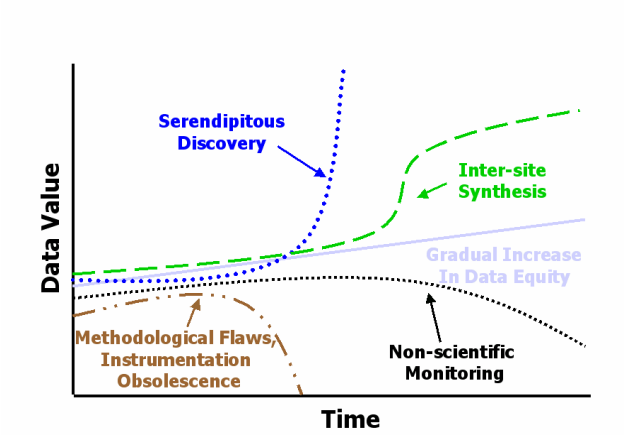

Figure 2. The coupling of comprehensive metadata with data generally facilitates the gradual increase in value of a data set over time.

Data sharing and collaboration can be further enhanced through the development of data distribution and archive centers, as well as information analysis centers (e.g., see [12]). Coupling comprehensive metadata with data in a data archive generally enhances the value of a data set over time (Figure 2). In some cases, well-documented

data will lead to an important serendipitous discovery whereby the data rapidly accrue new value. Similarly, utilization of documented data for an intersite synthesis project (e.g., broad-scale comparison) can increase the value of a particular data set for both the short term and long term. In contrast, however, well-documented data can lose value. For instance, discovery of methodological flaws or obsolescence of instruments may render a data set worthless. Likewise, monitoring without a clear scientific objective can lead to a data set that decreases in value over time. The next decade of LTER will likely see much more effort focused on partnering with environmental data archives and facilitating data integration and synthesis efforts through workshops, new technologies, and other mechanisms.

Rapid increases in environmental data holdings, as well as increasing demand for the data and information that are required for informed environmental decision-making, will require additional advances in data mining and knowledge discovery tools [13]. Particular emphasis in LTER will likely focus on value-added data products that can be more rapidly understood and used by decision-makers, resource managers, and educators. One example of progress in this area is the Ecology Explorers program at the Central Arizona Phoenix LTER site that directly involves K-12 students in the research process, including information management activities (see http://caplter.asu.edu/explorers/index.htm).

New challenges will confront LTER as new data types are increasingly integrated with ecological data. For example, increased collaboration with social scientists will require that information managers address the attendant data confidentiality issues of socioeconomic data. In addition, increased miniaturization of environmental sensors coupled with developments in wireless communication have enormous implications for LTER and will result in exponential increases in the volumes of data that sites manage. Such massive data streams will require new automated approaches for processing data streams and visualizing the resulting information.

Acknowledgments

LTER is supported by the National Science Foundation.

5. References

[1] Brunt, J.W. 2000. Data management principles, implementation and administration. Pages 25-47 in: *Ecological Data: Design, Management and Processing* (W.K. Michener and J.W. Brunt, eds.), Blackwell Science, Oxford, Great Britain.

[2] Franklin, J.F., C.S. Bledsoe, and J.T. Callahan. 1990. Contributions of the long-term ecological research program. *BioScience* 40:509-523.

[3] Cook, R.B., R.J. Olson, P. Kanciruk, and L.A. Hook. 2000. Best practices for preparing ecological data sets to share and archive. *Bulletin of the Ecological Society of America* 82:138-141.

[4] Porter, J.H. 2000. Scientific databases. Pages 48-69 in: *Ecological Data: Design, Management and Processing*. (W.K. Michener and J.W. Brunt, eds.), Blackwell Science, Oxford, Great Britain.

[5] Edwards, D. 2000. Data quality assurance. Pages 70-91 in: *Ecological Data: Design, Management and Processing* (W.K. Michener and J.W. Brunt, eds.), Blackwell Science, Oxford, Great Britain.

[6] Michener, W.K., J.W. Brunt, J. Helly, T.B. Kirchner, and S.G. Stafford. 1997. Non-geospatial metadata for the ecological sciences. *Ecological Applications* 7:330-342.

[7] Olson, R.J. and K.A. McCord. 2000. Archiving Ecological Data and Information. Pages 117-141 in: *Ecological Data:Design, Management and Processing,* (W.K. Michener and J.W. Brunt, eds.), Blackwell Science, Oxford, Great Britain.

[8] Baker, K.S., B.J. Benson, D.L Henshaw, D. Blodgett, J.H. Porter, and S.G. Stafford. 2000. Evolution of a multisite network information system: the LTER information management paradigm. *BioScience* 50:963-978

[9] Michener, W.K. 2000. Research design: translating ideas to data. Pages 1-24 in: *Ecological Data: Design, Management and Processing* (W.K. Michener and J.W. Brunt, eds.), Blackwell Science, Oxford, Great Britain.

[10] Robertson, G.P., D.C. Coleman, C.S. Bledsoe, and P. Sollins (eds.). 1999. *Standard Soil Methods for Long-Term Ecological Research.* Oxford University Press, New York.

[11] Porter, J.H. and J.T Callahan. 1994. Circumventing a dilemma: historical approaches to data sharing in ecological research. Pages 193-203 in: *Environmental Information Management and Analysis: Ecosystem to Global Scales* (W.K. Michener, J.W. Brunt and S.G. Stafford, eds.), Taylor and Francis, Ltd., London, England.

[12] Scurlock, J.M.O., R.J. Olson, R.A. McCord and W.K. Michener. 2002. Environmental data banks: archiving ecological data and information. Pages248-259 in: *Encyclopedia of Global Environmental Change* (E. Munn, ed), Vol. 2: The Earth system: biological and ecological dimensions of global environmental change, Ecosystems Section (H. Mooney and J. Canadell, eds.). John Wiley, Chichester.

[13] Michener, W.K. 2000. Ecological knowledge and future data challenges. Pages 162-174 in *Ecological Data: Design, Management and Processing* (W.K. Michener and J.W. Brunt, eds.), Blackwell Science, Oxford, Great Britain.