TOOLS FOR MANAGING ECOLOGICAL DATA*

J. H. Porter
University of Virginia
Charlottesville, Virginia, USA

R.W. Nottrott
University of Washington
Seattle, Washington, USA

Karen Baker
University of California, San Diego, SIO
La Jolla, California, USA

ABSTRACT

The Long-Term Ecological Research (LTER) Program is a set of 18 ecological research sites and a Network Office funded by the U.S. National Science Foundation. From the inception of the program in 1980, the LTER program has included a strong emphasis on data and information management as a crucial component of long-term research. The diversity of LTER sites (varying from arctic tundra to tropical rain forest), research questions and data collected (varying in size from <1 KB to >300 MB) demands a philosophy and structure which promotes flexibility and development of innovative solutions. This paper will focus on software tools which have become widely used within the LTER network. These include network information servers, geographical information systems, database and statistical packages as well as the myriad of smaller tools which link them together to work as an integrated whole. We will also discuss tools and approaches that we expect to become emergent standards in the future.

1.0 INTRODUCTION

The U.S. Long-term Ecological Research (LTER) is comprised of 18 ecological research sites and a Network Office supported by grants from the National Science Foundation. The LTER sites vary greatly in character from arctic tundra to tropical forest, with sites representing forest, desert, prairie, alpine, coastal and marine environments. The research topics addressed at sites are similarly varied, with research on ecological

* Presented at Eco-Informa '96, Lake Buena Vista, Florida, 4-7 November 1996.

succession, disturbance, landscape ecology, elemental cycling, trophic structure, biodiversity, organic matter and primary productivity taking place among the different sites. The sites also differ in their degree of centralization. Some sites are closely linked to investigators from only one university. Other sites have investigators from up to 15 different universities (Hayden, 1996).

The types of data collected at LTER sites are tremendously variable. Satellite images, scanned aerial photos, and output from spatially explicit models fall at the high-end of size, with an individual image requiring over 300 MB of disk space. At the other end of the spectrum are types of data, such as deep soil cores where a small number of samples are subjected to a large number of different tests. This may involve documentation files that are be orders of magnitude larger than the associated data files.

The diversity of the LTER sites, the scientific questions being addressed, the data being collected and the administrative structures under which they operate pose a significant challenge for information management efforts (Stafford et. al., 1994). The approach taken by the LTER Network since its inception in 1980 has been for each site to operate its own information management system under a data management policy that conforms to LTER Network guidelines, with the LTER Network Office taking the lead in linking those individual systems. This allows each site to focus on the types of data most prevalent at that site by choosing software and hardware compatible with the mission and the computational and administrative environments of the site.

2.0 SOFTWARE AND HARDWARE USED AT LTER SITES

Starting in 1992, annual surveys of the software and hardware environments used by LTER sites were conducted. Designed to facilitate cross site communication on computer tools, the 1995 survey showed almost all LTER sites made extensive use of "off the shelf" software for PC, Mac, UNIX and other computer platforms (Table 1). The greatest variety of packages was in the graphics software where the 28 solutions are distributed over PC, Macintosh and UNIX. This variety exists also with GIS and statistics software where approximately 20 different packages are used for each. Successful software transport is evident in the use of the statistics package SAS. This software, originally available only on the IBM computer, is now the most commonly used statistics package by LTER sites on PC and MAC platforms. The most common cross platform package by far was the GIS ARC/INFO (33 implementations on PC, Macintosh and UNIX). This is followed by the word processors Word (15 on PC and Macintosh) and the spreadsheet Excel (15 on PC and Macintosh). Availability on a variety of platforms does not appear to insure a package's use. Although some individual sites have adopted use of specific data-entry software, the trend toward use of microcomputers (Zinnel and Marozas, 1986) continues along with the use of generic spreadsheet software. PC are the most commonly

used platform in general but UNIX systems appear strongly used for GIS work and that the Macintosh is notable for diversity of graphics packages.

Table 1: Software and Hardware used at a minimum of 3 LTER sites.

	PC	Mac	UNIX			PC	Mac	UNIX
WORD-PROCESSORS*	5	3	5		**DATA-BASE***	5	2	9
wordperfect	16	-	1		dbase	7	-	-
word	8	7	-		ingress	1	-	3
framemaker	3	3	4		foxpro	3	-	-
clarisworks	-	3	-		oracle	3	-	-
latex	-	-	3		paradox	2	-	-
					ingress	1	-	3
BIBLIO-GRAPHIC*	6	1	7		foxpro	3	-	-
procite	6	-	-		oracle	3	-	-
papyrus	3	-	-		paradox	2	-	-
refmenu	3	-	-					
endnote	-	3	-		**GIS***	11	1	10
					arcinfo	8	10	15
SPREADSHEET*	4	2	4		erdas	5	-	7
quatropro	12	-	-		grass	1	-	7
lotus	9	-	-		arcview	3	-	2
excel	8	7	-		idrisi	2	-	2
					erdas-imagine	1	-	2
STATISTICS*	5	3	8					
sas	7	6	-		**DATA ENTRY***	12	1	2
systat	5	2	1		lotus	4	-	-
splus	3	-	2		excel	3	3	-
					dbase	3	-	-
GRAPHICS*	9	11	6					
sigmaplot	6	-	-					
sas-graph	4	-	-					
deltagraph	-	3	-					
cricketgraph	-	3	-					

*The first line gives the total number of packages used by platform while subsequent lines give the number of sites using a specific package.

3.0 MEETING DATA INPUT NEEDS

Collecting data and recording it in electronic form is difficult at LTER sites. The "field" is a notoriously dirty laboratory, with rain, snow, wind, heat and dust appearing when least needed. All LTER sites have meteorological stations with automated loggers which record weather data onto solid-state RAM packs (Ingersoll, 1994). In some instances, where a site operates multiple stations over a wide area (such as the Sevilleta LTER) radio transmitters are linked to the stations to provide real-time monitoring capabilities.

Some data is not amenable to direct electronic reading and requires a human to collect it. The Cedar Creek and Sevilleta LTER sites use small palm-top computers (such as the HP 200LX) to capture input in digital form in the field. There are also data that are collected on paper forms and entered electronically at a more secure location. Data entry for many North Temperate Lakes LTER data sets is performed on highly customized Excel spreadsheets. A special menu bar and locking of selected cells protects against entry errors. Error checking of values is also built into the spreadsheet through Excel formulas and lookup tables. At some other sites, data is input directly into a database. The Luquillo LTER site makes extensive use of the Paradox database on a PC for data input and manipulation.

4.0 METADATA AND DOCUMENTATION

One of the challenges for sites with a large number of investigators at different institutions is making the full power of the information management system available both locally and remotely. The Virginia Coast Reserve LTER site makes extensive use of on-line-forms, many of them linked directly to a relational database to transform the WWW into a two-way conduit for information, allowing on-line editing of WWW material for authorized users. The Konza LTER site, which has a more local user base, uses a Novell network server to allow investigators to directly manipulate and add metadata.

The Palmer LTER program operates off a ship near Antarctica. To meet information needs in a rigorous environment with only intermittent Internet access, a simple form of electronic notebook was developed . The electronic notebook fills the need for information in one location covering historical summaries, calibrations, procedures and methods, as well as past findings. Upon completion of a cruise, they reside in a shared, remotely accessible folder on a Macintosh.

Wide Area Information Servers (WAIS) have been used by the LTER Network Office search text databases from multiple LTER sites. WAIS servers are much simpler than relational DBMS, but their use is limited to metadata and other text-based

information such as bibliographies, catalogs and personnel directories that are located and maintained at widely dispersed sites.

5.0 DATA DISSEMINATION

Surveys of LTER sites a decade ago focused on storage media exchange (i.e. magnetic tapes, floppy disks), platform hardware (i.e. mainframe, mini, micro and supermicro) and telecommunication methods (Klopsch and Stafford, 1986). Currently the WWW is used at all LTER sites for the distribution of data. The Arctic and Coweeta LTER sites make use of scripts, written in the PERL language, to convert metadata files to HTML form and to assemble multiple files into an integrated WWW document.

The Andrews LTER site uses sets of standard metadata tables housed in a PC-based relational software package to integrate data input, quality assurance/quality control and automated generation of WWW pages. Metadata is input into a PC-database. These tables are then used to generate program code to read the data from its original form (if it were not already in database form) and reports and graphs highlighting inconsistencies in the data. The database is then used to produce a series of documentation files in HTML form which are uploaded to the site WWW server.

The Shortgrass Steppe LTER has progressed from locally-produced software for X-windows and towards packaged software designed for Internet-database connectivity. They have nearly all of their data stored in an ORACLE database. Using an ORACLE product (WebServer) they can pull any specific piece or set of data from the database and display it on the web. A similar WWW-based system is undergoing development at the North Temperate Lakes LTER site which currently uses Oracle Data Browser operating over a local area network.

6.0 DATA ARCHIVING

One of the challenges of managing environmental data is that the useful life of data almost always exceeds the life of the technologies used to store it. For this reason, LTER sites typically maintain archival copies of datasets in an ASCII form. These archival datasets are maintained on multiple media (e.g. floppy disks, CD-ROM, optical disk, magnetic tape). The Cedar Creek, Niwot Ridge, Sevilleta, and Shortgrass Steppe LTER sites have developed machine-readable forms of ASCII metadata and tools which facilitate transformation of archival ASCII files into alternative forms. For some small, but extremely valuable, datasets the Cedar Creek site has developed paper versions of archival datasets which incorporate checksum fields to allow automatic error detection for use with an optical character reader.

7.0 FUTURE DIRECTIONS

The LTER Network has always taken rapid advantage of improvements in information management technologies. The fastest growing area is the integration of SQL-based relational database management systems (DBMS) with the World Wide Web. This combination benefits from the strong query capabilities of the database software and the increasingly versatile interface provided by the WWW browsers incorporating the JAVA language. In addition, an emerging standard for network access to SQL-based DBMS, called ODBC for Open Database Connectivity, holds the potential for an infrastructure on which to build a distributed relational database from datasets maintained at individual sites. DBMS from all major vendors now have ODBC-compatible network database servers. ODBC clients are available "off the shelf" (e.g. Excel and Access). The Network Office has carried tests in which SQL tables served by Ingres servers located at Bonanza Creek LTER were viewed with an Access database running on a PC. Presently low effective bandwidth on many parts of the Internet imposes severe performance limitations. When that bottleneck is resolved, we expect this to be an extremely powerful technology.

This paper would not have been possible without the cooperation of information managers throughout the LTER Network.

LITERATURE CITED

B.P. Hayden "An LTER Profile: Fifteen years old, the LTER model is showing its worth" *LTER Network News,* Issue 19 p. 1, Spring/Summer 1996. Also available at: http://lternet.edu/about/program/table.htm

R. Ingersoll The management of electronically collected data within the Long-Term Ecological Research (LTER) Program. gopher://lternet.edu:70/11/newsletters/Reports/ElectronicData/Update94. 1994.

M.W. Klopsch and S.G. Stafford "The Status and Promise of Intersite Computer Communication" In *Research Data Management in the Ecological Sciences* ed. W.K. Michener University of South Carolina Press, 1986

S.G. Stafford, J.W. Brunt and W.K. Michener "Integration of scientific information management and environmental research" In *Environmental Information Management and Analysis: Ecosystem to global scales*, eds. S.G. Stafford, J.W. Brunt and W.K. Michener Taylor & Francis, Bristol, PA pp. 3-19, 1994.

K.C. Zinnel and M.F.Marozas. "Computer data entry techniques used Scientific Applications" In *Research Data Management in the Ecological Sciences*. ed. W.K. Michener. University of South Carolina Press, Columbia 1986.