DATA AND INFORMATION MANAGEMENT

IN THE ECOLOGICAL SCIENCES:

A RESOURCE GUIDE

CONTENTS

Data and Information Management

in the Ecological Sciences:

A Resource Guide

Edited by

William K. Michener, John H. Porter, and Susan G. Stafford

This publication should be cited as:

Michener, W.K., J.H. Porter, and S.G. Stafford. 1998. Data and information management in the ecological sciences: a resource guide. LTER Network Office, University of New Mexico, Albuquerque, NM.

Publisher: LTER Network Office, University of New Mexico, Albuquerque, NM

Preface

More than 100 individuals attended a two-day workshop (August 8-9, 1997) entitled "Data and Information Management in the Ecological Sciences" that was held at the University of New Mexico in Albuquerque. Objectives of the workshop were to: effect technology transfer, especially at biological field stations and marine laboratories; facilitate people networking; communicate training needs and opportunities; identify future needs for data management at field stations; and produce hard copy and digital versions of the proceedings. Workshop instructors provided comprehensive overviews of the technological infrastructure for a data management system (e.g., hardware, software, communications, and networking), data entry, quality assurance, database management systems, metadata, archival, the World Wide Web, and scientific visualization. Additional roundtable discussions focused specifically on software for field stations, challenges and opportunities at field stations, and site-specific data management implementation.

The editors are grateful to:

- the National Science Foundation Database Activities in the Biological Sciences Program for funding the workshop (DBI 97-23407);
- the LTER Network Office, Robert W. Woodruff Foundation, and Long-Term Studies Section of the Ecological Society of America for additional funding and support;
- William Michener, James Gosz (University of New Mexico), Arthur McKee (Oregon State University), and John Porter for writing the proposal and organizing the workshop;
- James Brunt and John Porter for WWW support;
- Jack Stanford (Organization for Biological Field Stations) and Frank Davis (National Center for Ecological Analysis and Synthesis) for their support and encouragement;
- staff members associated with the LTER Network Office,

Sevilleta LTER Project, and the University of New Mexico who facilitated the workshop;

- Jim Beach, Tom Callahan, and Scott Collins (current or former NSF program officers) who have been strong proponents of enhanced data management and computational capabilities at biological field stations and marine laboratories;
- Kathleen Parkhurst, Ellie Trotter, Paula Houhoulis, and Paula Johnson for coordinating onsite logistics;
- Paula Johnson, Patty Sprott, Paula Houhoulis, and Jean Turn for technical editing and production of the hardcopy version of the workshop proceedings;
- Jean Turn for design of the Web version of the workshop proceedings;
- the instructors who all gave 150% effort, particularly Hilary Swain whose caricatures of insects and synthetic abilities were sincerely appreciated; and
- the attendees for their attentiveness and enthusiasm.

To all of the above, we offer a heart-felt THANKS!

Data and Information Management

in the Ecological Sciences:

A Resource Guide

Edited by

William K. Michener, John H. Porter, and Susan G. Stafford

This publication should be cited as:

Michener, W.K., J.H. Porter, and S.G. Stafford. 1998. Data and information management in the ecological sciences: a resource guide. LTER Network Office, University of New Mexico, Albuquerque, NM.

Publisher: LTER Network Office, University of New Mexico, Albuquerque, NM

Preface

More than 100 individuals attended a two-day workshop (August 8-9, 1997) entitled "Data and Information Management in the Ecological Sciences" that was held at the University of New Mexico in Albuquerque. Objectives of the workshop were to: effect technology transfer, especially at biological field stations and marine laboratories; facilitate people networking; communicate training needs and opportunities; identify future needs for data management at field stations; and produce hard copy and digital versions of the proceedings. Workshop instructors provided comprehensive overviews of the technological infrastructure for a data management system (e.g., hardware, software, communications, and networking), data entry, quality assurance, database management systems, metadata, archival, the World Wide Web, and scientific visualization. Additional roundtable discussions focused specifically on software for field stations, challenges and opportunities at field stations, and site-specific data management implementation.

The editors are grateful to:

- the National Science Foundation Database Activities in the Biological Sciences Program for funding the workshop (DBI 97-23407);
- the LTER Network Office, Robert W. Woodruff Foundation, and Long-Term Studies Section of the Ecological Society of America for additional funding and support;
- William Michener, James Gosz (University of New Mexico), Arthur McKee (Oregon State University), and John Porter for writing the proposal and organizing the workshop;
- James Brunt and John Porter for WWW support;
- Jack Stanford (Organization for Biological Field Stations) and Frank Davis (National Center for Ecological Analysis and Synthesis) for their support and encouragement;
- staff members associated with the LTER Network Office, Sevilleta LTER Project, and the University of New Mexico who facilitated the workshop;
- Jim Beach, Tom Callahan, and Scott Collins (current or former NSF program officers) who have been strong proponents of enhanced data management and computational capabilities at biological field stations and marine laboratories;
- Kathleen Parkhurst, Ellie Trotter, Paula Houhoulis, and Paula Johnson for coordinating onsite logistics;
- Paula Johnson, Patty Sprott, Paula Houhoulis, and Jean Turn for technical editing and production of the hardcopy version of the workshop

proceedings;

- Jean Turn for design of the Web version of the workshop proceedings;
- the instructors who all gave 150% effort, particularly Hilary Swain whose caricatures of insects and synthetic abilities were sincerely appreciated; and
- the attendees for their attentiveness and enthusiasm.

To all of the above, we offer a heart-felt THANKS!

CONTRIBUTORS

Karen S. Baker

Scripps Institution of Oceanography, University of California at San Diego,

La Jolla, CA 92093-0218

Barbara J. Benson

Center for Limnology, University of Wisconsin-Madison, 680 N. Park Street,

Madison, WI 53706

Frederick A. Bierlmaier

Forest Science Department, Oregon State University, Corvallis OR 97331

Darrell Blodgett

Forest Soils Laboratory, University of Alaska, Fairbanks, AK 99775

John M. Briggs

Division of Biology/Ackert Hall, Kansas State University, Manhattan, KS 665026-4901

Scott E. Chapal

Joseph W. Jones Ecological Research Center, Route 2 Box 2324, Newton, GA, 31770

Don Edwards

Department of Statistics, University of South Carolina, Columbia, SC 29208

Hazel E. Hammond

U.S. Forest Service Pacific Northwest Research Station, 3200 SW Jefferson,

Corvallis, OR 97331

Mike Hartman

University of Colorado, INSTAAR, Campus Box 450, Boulder, CO 80309-0450

John J. Helly

San Diego Supercomputer Center, MS 0505, University of California, San Diego

La Jolla, CA 92093

Donald L. Henshaw

U.S. Forest Service Pacific Northwest Research Station, 3200 SW Jefferson,

Corvallis, OR 97331

Melissa E. Holmes

The University of Montana, Flathead Lake Biological Station, 311 Bio Station Lane,

Polson, MT 59860-9659.

Rick Ingersoll

Cornell University, Biometrics Unit, 441 Warren Hall, Ithaca, NY 14853

Matthew B. Jones

National Center for Ecological Analysis and Synthesis, 735 State St., Suite 300,

Santa Barbara CA 93101

Rudolf Nottrott

National Center for Ecological Analysis and Synthesis, 735 State St., Suite 300,

Santa Barbara, CA 93101

Raymond A. McCord

Oak Ridge National Laboratory, P.O. Box 2008 Oak Ridge TN 37831-6407

William K. Michener

Joseph W. Jones Ecological Research Center, Route 2, Box 2324, Newton, GA 31770

Richard J. Olson

Raymond A. McCord

Oak Ridge National Laboratory, P.O. Box 2008 Oak Ridge TN 37831-6407

Geoffrey C. Poole

The University of Montana, Flathead Lake Biological Station, 311 Bio Station Lane

Polson, MT 59860-9659

John H. Porter

Department of Environmental Sciences, University of Virginia, Charlottesville, VA 22903

Mark P. Schildhauer

National Center for Ecological Analysis and Synthesis, 735 State St., Suite 300, Santa Barbara. CA 93101

Susan G. Stafford

Department of Forest Science, Oregon State University, Corvallis, OR 97331-7501 (After September 1, 1998: Department of Forest Sciences, Colorado State University, Fort Collins, CO 80523)

Maryan Stubbs

Center for Limnology, University of Wisconsin-Madison, Madison, WI 53706

Hilary M. Swain

Archbold Biological Station, Lake Placid, FL 33862

ISSUES AND CONCEPTS OF DATA MANAGEMENT: THE H.J. ANDREWS FOREST SCIENCE DATA BANK AS A CASE STUDY

Susan G. Stafford

Department of Forest Science, Oregon State University, Corvallis, OR 97331-7501

Abstract. Managing scientific research information to promote ecological research and facilitate widespread availability to the broader scientific community can be an overwhelming task. The Quantitative Sciences Group in the Department of Forest Science at Oregon State University helps address this need within a context of integrating research information management into the research planning process. The history of the Forest Science Data Bank is described. Lessons learned and strategies for successful long-term ecological information management are shared.

INTRODUCTION

The scientific community is in the midst of an information explosion coupled with a technology revolution (Stafford et al. 1994). The amount of data beaming down from satellites over shorter and shorter time scales, over larger and larger regions, has been likened to receiving a "Library of Congress" worth of data everyday. Data acquisition is clearly not the problem anymore; managing the data is the challenge! The WWW and Internet connectivity have fueled the scientific community's and funding agencies' expectations for ready-access to on-line data and metadata (i.e., documentation essential for understanding the who, what, when, where, how of the data). Complex issues (e.g., global change, sustainability, biodiversity, and emerging diseases) require interdisciplinary collaboration and synthesis at much broader spatial and longer temporal scales (Levin 1992, Kareiva and Anderson 1988). These issues are faced by individual scientists working alone or in teams associated with personnel at independent biological field stations or as part of the National Science Foundation-funded Long-Term Ecological Research Network. The administrative organization and affiliation are independent of the growing expectation for sound data management policies and procedures.

THE FOREST SCIENCE DATABANK

It is important to have a well-defined statement of purpose for managing long-term research information. The mission of the Quantitative Sciences Group (QSG), staffed by both Oregon State University and U.S. Forest Service Pacific Northwest Station personnel, is threefold: to enable success, solve problems, and promote scientific exploration. Our goals are to facilitate research, as well as anticipate future needs. To be successful, research information management must be integrated into research planning. The systematic approach we have used at the H.J. Andrews Long-Term Ecological Research site (LTER) (Franklin et al. 1990) is comprised of: study planning, data production, data analysis, and data interpretation and synthesis (Stafford 1993). This approach can be implemented easily at other research stations.

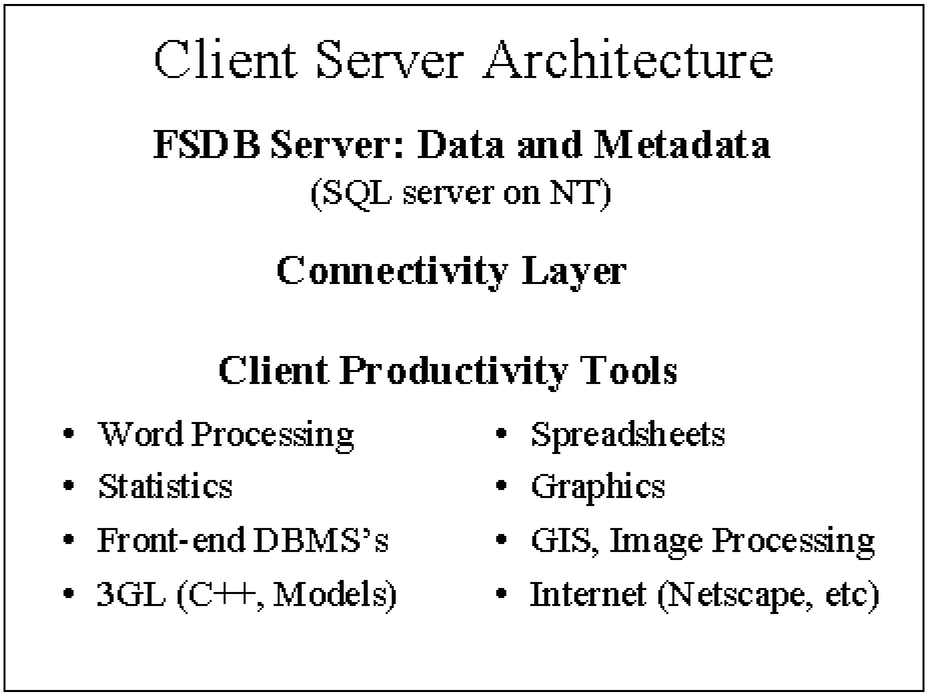
The Forest Science Data Bank (FSDB) (Stafford et al. 1984, 1986, 1988) was developed by QSG to house data generated by LTER and collaborating scientists. The FSDB has enjoyed a rich history beginning in 1948 when the Blue River Experimental Forest was established (renamed the H.J. Andrews Experimental Forest in 1953), through the decade of the 1970's and the International Biome Program. Our goal has always been to keep improving and changing, mirroring, to the best of our ability, ever-present changes in technology (Table 1). Figure 1 depicts the three layers comprising the current FSDB: the FSDB server housing data and metadata, a connectivity layer, and client productivity tools.

Table 1. History of the FSDB.

1973-80	Data on mainframe tapes, paper documentation, early abstract and format forms
1980-84	Tape library with automated access facility, documentation in CP/M databases, formalized abstracts, formats & codes.

1984-88	Transition to stand-alone PCs, metadata ported to Xbase, converted mainframe applications to PCs.
1988-93	Tape library ported to Novell server, restructured and cleaned LTER database, development of generic maintenance tools.
1993-96	Refined QC procedures, establish presence on World Wide Web
1997-	Planning port of FSDB onto SQL-server, normalize and expand metadata database.

Figure 1. FSDB client server architecture.



Client Server Architecture

FSDB Server: Data and Metadata
(SQL server on NT)

Connectivity Layer

Client Productivity Tools

- Word Processing
- Statistics
- Front-end DBMS's
- 3GL (C++, Models)
- Spreadsheets
- Graphics
- GIS, Image Processing
- Internet (Netscape, etc)

The FSDB "enterprise" encompasses several components: data; documentation (metadata); hardware/software, connectivity tools, and personnel. Space limitations preclude delving deeply into all aspects. Relevant literature is included in the bibliography. Taking a holistic, enterprise-view toward data management allows for a more balanced approach, including effective strategies for dealing with the various critical components and interrelated issues that must be addressed for continuity and long-term success.

Data

Recognizing that data and metadata are a "corporate asset" and need to be managed as such, the FSDB (*http://www.fsl.orst.edu/lterhome.html*, *http://www.fsl.orst.edu/fslhome.html*) currently houses over 2000 data sets from more than 250 studies. FSDB data include legacy data sets (e.g., IBP data sets), 500 Gb of spatial data (i.e., geographic information system (GIS) coverages and remotely sensed images), models, as well as text documents.

Legacy FSDB datasets include: aquatic/hydrology, geomorphology/vegetation, meteorology, terrestrial vegetation/litter decomposition, biodiversity, wildlife ecology, forest science/genetics/forest engineering, vegetation management and soils data. Specifically, we have over forty years of meteorological and hydrological records (see Henshaw, Bierlmaier, and Hammond, this volume), over eighty years of forest growth and mortality records from six western states, and over thirty years of continuous vegetation succession data.

The legacy data sets are a double-edged sword. Clearly, they add immensely to our wealth of long-term data and significantly increase our overall "portfolio" of data resources. They also pre-date, in many instances, computer technology, so data collection was done differently then and far more care is required to insure adequate documentation of field procedures and study objectives. In addition, the original researchers are frequently no longer on-site and some are deceased.

More recent acquisitions of the FSDB include: a multinational LTER woody debris decomposition project (Harmon 1991), large-scale bird and bird habitat surveys (McGarigal and McComb 1992), and a historical fish habitat database on the Columbia River Basin conducted by the U.S. Fish and Wildlife Service in the 1930's and 1940's (McIntosh et al. 1992).

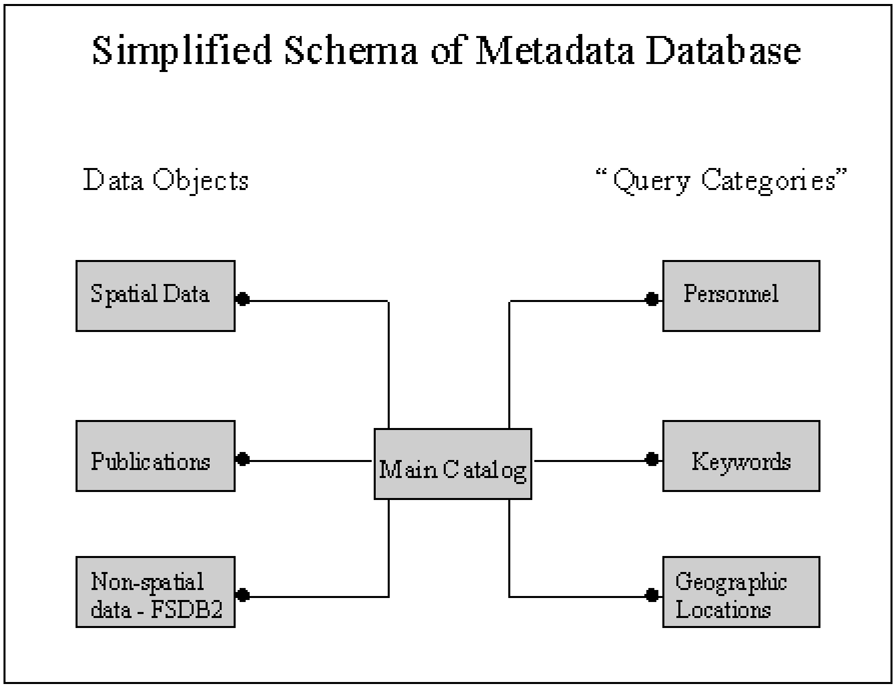


Figure 2. Metadata database structure.

Metadata

Metadata and data are equally important in the FSDB. The FSDB Metadata system includes: database catalogues, table definition files, domain tables, and tables containing database-specific rules records (for specific examples, see Henshaw, Bierlmaier, and Hammond, this volume; for a more generic discussion of data quality, see Edwards, this volume). We are in the process of further refining our metadata database structure (Figure 2). Examples of various metadata forms (abstract, variable format, variable definition, code definitions, etc.) have been published elsewhere (Stafford 1993).

We have used standardized metadata structures that are identical for every database. We use metadata for data presentation, guiding users in understanding database content, supporting global queries of data catalogs, generating data set documentation exports, and enabling generic access functions [e.g., web page creation, automatic import/export of flat files to relational database management systems (RDBMS) files]. We have used the metadata to develop project-specific "rules" for individual databases. For example, using metadata from the Andrews Reference Stand Monitoring Study, we can run quality assurance (QA) and quality control (QC) checks on newly entered data (for more on this topic, see Edwards, this volume). We can flag entries where trees changed species, shrank dramatically, grew dramatically, or came back to life after being dead for several years.

Hardware, software, connectivity, and personnel

Hardware decisions need to be considered in conjunction with software when assessing connectivity and making personnel decisions. We support six operating systems (AIX, SunOS/Solaris, Data General, Macintosh, Windows-NT, DOS/Windows 3.1x) on 5 platforms (IBM, Sun, Data General, Macintosh, Intel). Personnel are critical to the success of the whole operation. This equates to over 1300 user units. We have been extremely fortunate to hire individuals who are both computer-savvy, as well as interested in science. This has been a winning combination. Disciplinary interests of personnel include soils, statistics, entomology, GIS, remote-sensing, and computer science.

As a point of reference, from 1994 to 1995, the Novell LAN expanded from 180 to 280 PCs and more than tripled the amount of disk space from 5.8 Gb to 18.2 Gb. The UNIX network expanded from eighteen to 27 Suns and from 9.4 Gb public disc space to 25 Gb. In 1997, 450 PCs and 45 Sun Workstations were supported. All decisions need to be considered with an eye toward growth and scaleability (See Porter, this volume).

LESSONS LEARNED AND STRATEGIES FOR SUCCESS

- Strive to avoid creating an US versus THEM situation. Be creative with incentive programs to get buy-in from the research community served. Incentives can include user-friendly applications, "clean" data, shorter delays between collection and analysis, safe storage, and data back-ups.
- Data by themselves are of little value. Metadata must be maintained and recorded to insure the longevity and long-term utility of the data. We have found great variation between data managers' and researchers' expectations of what constitutes "adequate" metadata. Data managers must work with researchers to help educate and train the rest of the scientific community on what expectations must be met. It is also important to keep track of data requests.
- As you face the future, plan for growth. It is predictable that your systems will grow and expand.
- It is important to start the tradition of managing research information. Use the technology to your advantage and remember, the genius of the future lies not in technology, but in your ability to manage it.

ACKNOWLEDGMENTS

I would like to acknowledge National Science Foundation support from DEB No. 90-11663 and 96-32921 to the H.J. Andrews LTER project.

LITERATURE CITED

Franklin, J.F., C.S. Bledsoe, and J.T. Callahan. 1990. Contributions of the long-term ecological research program. BioScience 40(7):509-23.

Harmon, M.E. 1991. The long-term intersite decomposition experiment team (LIDET). Soil Ecology Meeting, April 1991, Oregon State University, Corvallis, OR (Abstract).

Harmon, M.E. 1992. Long-term experiments on log decomposition at the H.J. Andrews Experimental Forest. USDA Forest Service General Technical Report PNW-GTR-280.

Kareiva, P. and M. Anderson. 1988. Spatial aspects of species interactions: the wedding of models and experiments. Pages 38-54 in A. Hastings, editor. Community ecology. Springer Verlag, New York, NY.

Levin, S.A. 1992. The problem of pattern and scale in ecology. Ecology 73(6):1943-67.

McGarigal, K. and W.C. McComb. 1992. Streamside versus upslope breeding bird communities in the central Oregon Coast Range. Journal of Wildlife Management 56:10-23.

McIntosh, B. A., J. R. Sedell, and S. E. Clarke. 1992. Historical changes in anadromous fish habitat in the Upper Grande Ronde River Basin, Oregon, 1941-1990. Seventh Annual US Landscape Ecology Symposium, April 1992, Oregon State University, Corvallis, OR.

Stafford, S. G. 1993. Data, data everywhere but not a byte to read: managing monitoring information. Environmental Monitoring and Assessment 26:125-141.

Stafford, S. G., P. B. Alaback, G. J. Koerper, and M. W. Klopsch. 1984. Creation of a forest science data bank. Journal of Forestry 82(7):432-433.

Stafford, S. G., P. B. Alaback, K. L. Waddell, and R. L. Slagle. 1986. Data management procedures in ecological research. Pages 93-113 in W. K. Michener, editor. Research data management in the ecological sciences. The Belle W. Baruch Library in Marine Science No. 16. University of South Carolina Press, Columbia, SC.

Stafford, S. G., G. Spycher, and M. W. Klopsch. 1988. Evolution of the Forest Science Data Bank. Journal of Forestry 86(9):50-51.

Stafford, S. G., J. W. Brunt, and W. K. Michener. 1994. Integration of scientific information management and environmental research. Pages 3-19 in W. K. Michener, J. W. Brunt and S. G. Stafford, editors. Environmental information management and analysis: ecosystem to global scales. Taylor & Francis, London, UK.

TECHNOLOGICAL UNDERPINNINGS: HARDWARE

Scott E. Chapal

Joseph W. Jones Ecological Research Center, Route 2, Box 2324,

Newton, GA 31770

Abstract. Choosing appropriate computing hardware is challenging in this era of rapidly changing technologies. Hardware purchasing decisions affect long-term information management because of the large capital investment and the necessity to design a computing infrastructure which can withstand software upgrade cycles and provide operating system inter-operability. The increasing prominence of 'the network' in all aspects of information management has contributed to a re-alignment of hardware procurement budgets, with a larger proportion allocated to support 'bandwidth' requirements. Servers are re-establishing their positions of central importance in all computer networks while client computers are being standardized, simplified and increasingly required to perform terminal duties. The requirements of ecological information management are not extraordinary compared to other data-intensive endeavors, but the potentially contradictory demands of research, archival, analysis and collaboration can overwhelm an inappropriately designed computing infrastructure.

INTRODUCTION

The primary construct common to all modern collaborative computing is the Local Area Network (LAN). The dominant uses of LAN's are various forms of client/server computing (see Nottrott this volume, Schildhauer this volume). Client/server computing models allow flexibility in the allocation of data and processing resources and provide for the evolution of hardware use and network design. It is nonsensical to make hardware purchasing decisions without understanding operating system demands, application requirements, and the possible design alternatives available in various client/server models.

Perhaps more important than the technical considerations are site-specific needs that must be central to LAN design and, therefore, for hardware acquisition decisions. LAN design can be approached from several perspectives, but in the context of this chapter, three are paramount: 1) design to optimize the environment for research information management, 2) design to maximize system administration efficacy, and 3) design to maximize cost/benefit. Recommendations for specific hardware vendors, models or strategies are omitted herein because of the transient utility of that kind of information. The pace of change in these technologies renders specifics virtually obsolete by the time they can be drafted.

LAN design goals for research information management

A fundamental requirement of any computer environment is to provide access to the user population. Scientists, staff, students and temporary personnel must all have access to the computing resources of the site or project. A basic design goal that results from this need is to provide *interface consistency* and a *single log-on* paradigm for the user. The consequence of this goal for hardware acquisition is to minimize the number of supported platforms. The procurement ramification of this reasoning is to consolidate the number of vendors to a strategically selected minimum.

A related goal is to *accommodate mobility*, or to give users access to resources from all points on the network. This means that the design will achieve a many-to-many relationship of people to computers rather than a one-to-one relationship that was the hallmark of the stand-alone PC. Although there are researchers who do primarily use a single computer in an office, the flexibility and utility of the LAN is enhanced dramatically when this many-to-many relationship is established. Security implications of this arrangement are immediately apparent, but the details of security planning are best left to another forum (see Nottrott this volume).

It should be obvious that the facilitation of data collection and processing in the research environment is absolutely essential. Therefore, data entry protocols (see Briggs et al. this volume) must be available and integrated into the network design, and instruments must be interfaced to their respective computers and those computers to the network. Although details will vary from situation to situation, it is desirable to simplify and standardize, and this applies to hardware as well as to network protocols and applications.

In order to support a cogent data management framework, data storage, data organization and data security must be thoroughly considered and incorporated into fileserver design. Data can be centralized, secure, accessed by multiple client computers, adequately backed-up and redundantly configured on servers. In contrast, a completely de-centralized collection of client computers, all serving as data repositories in peer-to-peer relationships, is extremely difficult to manage and use. The utility of the client/server architecture is obvious given the lack of alternatives that can scale to accomplish increasing research demands.

Data management/analytical software tools represent an arena where standardization and consolidation should also be design goals. Planning for these tools should be on a time cycle that is longer than the upgrade cycle of operating systems and probably longer than the turnover frequency of hardware. The investment in these tools via programming and data structures can be quite high and should provide longevity and continuity to meet the long-term information and research demands. Much of the research agenda of individual scientists and institutions is now interdisciplinary and requires synthesis to address broad-scale and long-term questions. The simple fact that collaboration is necessary should be designed into computing infrastructure planning at all levels, including hardware specification and procurement.

LAN design goals for system administration

Another perspective from which to address infrastructure development, is from the system administrator. Given that resources for system administration are often limited and difficult to expand, it is prudent to make decisions that reduce administration workload. Obvious ways to simplify network operations are to centralize administration, and standardize the hardware, operating systems, and software supported. Applying conventions to all system administration functions (userID's, name service conventions, filesystem layout, computer names, IP address allocation, mail aliases), is essential to laying the groundwork for automation of tasks. Automation is a powerful way to accomplish repetitive tasks, thereby freeing the system administrator's time for problem-solving or project development. Simplifying installations of client operating systems and applications is especially important for organizations that have more than a dozen or so computers.

Cost/benefit perspective

Addressing hardware from a cost/benefit perspective is relatively straightforward -- maximize network functionality per dollar invested. This goal is conceptually simple, but its implementation is rather more complicated, and depends largely on the information management goals and system administration constraints. A common strategy is to 'Right Size' which translates to: 'Don't buy what you don't need'. This is more challenging when decision-making for technology purchases is distributed. Individuals may not be aware of the broader organization's needs or priorities and often make decisions based only on a single project, investigator, or end-of-year surplus budget. Avoiding redundant purchases and budgeting across project boundaries can be difficult, but given the extreme cost of the technology, it can be well worth the effort. A structured approach to building consensus through a committee can be used to help with these decisions.

Another economic motivation is to attempt to *future proof* the investment in computer hardware. This can often be likened to forecasting the future with a crystal ball, but there are some basic assumptions that hold true. 1) Delay acquisitions to the extent possible because hardware gets cheaper, faster and better every day. 2) Try to extend the longevity of components by predicting their useful life-span and their potential to be re-deployed to secondary functions. The consequence may be to buy fewer, better components.

Hardware is the most persistent part of the infrastructure (if hardware components are purchased rather than leased), and therefore must be able not only to accomplish today's needs, but be sufficiently upgradeable or re-deployable to have enduring utility. Leasing may be a valid option for some organizations, projects, or individuals and must be analyzed on a case-by-case basis. Non-profit organizations, for example, may not have the tax incentives that make leasing attractive to some businesses.

RIGHT-SIZING THE COMPUTER INFRASTRUCTURE

As alluded to in the previous section, right-sizing the computer infrastructure means, fundamentally, to avoid investing in unnecessary hardware or technology that does not appropriately fill the need. Needs-based planning, to use both an ecological and utilitarian metaphor, should be both top-down and bottom-up. Staff may argue as to who or what's on top versus bottom, but regardless, planning needs to occur from both the vantages of: 1) budgetary and personnel constraints, and 2) projects, goals, and other aspirations. A technology decision-maker may have to compromise incongruous demands -- usually too much ambition for too little money. To consciously avoid this decision process, however, may undermine the utility of the network over the long-term by specifying too little, or may unnecessarily inflate the technology budget by specifying too much. An example of this short-sightedness would be if budgetary allocation were too heavily skewed toward client computers and not sufficiently devoted to network hardware (hubs, switches): bandwidth could become a bottleneck, ironically just when high powered CPUs could take advantage of it! It may be difficult to strike a balance, and that balance, once achieved, will definitely change quickly.

Most organizations, by necessity, have to incrementally improve their computer systems by building on legacy systems (the existing hardware and software that comprise the computing infrastructure). Usually, legacy systems are both an asset and a constraint, but the fact that they do accrue over time underscores the need to assess technology acquisitions (hardware purchases) for their entire life cycle and total cost of ownership. Evaluating cost of acquisition without regard to longer term personnel, budgetary, applications and research realities can result in more expensive solutions over time. A key is to implement for change. Hardware can be cycled to secondary functions as it ages, for example, but this perspective assumes institutional-level planning and coordination.

It is important to understand the implications of standards on interoperability. Today's market leader can become tomorrow's albatross in the realm of proprietary network protocols and database technology. Therefore, it is a good idea to have at least a cursory understanding of the existence of standards in various areas of networking and database inter-operability. This understanding will influence hardware purchases ultimately, since all component devices will need to communicate throughout their life-cycle. One only needs to talk to the systems administrator of a multi-protocol LAN to understand the complications that can result from a proliferation of proprietary network operating systems. Some of the standards that have evolved to address these issues are:

OSI - Open Systems Interconnection reference model

TCP/IP - Transmission Control Protocol/ Internet Protocol

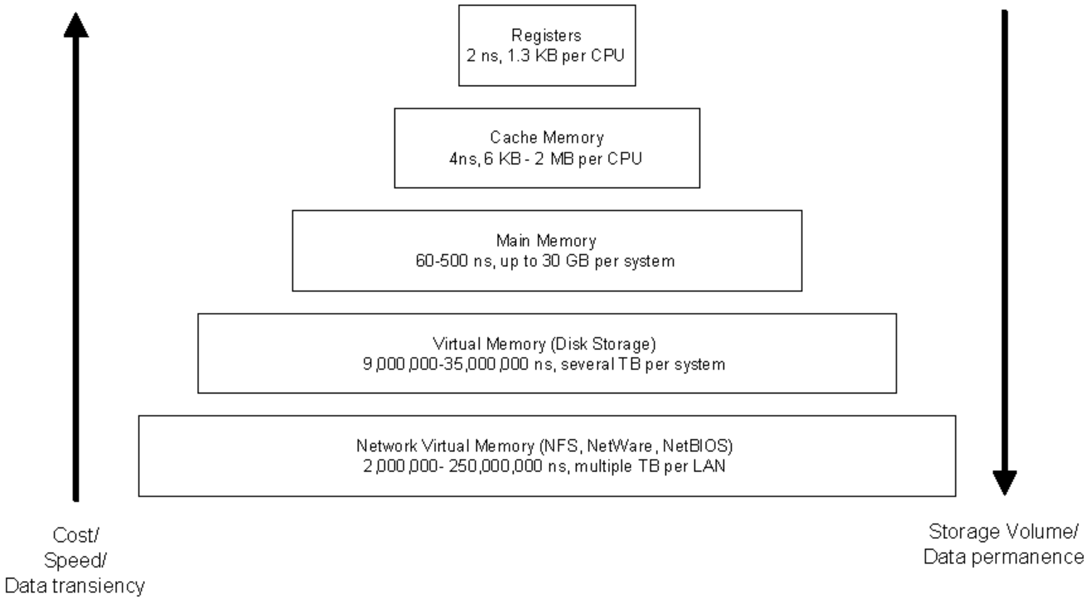
POSIX - Portable Operating System Interface for Computing Environments

CORBA - Common Object Request Broker Architecture

The salient point is that standards can persist to a greater extent than proprietary implementations, and the universal motivation for them to do so is the need to interoperate. TCP/IP is the most recognized example of an accepted standard, and has become essentially ubiquitous in the Internet and in LANs. SNA and NetWare are testaments to proprietary architectures that functioned well in organizations, but their vendor-specific nature compromised their scalability and rendered them inelegant to incorporate into the Internet. For the Internet/WWW to continue to interoperate at ever greater levels of complexity, these *dejure* standards must continue to be respected and evolve, notwithstanding *defacto* standards that do achieve a level of interoperability such as Microsoft's operating systems dominance. It is worth noting that the address space afforded by the 32 bit Internet Protocol is quickly becoming saturated and the transition to IPv6 with it's128-bit address scheme is only achievable because of the acceptance of the standards process.

Right-sizing can also be viewed from the perspective of balancing the elements of the 'Virtual Storage Hierarchy' (Wong 1997). This is useful, because the utility of an individual computer or a LAN can be understood in terms of its ability to move data to where the data are needed: in the CPU, into RAM, cached on disk, on a server, or archived to some storage media. This illustrates the balance and distribution of processing and storage resources on a network and further emphasizes the advantage of designing functionality into the entirety of the LAN instead of into individual computers.

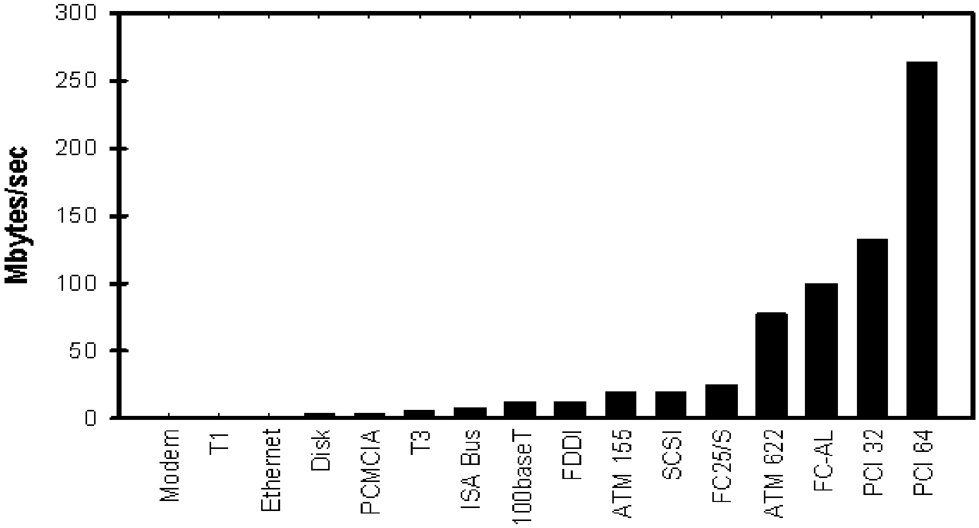
Figure 1. The virtual storage hierarchy (from Wong 1997).


Registers
2 ns, 1.3 KB per CPU
Cache Memory
4ns, 6 KB - 2 MB per CPU
Main Memory
60-500 ns, up to 30 GB per system
Virtual Memory (Disk Storage)
9,000,000-35,000,000 ns, several TB per system
Network Virtual Memory (NFS, NetWare, NetBIOS)
2,000,000- 250,000,000 ns, multiple TB per LAN
Cost/
Speed/
Data transiency
Storage Volume/
Data permanence

A complete description of LAN topologies, media and types is beyond the scope of this chapter, but it is important to briefly describe the models that have become standard. Ethernet (10 Mbps) over twisted pair wiring is a dominant model in modern LANs. In fact, over 80 percent of all network connections were Ethernet by the end of 1996. Star topologies are commonly designed into buildings where the media can be easily reconfigured to centralized hardware. Token Ring persists in IBM environments, and FDDI and (increasingly) ATM provide backbone capabilities. Heightened demand for bandwidth has resulted in 100 Mbps ethernet to the desktop and even gigabit (1 Gbps) ethernet is slated for standards adoption [IEEE 802.3z] in early 1998 *(http://www.gigabit-ethernet.org)*. Just as important is the transition to switched technology, which is rapidly replacing older shared ethernet segments, providing dedicated bandwidth improvements of an order of magnitude.

Although the network is increasingly critical to collaborative data access and processing, it is also the slowest part of client/server transactions (Figure 2). This is especially true in WANs, but even on LANs, data transfer is slow relative to the internal components of the computer. The situation is improving quickly, however, with gigabit ethernet, Fiber Channel Arbitrated Loop [FC-AL] *(http://www.fiberchannel.com)* peripheral interfaces (rapidly replacing SCSI), and PCI buses with dramatically increased throughput. If transfer rates (bandwidth) are represented graphically in common units of measure, the comparison is dramatic (Figure 2).

Figure 2. Bandwidth vs. component expressed in common units illustrating the relatively slow performance of the network.



Rationale for a client/server architecture

Given the constraints of network speeds and the costs involved in providing bandwidth and processing enhancements, the motivation to use client/server solutions may be obscure to many people. There are many reasons for the success of client/server as the dominant computing paradigm of our time, most of which derive from the economics of computers and the universal (corporate) requirement for integration and collaboration of many people into large projects. The business and technical incentives that favor client/server may be tangential to research information management concerns, but the application of client/server solutions to infrastructure needs in our domain is inevitable.

The benefits of client/server computing are simple to understand. Primary among these strengths is scalability: i.e., the ability to enhance or reconfigure components of the client/server architecture simply and in proportion to need. Scaleable designs are very important, both economically and for performance tuning and problem resolution. The fact that client computers have powerful processors is key to many client/server implementations and contributes to the scalability of the system by dedicating significant processing capabilities to the user. While the basic role of the client computer is understood to mean providing the computer to appropriately address user requirements, the optimal size and description of the client is hotly debated.

The 'thin client' as represented by the Net Computer Initiative, proposes to simplify client hardware and configuration to provide basic network access. This trend is premised on the increasingly central role of the server to Intranets via Web paradigms and the potential of computing platforms such as Java^{TM}. The criticism of traditional PCs is that they are difficult and expensive to manage and are inappropriate for many users who are mainly 'data consumers'. The lack of early adoption of the Net Computer as a PC replacement is at least in part because of Microsoft's antagonism to the model, until very recently. Even Microsoft, through its development of a multi-user version of NTServer - [Windows-Based Terminal Server - *Hydra*] is addressing the need to provide client/server computing to thin clients and aging, underpowered PCs, albeit in their characteristically proprietary manner. There are at least two end points on the spectrum of client computer evolution. The traditional PC is now characterised as a 'Fat' client while the Net Computer typifies the 'thin' client. Realistically, the entire spectrum will be represented for the foreseeable future.

Another consequence of the asymmetry of client/server design is that the client becomes increasingly generic and interchangeable while the server is managed for high availability. The partitioning of logic between the client and server is inherently flexible and further enhances the scalability of these designs. While the economics of computer system design dictate that it is much easier and less expensive to build 100 small computers than a single system that is 100 times as powerful, consolidation of processing on servers does occur to provide specialized functions. Specifically, fileserver, DBMS and computational server functions have been traditional to client/server, but increasingly, Internet/Intranet/Web functions are becoming important (e.g. http, email, ftp, usenet and firewall servers).

BANDWIDTH AND BOTTLENECK AVOIDANCE

Chief among issues which must be addressed in the evolution of the LAN are balancing data transfer demands on the network, in other words, how to size the 'Network Plumbing'. Managing the growth of bandwidth demands is complicated by the fact that is quite difficult to predict future need based on historical data. The introduction of http has further exacerbated the situation by accelerating the rate of growth and introducing even more extremes in use patterns. It is undisputed that bandwidth demand will continue to grow at an accelerated rate so sizing solutions based on available and affordable technology must be balanced against predictions of availability of future cheaper technology. Sub-netting and segmentation are useful in managing traffic by isolating data to only those portions of the network where they are needed. Repeaters, bridges, routers and switches can be strategically implemented to achieve the necessary grouping of traffic. Bandwidth vs. latency (response time a requestor spends waiting for a result; Wang 1997) is an important distinction for planners to keep in mind because increases in bandwidth may not necessarily have the concomitant reduction in latency that users require.

A Redundant Array of Inexpensive Disks (RAID) is an important component of the data storage and transfer equation providing centralized, fast, fault tolerant disk space. RAID has been used to provide access to large amounts of data storage arranged on multiple physical devices. Large disks by themselves are not necessarily good solutions for providing access to large filesystems or databases. This is because the bandwidth available to an individual disk is not usually proportional to the size of the disk. Therefore, providing more, smaller component disks, with data striped across them and increased aggregate bandwidth is what RAID can accomplish. RAID devices have been largely responsible for the rapid adoption of the Fiber Channel Arbitrated Loop *(http://www.fiberchannel.com)* interconnect, which is replacing SCSI for data intensive peripherals and is poised as a future network transport technology. Data redundancy and fault tolerance are also accomplished through RAID via mirroring [RAID level 1] or parity calculations [RAID level 5]. Software and hardware RAID solutions can be implemented alone or hybridized.

BACKUP AND ARCHIVAL

Data backup and archival are central to data management in the ecological research environment. The multitude of tape drive formats available presents a confusing array of choices including 8mm, DLT, DAT, and QIC. The highest volume/speed backup devices are now being built around DLT drives and 8mm drives to a lesser extent.

Table 1. Popular tape formats, 1997.

Format	Capacity	Transfer Rate
QIC/Travan	<1GB - 8GB	£ 1MB/s
8mm	2GB - 125GB	1 - 6MB/s
DAT (4mm)	4GB - 24 GB	<1-3MB/s
DLT	10 - 100GB	1.5 - 10 MB/s

Future tape library migration path is at least as important a consideration as the technical attributes of an individual format. Robotic tape library devices are necessary for unattended backup of large data repositories. Care should be taken with the location of the tape library including plans for redundant storage, to ensure disaster recovery. Tape servers should be in secured locations and should be stable, competent computers.

Archival means different things in various environments. The exponential growth in data volume has further blurred the distinction between backup and archival. In particular, the shift from predominantly character data to object data, especially images, has largely been driving the massive increase in total volume. Images are large and in many cases are prime candidates for migration to archival media. The access to archive media can be provided via on-line, near-line or off-line solutions, using jukebox technologies and hierarchical storage software, or manual management.

OTHER CONSIDERATIONS AND REMOTE SITE SPECIFICS

Printers, plotters, film recorders, etc. should be shared among workgroups and the incentive to do this is both budgetary and practical, i.e. to simplify administration. Providing the printer with a network interface and queuing it from a capable print server will allow for relatively flawless printer access. High quality printing capabilities (high resolution, PostScript, etc.) can then be costed across many staff members and projects.

Fault tolerance and redundancy are qualities that should be prioritized relative to the institutional dependence on the equipment or service. For example, if a fileserver holds all research data, the availability of those data is critical to daily operation. Steps should be taken to ensure a level of fault tolerance that can be justified in the budget and in relation to all other priorities. These two attributes, fault tolerance and redundancy, are essentially two sides of the same coin from the hardware perspective. They can take the form of dual power supplies, mirrored disk systems, fail-over network paths, etc. Service contracts and spare-parts agreements can ameliorate the cost of redundancy. The relative merit of these approaches is highly site-specific and cost/benefit analyses are quite difficult. The need for UPS protection can not be overstated, especially in sites prone to brown power. Surge suppression on network cabling should not be overlooked either, as lightning can wreak havoc on twisted-pair or any non-inert media.

Table 2. LAN reliability needs assessment (MTBF = mean time before failure).

Reliability Value = Cost of Downtime x System MTBF x Site Risk Probability
Cost of Downtime = (System Time Value X Mean Time To Repair) + Cost to Repair

In the realm of field computers, it can be categorically stated that traditional laptops are ill-suited to dirty field work. There are many varieties of Electronic Data Recorders or ruggedized handheld PC's which can be exposed to water, dirt, etc. The cost of these units may be higher than a standard laptop, but the life expectancy in extreme conditions can be significantly longer. Inexpensive palm top computers can serve some purposes, but are notably fragile and have cramped keyboards.

Because of the remote location of many field stations, other logistic and budgetary considerations contribute to hardware specification and network design. The cost of telecommunications connectivity, especially for leased digital services, are distance-sensitive. This means that there may be strong motivation to merge voice and data over the same line, a T1 for example, in order to leverage the investment for multiple functions. The investment in multiplexing/channel bank hardware to do this can often be recouped in a year or less. Expect radical changes in this arena over the next couple of years.

Budgetary limits are usually a major implementation constraint in research. Fractured budget sources and authority can further exacerbate the inability to execute information technology strategy and evolution. The cost of computing is often not fully integrated into planning, proposal development and budgeting in research projects or institutions. It is critical to bear in mind the total cost of ownership of technology, and not just the acquisition cost of hardware and software. Hardware and software management, system administration, user support, design planning, and alternatives analysis are all personnel costs which must be factored in the total cost of ownership equation. Balancing these demands for resources against the cost of research and operations is a challenge, but the necessity of information technology is now indisputable. Computer hardware provides technological underpinnings for virtually every aspect of research, communication and publishing, as we currently know it.

LITERATURE CITED

Internet Engineering Task Force. 1998. *http://www.ietf.org*

Microsoft, Inc. 1996. Microsoft Windows NT Server Resource Kit. Microsoft Press, Redmond, WA.

Ryan, H.W., Sargent, S.R., Boudreau, T.M., Arvantis, Y.S., Taylor, S.J., Mindrum, C. 1998. Practical guide to client/server computing. Auerbach, Boston, MA.

Stone, J.P. 1997. Handbook of local area networks. Auerbach, Boston, MA.

Wong, B.L. 1997. Configuration and capacity planning for Solaris servers. Sun Microsystems Press, Mountain View, CA.

TECHNOLOGICAL UNDERPINNINGS: COMMUNICATIONS AND NETWORKING

Rudolf Nottrott

National Center for Ecological Analysis and Synthesis, 735 State St., Suite 300,

Santa Barbara, CA 93101

Abstract. At most biological field stations there are few formal provisions for on-going data exchange after individual investigators have returned to their home institutions. However, opportunities for such long distance collaboration and information exchange have recently increased with the development of wide area network technology. Wide area networks have the potential to change the culture of collaborative research in ecology. To familiarize ecologists working at biological field stations with the mechanics of Internet communication and data exchange, this chapter provides a brief review of the history of electronic networking; the architecture, protocols and common client/server applications of the Internet; and basic network security issues.

INTRODUCTION

Ecologists have recognized an increasing need for long-distance collaboration, rapid communication, and increased data access. Biological field stations, which may host hundreds of scientists and research studies over the course of decades, have a clear need for data archiving and access to those data by geographically and temporally dispersed researchers. Once biological field stations and other research institutions have established standards for metadata, protocols, and network software, archiving of long-term data (e.g., relating to the history of a site) for efficient retrieval will be possible. In this chapter, key aspects of network infrastructure such as network functions, hardware, client/server mechanisms, and security are reviewed.

HISTORY

1989 and before - a jungle of networks

Until 1989, there were few opportunities for ecologists to utilize wide area networks for data exchange or collaborative research. The myriad of incompatible networks with variable longevity (Frey and Adams 1989) made it impractical for scientists, both nationally and internationally, to use this technology for information exchange. However, once the potential significance of wide area networks for data exchange between geographically dispersed researchers was recognized, demand for enhanced network functions and increased network access ensued.

A 1989 survey of 18 Long-Term Ecological Research (*www.lternet.edu*) sites (with more than 500 widely-dispersed researchers) indicated three primary needs related to increased network capabilities (Brunt et al. 1990):

- **Local Area Networks** (LANs: Ethernet, Appletalk, PC Networks, etc.) - Resource sharing: files, programs, printers; high-speed links to higher level networks
- **Institutional Networks** (including campus networks) - Access to mainframe computers, Wide Area Networks, and files and printers on other LANs
- **Wide Area Networks** (WANs) - Instantaneous and reliable email; access to supercomputers and national information and software repositories; rapid long-distance transfer of data and information (e.g., graphics and other binary information); electronic infrastructure for long-distance collaboration.

THE INTERNET

Beginning as early as 1969, but accelerating exponentially around 1989 (**Rutkowski 1997, Leiner et al.1997**), the growth of the Internet, and its current position as a *de facto* global standard, has now made it feasible for widely distributed researchers to utilize wide area technologies. The key elements for the Internet's success were the openness and expandability of the Internet protocols, and their scalability from Local Area Networks (LAN) to global Wide

Area Networks (WAN). Although Internet access is not yet truly global, it currently is widely available in the U.S., Europe, Japan, Southwestern Australia, and parts of South America and South Africa (**Landweber** (*ftp://ftp.cs.wisc.edu/connectivity_table*), and Matrix Information and Directory Services (*http://www.mids.org/mapsale/world/index.html)*).

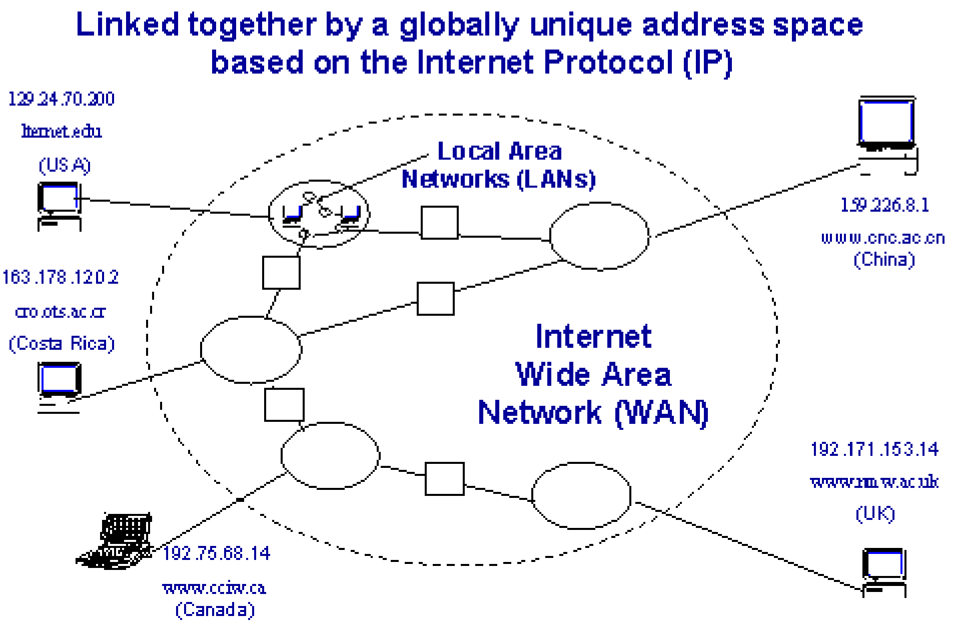
Definition of the Internet - the foundation of TCP/IP

The Federal Networking Council (Federal Networking Council 1995, *http://www.fnc.gov/*) defines the term Internet as the global information system that:

i. "is logically linked together by a globally unique address space based on the Internet Protocol (IP) or its subsequent extensions/follow-ons;
ii. is able to support communications using the Transmission Control Protocol/Internet Protocol (TCP/IP) suite or its subsequent extensions/follow-ons, and/or other IP-compatible protocols; and
iii. provides, uses or makes accessible, either publicly or privately, high level services layered on the communications and related infrastructure described herein."

The global address space of (i) is illustrated in Figure 1 which shows five computers (Internet hosts) on four continents connected via the Internet.

Figure 1. IP global address space.



Each computer is identified by a unique address, called IP number. For better readability, IP numbers are usually shown as four sets of decimal numbers separated by periods (e.g., 128.85.36.9), but they simply represent 32-bit binary numbers, allowing for 2^{32} = 4,294,967,296 computers. In practice the number is smaller, because blocks of IP numbers are reserved for various technical reasons (Hunt 1992). Also, organizations are allocated whole blocks of numbers (usually 256 or 65,536 numbers), which they can use at their discretion. Some argue that present trends indicate a leveling off in the number of host computers at approximately 38 million hosts around the year 2002 (Hilgemeier 1997)**.** However, the future internet may well have 128-bit IP numbers, to avoid the bottleneck of address shortages, and thus keep growing into the foreseeable future.

To further simplify use of the system, IP numbers are commonly represented in the form of hierarchical domain names, e.g., LTERnet.edu instead of 129.24.70.200, and Domain Name Service software is used to facilitate the conversion to numeric IP addresses. The basis of the IP protocol is that all information sent over the network is in small packets (e.g., 1000 characters) complete with destination and sender IP numbers plus other data (e.g., sequence number) needed during and on delivery. The packets may arrive at their destination in arbitrary order, but software on the destination computer can put them back together as needed, using the sequence number. Imagine a colleague in Australia sending you a 300-page story in sequentially numbered, one-page letters, at a rate of one per day. After nearly a year, you compile them to get the full story. Fortunately, the Net is faster than that.

Layers and protocol stacks

Network architects conceptualize, design and implement their network software in what they call "layers." For the Internet, the layers represent protocols including IP and TCP/IP. In schematic diagrams the software layers resemble stacks of bricks. Hence, they are often called protocol stacks. Before protocol stacks came preinstalled with most computers, one would have to install them before attempting a connection to the Internet (the Trumpet Winsock stack is a well-known example for Windows® 3.1). Figure 2 illustrates the Internet Network layers and how they relate to the Internet definition described above (this is a special case of the ISO/OSI reference model as detailed in Hunt 1992).

Figure 3 illustrates the same layers with an e-mail handling program at the top application layer. The Internet Layer, IP, corresponds to (i), the Host-to-Host Transport Layer, TCP/IP, corresponds to (ii), and the Application Layer, Telnet, SMTP, FTP, HTTP, correspond to (iii). Most network users work at the level of the "high-level services" of the Application Layer, such as Telnet, FTP or HTTP (through Web browsers). By examining some examples of high-level applications, and considering how layers pass data back and forth, the mechanics of the Internet become clearer.

Clients and servers - present-day workhorses of the Internet

The example of e-mail delivery in the previous section illustrates another concept that has found widespread use in software architecture for network-distributed applications - client/server architecture. A client is generally a software program that requests a "service" from another program called server. In the example of e-mail delivery, the client might be a program such as Eudora®, pine, Microsoft Outlook® or the original Unix program called 'mail', all available for many different kinds of operating systems and computer platforms. It's server counterpart has historically been a program called 'sendmail' running on Unix machines (alternative mail servers are now available). The client and the server communicate using standard keywords and formats, which are called a "protocol." In the case of e-mail delivery the protocol is called Simple Mail Transfer Protocol (SMTP). Mail delivery using this scheme is not unlike a Telnet session

Figure 2. Internet network layers.

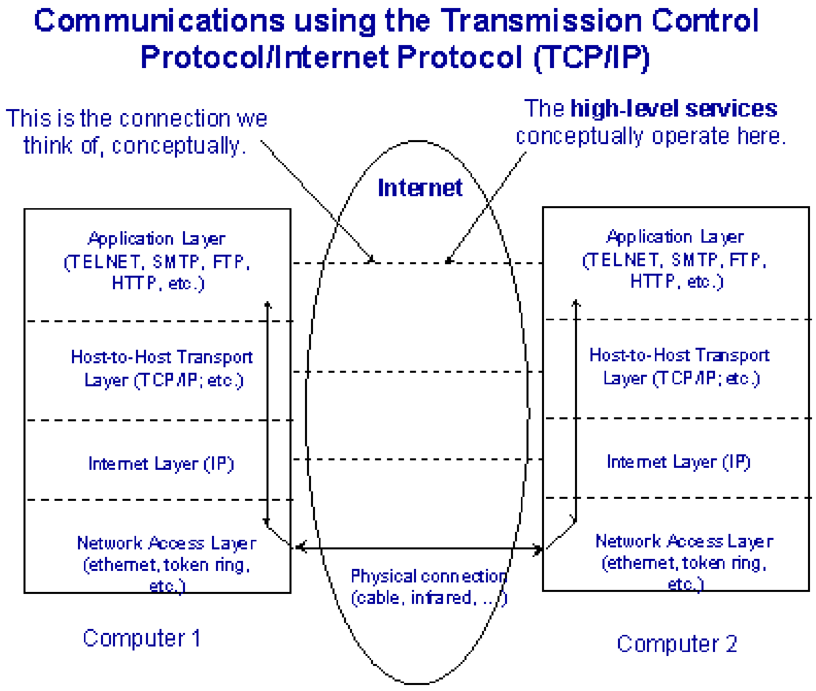
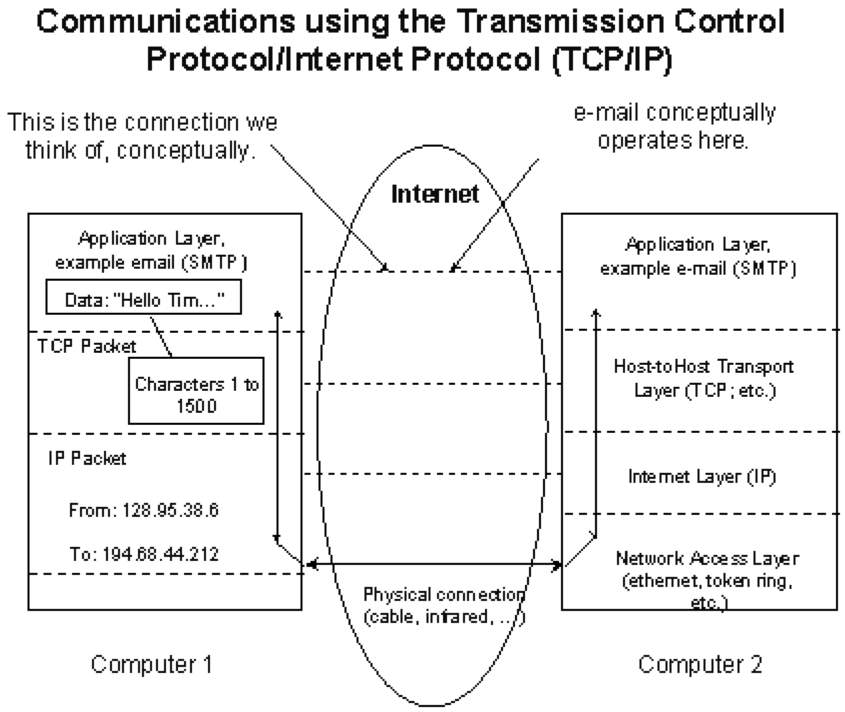

Communications using the Transmission Control Protocol/Internet Protocol (TCP/IP)
This is the connection we think of, conceptually.
The high-level services conceptually operate here.
Internet
Application Layer (TELNET, SMTP, FTP, HTTP, etc.)
Host-to-Host Transport Layer (TCP/IP; etc.)
Internet Layer (IP)
Network Access Layer (ethernet, token ring, etc.)
Application Layer (TELNET, SMTP, FTP, HTTP, etc.)
Host-to-Host Transport Layer (TCP/IP; etc.)
Internet Layer (IP)
Network Access Layer (ethernet, token ring, etc.)
Physical connection (cable, infrared, ...)
Computer 1
Computer 2

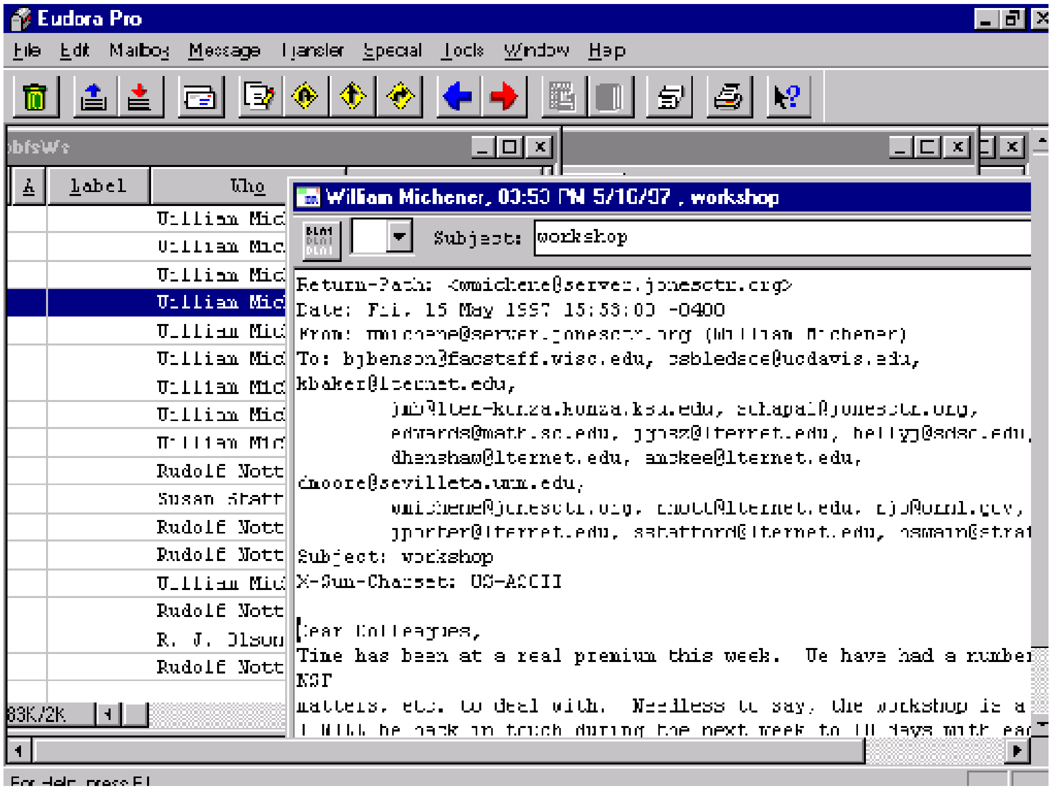
Figure 3. Internet network layers with an e-mail handling program.


Communications using the Transmission Control Protocol/Internet Protocol (TCP/IP)
This is the connection we think of, conceptually.
e-mail conceptually operates here.
Internet
Application Layer, example email (SMTP)
Data: "Hello Tim..."
TCP Packet
Characters 1 to 1500
IP Packet
From: 128.95.38.6
To: 194.68.44.212
Application Layer, example e-mail (SMTP)
Host-to-Host Transport Layer (TCP; etc.)
Internet Layer (IP)
Network Access Layer (ethernet, token ring, etc.)
Physical connection (cable, infrared, ...)
Computer 1
Computer 2

(although in practice a software mechanism called "sockets" is used, with Winsock and BSD Unix sockets being most common), and it is indeed possible to 'talk' by telnet directly to e-mail servers, as well as many other servers (such as Web, WAIS and News servers).

Naturally, people have come to expect much more user-friendliness, and consequently modern clients hide the protocol exchange behind a façade of windows and menus, as Figure 4 shows for the Eudora® mailer client. With the widespread use of desktop workstations, clients and servers for different application areas are becoming increasingly common. Table 1 gives an overview of the most common types of application protocols. Most Internet protocols are described in Request for Comments (RFC *http://ds.internic.net/rfc/*). A very comprehensive list of Winsock clients, together with reviews and source links, can be found at *http://cws.internet.com*.

Figure 4. Eudora windows.



```
Return-Path: <wmichene@server.jonesctr.org>
Date: Fri, 15 May 1997 15:53:03 -0400
From: wmichene@server.jonesctr.org (William Michener)
To: bjbenson@facstaff.wisc.edu, csbledsoe@ucdavis.edu,
kbaker@lternet.edu,
        jmb@lter-konza.konza.ksu.edu, schapal@jonesctr.org,
        edwards@math.sc.edu, jgosz@lternet.edu, hellyj@sdsc.edu,
        dhenshaw@lternet.edu, amckee@lternet.edu,
dmoore@sevilleta.unm.edu,
        wmichene@jonesctr.org, rnott@lternet.edu, cjo@ornl.gov,
        jporter@lternet.edu, sstafford@lternet.edu, osmain@strat
Subject: workshop
X-Sun-Charset: US-ASCII

Dear Colleagues,
Time has been at a real premium this week.  We have had a number
NSF
matters, etc. to deal with.  Needless to say, the workshop is a
I will be back in touch during the next week to 10 days with eac
```

NETWORK SECURITY

Security of data and other information on wide area networks is a key concern among scientists (Brunt et al. 1996). However, in most circumstances, solutions are available to ensure security of data and information, within reasonable limits. Most of the original Internet applications were developed by engineers and academics with little need for security, and only recently have commercial applications such as bank transactions and online sales necessitated development of extremely secure network applications (including features such as encryption). With the growth of the number of Internet hosts to tens of millions, the atmosphere has changed from that resembling a small town, where few residents lock their doors, to that of a big city, where some doors may need dead-bolts and chains. It is important to keep in mind that network growth is good, and security consciousness is a small price to pay for the increased services that have come with network growth.

Table 1. The most common client/server pairs at the Internet application layer

Application Layer Protocol	**Common Name**	**Example clients**	**Example servers**
SMTP, Simple Mail Transfer Protocol	e-mail delivery	Eudora®; MS Outlook®; Pegasus®; mail (built-in on most Unix systems)	Sendmail, built-in on most Unix systems

POP, Post Office Protocol	e-mail pickup box	Eudora®; MS Outlook®; Pegasus®	POP3
FTP, File Transfer Protocol	Ftp	ws_ftp; built-in on Unix systems	ftpd - built-in on most Unix systems
NNTP	News	Free Agent®;	Nntpd
TELNET	Telnet	Ewan; Built-in on most Unix systems	Telnetd - built-in on most Unix systems
HTTP	WWW, Web	Netscape Navigator®; MS Internet Explorer®	Httpd from NCSA
	Video conferencing	CU-SeeMe®	Reflector
Gopher	Mostly superseded by the Web		
ODBC, Open DataBase Connectivity	ODBC database Access	MS Access®; Excel®	Oracle®, Ingres®, MS SQL server

Sources for security information

The two most widely used operating systems with built-in TCP/IP capabilities are Unix and Windows NT (Windows 95® was developed with much lesser network capabilities, mostly for use in proprietary LANs). Because Unix is a much older and more mature operating system, several organizations (e.g., Computer Emergency Response Team, *http://www.cert.org* and the Internet Society, *http://www.isoc.org*) have considerable experience with Unix security issues, and a wealth of literature is available on Unix system security.

Recently, numerous books dealing with NT security issues have become available. An easy way to find the latest books is to do an online search at one of the electronic bookstores. For example, a search at *http://www.amazon.com* for "windows and security" retrieved 9 items, including Rutstein (1997) and Dalton et al. (1997). Similarly, a search for "unix and security" returned 7 items including Garfinkel and Spafford (1996). In addition, several newsgroups have been established to discuss Unix-related security issues (comp.security.unix), NT (comp.os.ms-windows.nt.admin.security), and miscellaneous other security issues (all in comp.security).

Simple precautions

Simple measures can help prevent most security problems:

- Choosing safe passwords is the simplest part of network security. Bad passwords cause >80% of all security problems (RFC 1244 1991); change your password at regular intervals!
- Perform the following activities on a regular basis (or choose a good systems administrator who will do it for you): use password software to make sure passwords are "good"; perform password aging; get your software from trusted sources only; keep software updated.

Keep an eye on your system (unusual files, processes, login activity, etc.).
- Possibly limit access (allow logins only for certain machines, or domains).
- Use programs that help with security checks (e.g., COPS).
- Consider the use of encryption.

Gated Communities - firewalls and intranets

With the exponential expansion of the Internet, some organizations with strict security needs have partially separated themselves from the Internet. They have done this by using a TCP/IP-based LAN internally, called an Intranet, which is connected to the Internet through a separate machine running "firewall" software. Firewalls can effectively protect your institutional network from the outside world and still allow your users access to the Internet. Firewalls obscure the internals of your Intranet from the outside world by refusing to provide name or address information about internal machines, by replacing internal users' login names with aliases (for email), by allowing FTP and other services only to/from the firewall and by allowing telnet or remote log-ins only to/from the firewall.

LITERATURE CITED

Brunt, J., J. Porter, R. Nottrott. 1990. Internet connectivity in the Long-Term Ecological Research Network (LTER): assessment and recommendations. LTER Network Office, University of Washington. Seattle, WA.

Dalton, W., S. Fuller, B. Kolosky, J. Millecan, Nachenberg, C. Goggans. 1997. Windows NT server 4: security, troubleshooting, and optimization. New Riders Publishing

Frey, D., and R. Adams. 1989. !%@:: a directory of electronic mail. O'Reilly & Associates, Cambridge, MA.

Garfinkel, S., and G. Spafford. 1996. Practical UNIX & internet security, 2nd edition. O'Reilly & Associates, Cambridge, MA.

Hilgemeier, M. 1997. Internet growth - host count turning point in June 1997. *http://www.is-bremen.de/~mhi/inetgrow.htm*

Hunt, C. 1992. TCP/IP network administration. O'Reilly & Associates, Cambridge, MA.

Leiner, B., V.G. Cerf, D. Clark, R. Kahn, L. Kleinrock, D. Lynch, J. Postel, S. Roberts, S.Wolff. 1997. Brief history of the Internet. Internet Society. *http://www.isoc.org/internet/history/brief.html*

Porter J., R. Nottrott, D. Richardson. 1996. Ecological databases: new tools and technologies. In Long-Term Ecological Research. Ecological Society of America annual meeting, Providence, RI.

Porter J., K. Baker, R. Nottrott. 1996. Tools for managing ecological data. Eco-Informa '96 Conference. Lake Buena Vista, FL.

RFC 1244. 1991. Request for comment. Site security handbook. *http://ds.internic.net/rfc/rfc1244.txt*

Rutkowski, A.M. 1997. Internet Trends. General Magic, Inc., Sunnyvale, CA. *http://www.genmagic.com/Internet/Trends*/; and *http://www.genmagic.com/Internet/Trends/slide-4html*

Rutstein, C.B. 1997. Windows NT security: a practical guide to securing Windows NT servers and workstations. Computing McGraw-Hill, New York, NY.

TECHNOLOGICAL UNDERPINNINGS: SOFTWARE

Karen S. Baker

Scripps Institution of Oceanography, University of California at San Diego,

La Jolla, CA 92093-0218

Abstract. Survey results from Long-Term Ecological Research (LTER) sites provide an overview of the variety of software choices made at individual locations where PC, Macintosh and UNIX platforms predominate. The survey considered software categories including bibliographic, data entry, database management systems, drawing, geographic information systems, graphics, spreadsheets, statistics and word processing. The objective of the survey was to assess the heterogeneity of software used by the LTER community.

INTRODUCTION

Decisions with respect to software selection must balance the often conflicting requirements of addressing immediate local community needs and meeting broader, long-term institutional objectives. Research institutions, including biological field stations, often have very specific needs that require further balances between simple versus complex, individual versus standard, and the current state-of-the-art versus emergent technologies. Although software ideally should be extensible and have clear export paths, there are few other specific rules to guide software infrastructure. Software choices often depend upon several factors:

- What are the computational, data management and storage priorities?
- What software options exist? Do options vary by platform?
- What are the costs of hardware, software, and support?
- Where will the different functions of data collection, management (e.g., entry, processing, archival), and analysis be carried out?
- Who is available for technical support (e.g., local support environment, consultants, vendors)?

Consideration of such questions permits a definition of priorities. Subsequently, more task-specific questions can be asked, such as whether scientific visualization tools are a high priority and how data can be accessed.

Available resources play a significant role in discussions of the organizational approach of a research group. Some sites identify and encourage use of a common set of software tools. In such cases, the availability of training can help those users not familiar with the supported tools. Other sites find a range of diverse software to be advantageous. It is important, given the variety of software available, to consider the collective consequences of software choice and to develop a policy regarding which packages will be site supported in order to maintain realistic user expectations.

Factors such as cost (including the availability of academic discounts), stability, marketing, interoperability, power and ease of use influence decisions about software. Education helps build consensus, but it is important to recognize how software acquisition is influenced by a diversity of legacies (hardware and software), interfaces (human and hardware), and data volumes (small to large). Familiarity also plays a role in the decision-making process. The LTER software survey permits sites to place their own decisions into a network-wide context, i.e., a survey extends a single site's experience to a network of sites.

LTER SOFTWARE SURVEY (1992 to 1997)

The LTER software survey began in 1992 with eighteen sites in addition to the LTER Network Office. By 1997, the survey included twenty-one participants. Results from an earlier survey have been discussed previously (Porter et al. 1996). A yearly LTER software survey of more than nine categories quantifies the diversity and trends of software within the LTER community. Table 1 summarizes bibliographic, data entry, database management systems, drawing, geographic information system (GIS), graphics, spreadsheets, statistics and word processing software products that were employed in 1997. The first line of each category in the table gives the total number of packages used throughout the LTER Network. This is followed by a list of the software packages with at least three site implementations along with the number of implementations at all the sites by platform type.

Table 1. Software and software summary for LTER sites (minimum of three installations) for 1997.

Original form is found at *http://lternet.edu/im.* The first line gives the total number of packages used by platform followed by lines with the number of sites using a specific package.

	PC	MAC	UNIX		PC	MAC	UNIX
Bibliography*	**7**	**1**	**5**	**Graphics***	**16**	**8**	**7**
Procite^{TM}	6	--	--	Excel^{TM}	8	7	--
Papyrus^{TM}	5	--	--	Deltagraph^{TM}	2	3	--
Endnote^{TM}	4	4		Lview^{TM}	5	--	--
Bibtex^{TM}	--	--	3	Matlab^{TM}	--	2	2
Data entry*	**11**	**2**	**4**	Quattropro^{TM}	8	--	--
Excel^{TM}	12	10	--	SAS-graph^{TM}	4	--	4
Quattropro^{TM}	5	--	--	Sigmaplot^{TM}	8	--	--
Lotus^{TM}	4	--	--	Slidewrite^{TM}	3	--	--
SAS^{TM}	--	--	4	Cricketgraph^{TM}	--	4	--
Quickbasic^{TM}	3	--	--	Framemaker^{TM}	--	--	3
Database*	**10**	**0**	**11**	**Spreadsheets***	**6**	**3**	**1**
Access^{TM}	8	--	--	Excel^{TM}	20	10	--
Dbase^{TM}	5	--	--	Lotus^{TM}	10	--	--
SQLserver^{TM}	3	--	--	Quattropro^{TM}	15	--	--
ArcInfo^{TM}	--	--	4	**Statistics***	**11**	**9**	**10**
Ingres^{TM}	--	--	4	Excel^{TM}	16	--	--
Oracle^{TM}	--	--	4	Matlab^{TM}	1	--	2
Msql^{TM}	--	--	3	SAS^{TM}	12	--	10
Drawing*	**8**	**8**	**3**	Sigmaplot^{TM}	7	--	--
Photoshop^{TM}	7	6	--	Splus^{TM}	3	--	6

Freehand^{TM}	2	1	--	Systat^{TM}	10	3	1
Coreldraw^{TM}	3	--	1	Cricketgraph^{TM}	--	4	--
GIS*	**9**	**1**	**10**	Powerpoint^{TM}	2	2	--
ArcInfo^{TM}	13	7	18	Statview^{TM}	--	3	--
ArcView^{TM}	12	--	13	SPSS^{TM}	1	--	2
Erdas^{TM}	6	--	9	**Wordprocessors***	**5**	**4**	**4**
Erdas-imagine^{TM}	4	--	5	Framemaker^{TM}	3	2	4
Idrisi^{TM}	6	--	2	Word^{TM}	18	9	--
Grass^{TM}	1	--	7	WordPerfect^{TM}	17	2	4
				LaTex^{TM}	--	--	5

Software categories with the highest diversity (number of different packages) across all platforms within the LTER Network included graphics and statistics, whereas the lowest diversity was associated with drawing, wordprocessing and spreadsheet software (Figure 1). Given the differences in hardware, it is interesting to consider distributions by platform type. In general, there were more packages used by the PC than the Mac, except in the case of drawing software. In 1997, the largest variety of packages was related to graphics on the PC. The greatest variety of packages for the UNIX system was associated with database, statistics and GIS categories. Examination of temporal trends from 1992 to 1997 (Figure 2) reveals that the software diversity for some categories increased (e.g., graphics, statistics and drawing) while the diversity for other categories remained relatively unchanged (e.g., word processors and spreadsheets). While UNIX was the platform of choice for most database applications in earlier years, there has been an increase recently in GIS and database software use on both PC and Unix platforms.

Figure 1. The number of different software packages used by LTER sites in 1997 for each platform type.

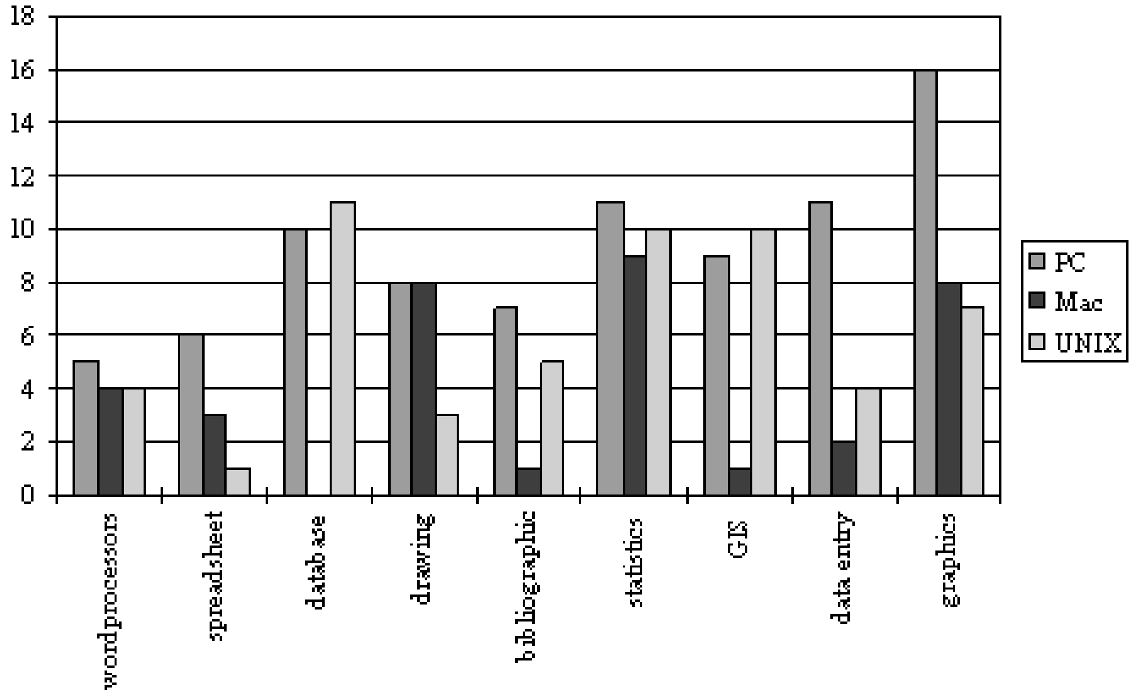
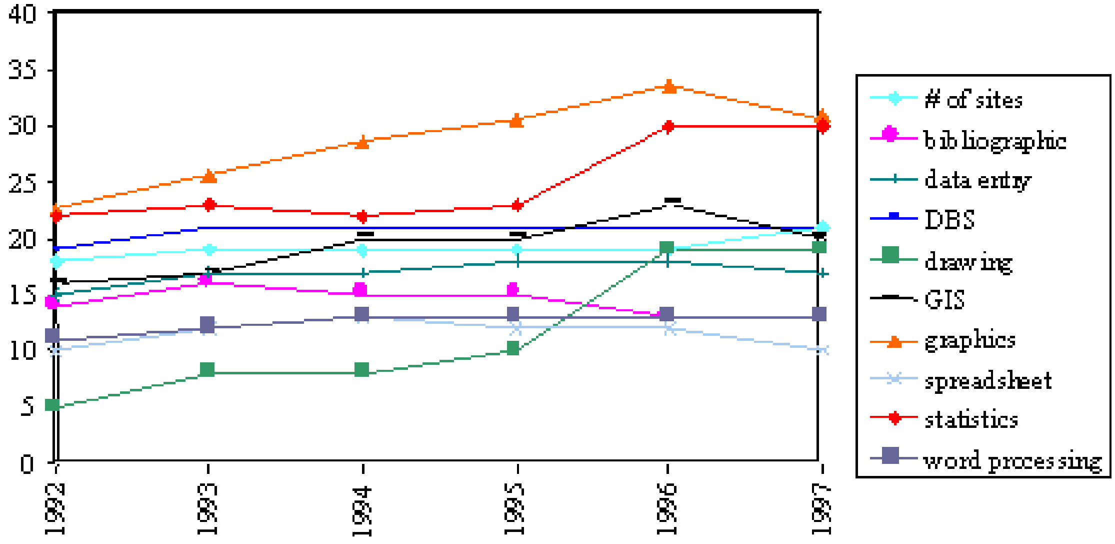


Figure 2. The total number of different software packages (by category) used by LTER sites during 1992-1997.



CONCLUSION

Several issues have and will continue to influence the software infrastructure at long-term environmental and ecological research sites, including scientific objectives, software policies, cost, and standardization. Given the diversity of software available, a general survey facilitates consideration of a wide range of potential solutions. The interplay of objectives and approach is unique to each research site, so decisions with respect to software vary.

ACKNOWLEDGEMENTS

This work was funded by NSF Grant OPP90-11927. This is Palmer LTER Contribution #151.

LITERATURE CITED

Porter, J.H., R.W. Nottrott, and K.S. Baker. 1996. Tools for managing ecological data. Proceedings of the Eco-Informa Workshop, Global Networks for Environmental Information, November 4-7 1996, Lake Buena Vista, FL, 11:87-92.

DATA ENTRY

John M. Briggs

Kansas State University, Division of Biology/Ackert Hall, Manhattan, KS 665026-4901

Barbara J. Benson

University of Wisconsin-Madison, Center for Limnology, 680 N. Park Street,

Madison, WI 53706

Mike Hartman

University of Colorado, INSTAAR, Campus Box 450, Boulder, CO 80309-0450

Rick Ingersoll

Cornell University, Biometrics Unit, 441 Warren Hall, Ithaca, NY 14853

Abstract. This chapter summarizes the conversion of field- or laboratory-collected data into an electronic form. Techniques to make this conversion as quick and error-free as possible are illustrated, including descriptions of data entry software. Finally, the use of field data recorders and newer technological advances such as optical character recognition for data entry are discussed.

INTRODUCTION

One of the many issues that an information management specialist must consider is the need to convert data into a useable electronic format. This frequently means converting data collected in the field, usually on paper, into an electronic form that can then be used in a statistical or graphical package by the researcher. The purpose of this paper is to present guidelines that we have found useful in making this conversion as quick and error-free as possible.

PLANNING

The first and most critical aspect of data entry is planning. If at all possible, the information manager should be involved in the development of data entry strategies. Much like a statistician is consulted prior to a proper experimental design, an information manager can assist scientists in designing the means of data capture (recorded on paper in the field or laboratory). Guidelines for designing the forms on which data are recorded include:

- Field and laboratory data collection forms should facilitate data collection.
- On-screen forms should facilitate data entry in the computer.
- Field and laboratory forms should be as similar as possible to on-screen forms.
- Include fields for initials of data collection personnel and date.
- Allow room for qualifying comments and metadata.

On-screen data entry forms should incorporate the following features:

- Forms should be easy-to-read and arranged to facilitate data entry.
- Use color only to improve readability. Use of color can be effective, but should not be overused. Excessive use of color can generate eyestrain. In addition, color schemes may not "translate" if other systems don't support the same palette of color choices.

Use automatic duplication, manual duplication, default values, and other keystroke-saving methods to speed entry and reduce tedious aspects of data entry. Data entry is a tedious job; anything you can do to speed it up and to reduce keystrokes is useful.

- Use quality-control features such as range checks, internal and external table lookup, as well as re-key verification. If properly established and implemented at the time of entry, these steps can greatly reduce the number of errors introduced into the data.

SOFTWARE TOOLS

Historically, data entry tools associated with mainframe computers were limited to data punch or Teletype machines. However, numerous options were introduced with the advent of personal computers. The tools we describe by no means comprise a comprehensive list, but have been found to be useful.

Spreadsheets are probably the most common software tools used to enter data. Their generic interface of rows and columns is familiar to most scientists and, with a little modification, can be quite powerful. For example, the North Temperate Lake LTER site uses EXCEL^{TM} spreadsheets which have been customized to make date entry easier and incorporate error checks. To protect the data entry template from modification, the menu bar has been simplified to permit only a limited set of spreadsheet operations. *In addition, cells that define the form (and should not be changed) are locked*. There is a considerable amount of error checking built into the data entry sheet through formulas and look-up tables. After data entry, the data entry staff scrolls down the spreadsheet to an area that shows a duplicate of the entry area but with errors marked. For example, categories of error checks for the North Temperate Lake LTER fish data include: range check, checks of spelling of character-valued parameters, and comparison of length and weight.

For relatively small data sets, or for sites that can afford to hire people to do double entry of data sets and/or for research sites that use SAS^{TM} for their data analysis, there is an excellent application called SAS^{TM} DUALDATA. It is available at *www.npsc.nbs.gov/resource/tools/software/dualdata/dualdata.htm*. If you are going to use SAS^{TM} as your analysis tool, this application will automatically create a SAS^{TM} dataset for you. It uses double-entry techniques (i.e., you enter the data twice) for validating data entry. The user defines the variables, enters the values comprising the data set, and then reenters to validate. During validation, each value entered is compared against the corresponding value from the initial data entry. If discrepancies occur, the field is flagged and the user is prompted to enter the correct value. The application keeps track of validated observations so that the validation process can span multiple data entry sessions. The application also allows for the possibility that observations may be omitted or duplicated.

A powerful commercial package designed solely for data entry is EasyEntry^{TM} (P.O. Box 2464, Chapel Hill, NC 27515-2464; Phone 919-933-3113; Fax 919-968-1350; Toll free 1-800-532-7573; Email: info@easyentry.com; Web: *http://www.easyentry.com*).

EasyEntry^{TM} is easy to learn, thereby reducing training time and allowing for rapid data entry (minimizing of keystrokes). Numerous quality control features are included:

- Full screen design and modification
- Data field specifications
- Field validation
- Entry and modification
- Keypunch emulation

Under the data field validation, this package allows for:

- range tests
- validity checks
- internal and file table lookup

selective and full re-key verification

- error messages-standard as well as user-defined messages

EasyEntry^{TM} interfaces with SAS^{TM}, Oracle^{TM}, Rdb^{TM}, Informix^{TM}, and other software packages, thus allowing the data to be ported into almost any data analysis package. EasyEntry^{TM} operates on a variety of platforms including: AS/400^{TM}, Unix^{TM} [IBM^{TM}, DEC^{TM}, HP^{TM}, SUN^{TM}, SGI^{TM}], Windows^{TM} (XVT), OS/2^{TM}, MS-Windows^{TM}, X-Windows^{TM} and Mac^{TM}. Thus, it is truly hardware-independent. Future plans for EasyEntry^{TM} include interfaces to new data input devices such as scanners, optical character recognition (OCR), and barcode devices.

Another option is for users to write custom programs. These can range from customizing spreadsheets (as described above) to computer-language specific (or proprietary) data entry. Custom programs are particularly well suited to long-term research for which data collection and entry protocols undergo little change. Nonetheless, the "hidden costs" associated with development and maintenance of custom data entry programs should not be overlooked. For example, evolution of modifications to data collection, programmer turnover and inadequate documentation, as well as the rapid evolution of computer technology can result in high maintenance costs. For these reasons, Konza Prairie LTER is relying less on custom data entry approaches than they have in the past and more on commercially available data entry packages.

Field data recorders, commonly used to make meteorological and hydrological measurements often capture data electronically and do not have to be manually entered. These are very common and useful tools. However, the fact that data are collected electronically does not imply that those data are accurate. Where appropriate, tools used to ensure reliability of manually entered data, such as field range checks, should be employed for electronic data collection as well. In addition, electronically collected data are excellent candidates for many of the approaches Edwards (this volume) has advocated.

FUTURE DIRECTIONS

Future technological advances such as optical character recognition (OCR), voice recognition, electronic "notebooks" and electronic writing devices will reduce the need for manual data entry. The decrease in hardware size and cost associated with an increase in computational power should augment this trend. However, as long as ecologists must collect their data under "field" conditions, there will always be a need for at least some manual data entry.

DATA QUALITY CONTROL / QUALITY ASSURANCE

Don Edwards

Department of Statistics, University of South Carolina, Columbia, SC 29208

Abstract. Some basic concepts and strategies for data quality are discussed, specifically: management philosophies; outlier detection for the purpose of elimination of data contamination; keypunch errors; illegal data filter programs; detection of outliers in samples; and detection of outliers and leverage points in simple linear regression.

INTRODUCTION: PREVENTION FIRST

The importance of data quality assurance strategies to long-term ecological research cannot be understated, yet the topic receives surprisingly little attention in the scientific literature. In the short space allotted here, little can be done to comprehensively alleviate this lack of guidance, so one particular issue will be focused on which is highly statistical in nature: the detection of "outliers" in data, as an intermediate step in the elimination of contamination. Before beginning that discussion, though, it must be emphasized that this particular issue is not the most important one to data quality. It is, however, one that has been abused, and one, which this author is qualified to discuss.

Prevention of data contamination is clearly preferable to after-the-fact heroics, but prevention issues are largely management issues. American industry learned the prevention lesson the hard way in the 1960's and 70's, when advancements in quality science in Japan erased American worldwide dominance in the electronics and automobile industries. Ironically, Americans Joseph Juran and W. Edwards Deming, sent to Japan after World War II to help reconstruction, played huge roles in the Japanese coup. As for the relative importance of prevention, no one has expressed it more succinctly than the ever-acidic Deming: "Let's make toast the American industry way - you burn, I'll scrape."

Many management strategies for data quality assurance in scientific settings could be borrowed from industrial quality science. For example, Flournoy and Hearne (1990), in a cancer research center, stress the importance in a multi-user database setting that *all* users and data contributors have a stake in data quality. In fact, this is also one of Deming's (1986) foundational principles: all company employees, from upper level management (i.e., principal investigators) to line workers (i.e., data entry technicians), must feel a responsibility for, and a pride in, product (i.e., database) quality. Of course, the real challenge lies in inspiring this universal motivation. Along these lines, another surprising Deming principle is that no worker should ever be penalized for poor quality, as poor quality is usually the result of a poorly designed manufacturing (i.e., data collection) process; punishment is unfair and destroys worker-management (i.e., technician-scientist) trust. A successful organizational structure promoted by Deming, which could be adopted immediately for database quality assurance, is the use of "quality circles": these would be regular (e.g., weekly) meetings of scientists, field technicians, systems specialists, and data entry personnel for the purpose of discussing data quality problems and issues. These brief regular meetings build teamwork-attitudes while focusing brain power on data quality issues; participants become constantly aware of quality issues and learn to anticipate problems. Not surprisingly, some of the best ideas come from the lowest-ranking members of the circle!

Incidentally, another of Deming's principles is that everyone, from upper-level management to line workers, should have a basic understanding of natural variability and simple statistical methods for dealing with it. It has been said that one can stop a Japanese at random on the street, and he/she will know the meaning of "standard deviation". In America, asking that question to a random passerby is likely to result in a less desirable outcome!

OUTLIER DETECTION PHILOSOPHY

The term "outlier" is not formally defined. An outlier is simply an unusually extreme value for a variable, given the statistical model in use. What is meant by "unusually extreme" is a matter of opinion, but the operative word here is "unusual"; some extremes are to be expected in any data set. It must also be emphasized, and will be demonstrated, that the "outlier" notion is model-specific: a particular value for a variable might be highly unusual under, say, a linear regression model, but not unusual at all in a model without the regressor. So, outlier detection is part of the process of checking the statistical model assumptions, a process that should be integral to any formal data analysis.

"Elimination of outliers" should *not* be a goal of data quality assurance. Many ecological phenomena naturally produce extreme values, and to eliminate these values simply because they are extreme is tantamount to pretending that the phenomenon is "well-behaved" when it is not. To mindlessly or automatically do so is to study a phenomenon other than the one of interest. The elimination of data *contamination* is the appropriate phrasing of this data quality assurance goal. Data contamination occurs when a process or phenomenon other than the one of interest affects a variable's value. If this contamination is undetectable at observation time, it can usually only be detected if it produces an outlying value. Hence, the *detection* of outliers is an intermediate step in the elimination of contamination. Once the outlier is detected, attempts should be made to determine if some contamination is responsible. This would be a very labor-intensive, expensive step if outliers were not by definition *rare*. Note also that the investigation of outliers can in some instances be more rewarding than the analysis of the "clean" data: the discovery of penicillin, for example, was the result of a contaminated experiment. If no explanations for a severe outlier can be found, one approach is to formally analyze the data both with and without the outlier(s) and see if conclusions are qualitatively different.

DATA ENTRY ERRORS AND ILLEGAL DATA CHECKS

Sources of contamination due to data entry errors can be eliminated or greatly reduced in several ways. One excellent strategy is to have the data independently keyed by two data entry technicians, and then computer-verified for agreement. This practice is commonplace in professional data entry services, and in some service industries such as the insurance industry (Lepage 1990). Sadly, scientific budgets for data entry are usually inadequate to allow for double-keying of data, though other means of detecting keypunch errors are less effective and probably more expensive since they involve higher-paid personnel.

Illegal data are variable values or combinations of values that are literally impossible for the actual phenomenon of interest. For example, non-integer values for a count variable (e.g., the number of flowers on a plant) or values outside of the interval [0,1] for a proportion variable would be illegal values. Illegal combinations occur when natural relationships among variable values are violated, e.g., if Y_1 is the age of a banded bird in last year's census, and Y_2 is the same bird's age in this year's census, then Y_1 had better be less than Y_2. These kinds of illegal data often occur as data entry errors, but also for other reasons, e.g., misreading of gauges or miswriting of observations in the field or laboratory due to fatigue.

A simple and widely-used technique for detecting these kinds of contamination is an illegal data filter (or "rules," see Henshaw, Bierlmaier, and Hammond, this volume). This is a program which simply checks a laundry-list of variable value constraints on the master data set (or on an update to be added to the master) and creates an output data set including an entry for each violation with identifying information and a message explaining the violation. Table 1 shows the structure of such a program, written in the SAS^{TM} language (SAS 1990). The filter program can be updated and enhanced to detect new types of illegal data that may have been unanticipated early in the study. A word of caution, however: the operative word *here* is "illegal". Simply because one has never observed, say, an ozone concentration below a given threshold, and can't imagine it ever happening, does not make such an observation an illegal data point. One of the most famous data QA/QC blunders occurred when NASA computers were programmed to delete satellite observations of ozone concentrations below a specified level, and thus failed to discover the "ozone hole" over the south pole (Stolarski et al. 1986).

Table 1. An illegal-data filter, written in SAS (the data set "All" exists prior to this DATA step, containing the data to be filtered, variable names Y1, Y2, etc., and an observation identifier variable ID).

```
Data Checkum; Set All;

message=repeat(" ",39);

If Y1<0 or Y1>1 then do; message="Y1 is not on the interval [0,1]"; output; end;
If Floor(Y2) NE Y2 then do; message="Y2 is not an integer"; output; end;

If Y3>Y4 then do; message="Y3 is larger than Y4"; output; end;

:

(add as many such statements as desired...)

:

If message NE repeat(" ",39);

keep ID message;

Proc Print Data=Checkum;
```

OUTLIERS IN SAMPLES: GRUBBS' TEST

One of the oldest and most widely used procedures for detecting contamination in samples is Grubbs' test (Grubbs and Beck 1972, ASTM E 1994). By "samples" we mean that, if the data are uncontaminated, we would have several (say, n) independent observations on the variable from the same repeatable, well-defined, stable experimental process. Grubbs' test assumes that the uncontaminated process produces data which follow a Normal (or Gaussian) distribution, and it is very sensitive to that assumption; if the "clean" data are grossly non-Normally distributed, one should not use Grubbs' test. In fact, to this author's knowledge, every formal outlier detection rule / test has the serious drawback that it makes a distributional assumption and is sensitive to that assumption. This is not the case for all statistical procedures that nominally assume Normality; for example, t-tests are typically robust to this assumption.

Grubbs' test is performed as follows: let $Y_1<Y_2<...Y_n$ denote the ordered sample values, and $\bar{Y}$ and *S* the sample mean and standard deviation, respectively. If it is only of interest to detect unusually large outliers, then compare the test statistic

$$T_n = (Y_n - \bar{Y}) / S$$

to the appropriate tabled one-sided critical point (Grubbs and Beck 1972, ASTM E 1994), which depends on n and an error rate which we will call a $_G$. If it is only of interest to detect unusually small outliers, compare the test statistic

$$T_1 = (\bar{Y} - Y_1) / S$$

to the appropriate one-sided critical point. If either large or small outliers are to be detected, compare the larger of T_n and T_1 to the two-sided critical point.

The probability a $_G$ is in this case a *per-sample* error rate. So, for example, if a $_G$ is chosen to be .05, then in 5% (1 in twenty) of repeated *uncontaminated* samples of this size, we would falsely declare a contamination to exist. Users are encouraged to choose a $_G$ thoughtfully, as it has a different meaning than the "a -level" one uses in testing research hypotheses. What fraction of the clean data are you willing to lose, or at the very least investigate, for the sake of detecting possible contamination? Bear in mind that if such contamination is really severe, it would be detected using a smaller a$_G$, as well. ASTM E (1990) recommends a "low significance level, such as 1%". It should also be noted that Grubbs' test cannot be done at all for n=2, and for n=3 the critical points do not differ for choices of (two-sided) a $_G$

less than .05.

As an example of the (mis-)application of Grubbs' test, consider the seeded-cloud rainfall data of Simpson and colleagues (1975) shown in Table 2. The mean and standard deviation for these data are $\bar{Y}$ =442 and S = 651. With n = 26 and a $_G$=.01, the one-sided critical point for Grubbs' test is 3.029, and the test statistic for detecting large outliers is T_{26}=(2745.6 - 442)/651 = 3.539, hence (if being careless) we would assert contamination.

Table 2. Rainfall in acre-feet from seeded clouds (Simpson et al. 1975).

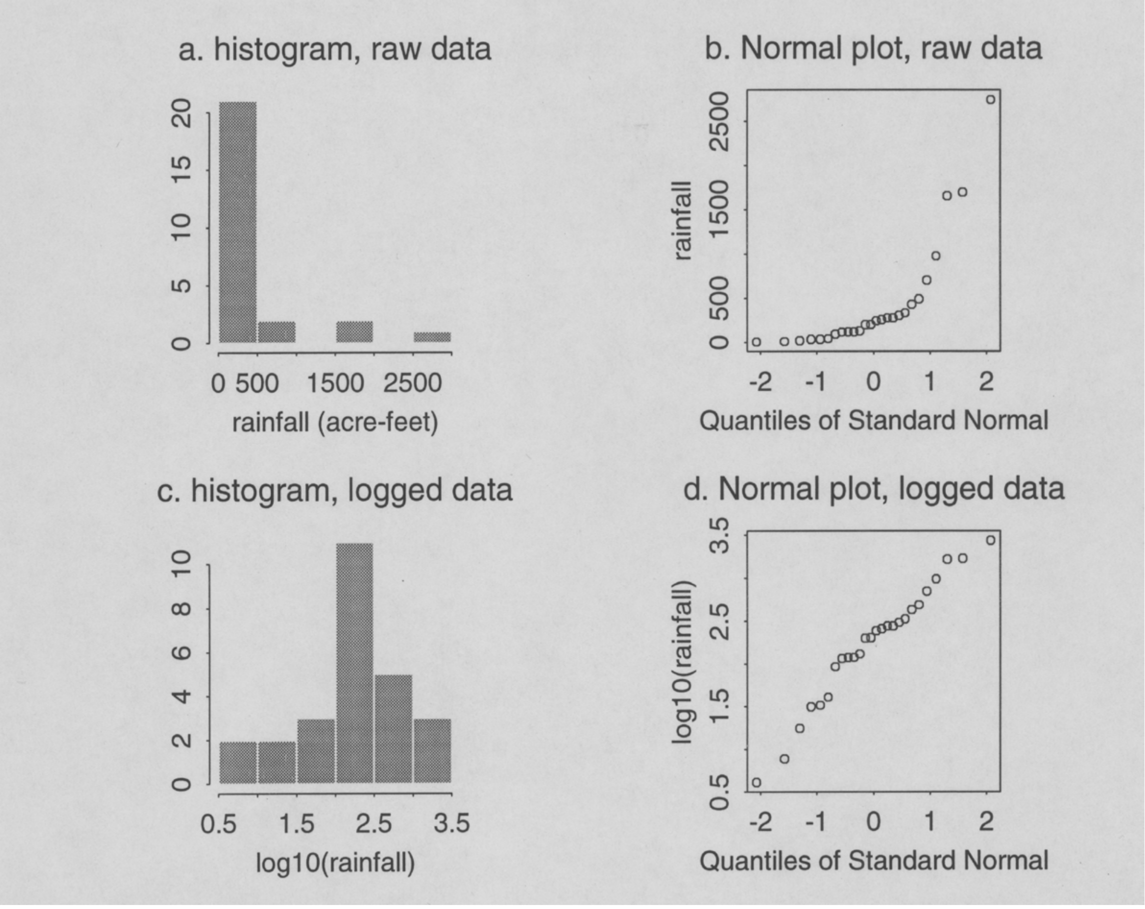
4.1 7.7 17.5 31.4 32.7 40.6 92.4 115.3 118.3

119.0 129.6 198.6 200.7 242.5 255.0 274.7 274.7 302.8

334.1 430.0 489.1 703.4 978.0 1656.0 1697.8 2745.6

Of course, the assumption that the uncontaminated sample follows a Normal distribution is grossly violated here; Figures 1a and 1b show a histogram and Normal probability plot for the raw data, which clearly show that the sample as a whole follows a severely right-skewed distribution (readers unfamiliar with Normal probability plots can find discussion of them in many modern intermediate statistics texts, e.g., Chambers et al. 1983, Sokal and Rohlf 1981). Figures 1c and 1d show a histogram and Normal plot for the $\log_{10}$-transformed rainfall data. Clearly, these rainfall data are very nearly log-Normally distributed, and there is no evidence of contamination.

Figure 1. Distributional checks of data on rainfall from seeded clouds (Simpson et al. 1975).



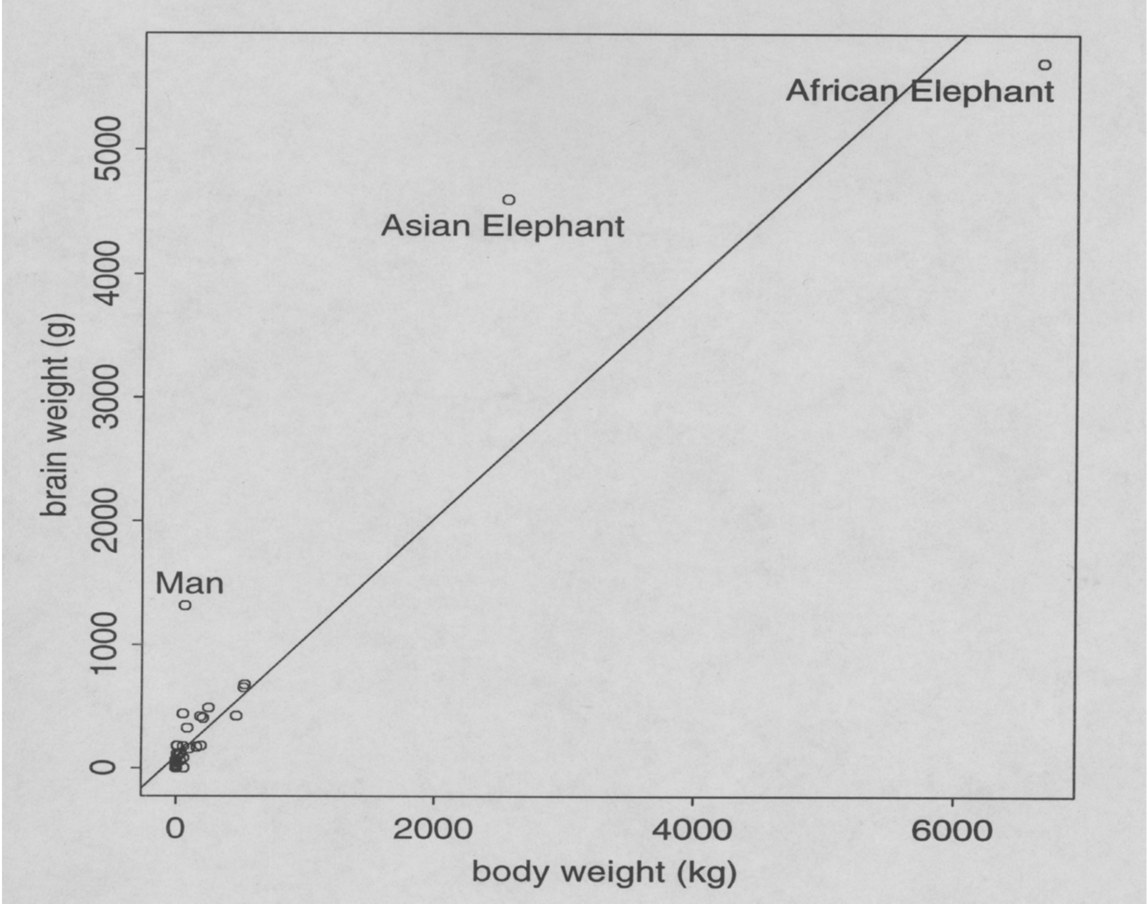
OUTLIERS AND INFLUENTIAL POINTS IN REGRESSION

As an example of outlier detection in a multivariable setting, consider the data on 63 species of terrestrial mammals shown in Figure 2, from Allison and Ciccheti (1976). In any study comparing brain weights of animal species, some correction should be made for body weight. One approach to doing this would be to regress brain weight Y on body weight X in some way, and use residuals. Of course, data in a simple linear regression analysis comes in pairs (X_1,Y_1), (X_2,Y_2), ..., (X_n,Y_n). A particular pair can be unusual in at least two ways: Its X-value can be unusually extreme, in which case the pair is referred to as a "leverage point", and/or its Y-value can be unusually extreme relative to the regression line, in which case the point is labeled an outlier. Diagnostics have been defined to measure / detect each of these conditions (Belsley et al. 1980). For example, the *leverage* of the ith point is defined to be

$$h_i = (1/n) + (X_i - \overline{X})^2 / (n-1)S_X^2$$

i=1,2,...,n, where $\overline{X}$ and S_X^2 are the mean and variance of the regressor. The average value of these h_i values in simple linear regression is 2/n, and the ith data point is (under some conventions) labeled a "leverage point" if $h_i > 4/n$. Some authors prefer a more stringent cutoff value, 6/n. At any rate, leverage points are not necessarily bad; they are just more influential in determining the regression line than the other data points. In the regression shown in Figure 2, both the Asian Elephant (*h*=.1279) and African Elephant (*h*=.8612) are leverage points.

Figure 2. Brain weights and body weights of 63 species of terrestrial mammals (Allison and Cicchetti 1976).



Outliers in regression can be detected by means of *studentized residuals*. Several varieties have been defined, but the so-called externally studentized residual is recommended:

$$r_i = e_i / \sqrt{MSE_{(-i)}(1 - h_i)}$$

where e_i is the ith ordinary residual (actual Y_i - predicted Y_i) and $MSE_{(-i)}$ is the error mean square for the regression excluding the ith pair. Both studentized residuals and leverage points can be obtained (for example) from SAS' PROC REG by requesting their creation in an output data set (SAS 1990).

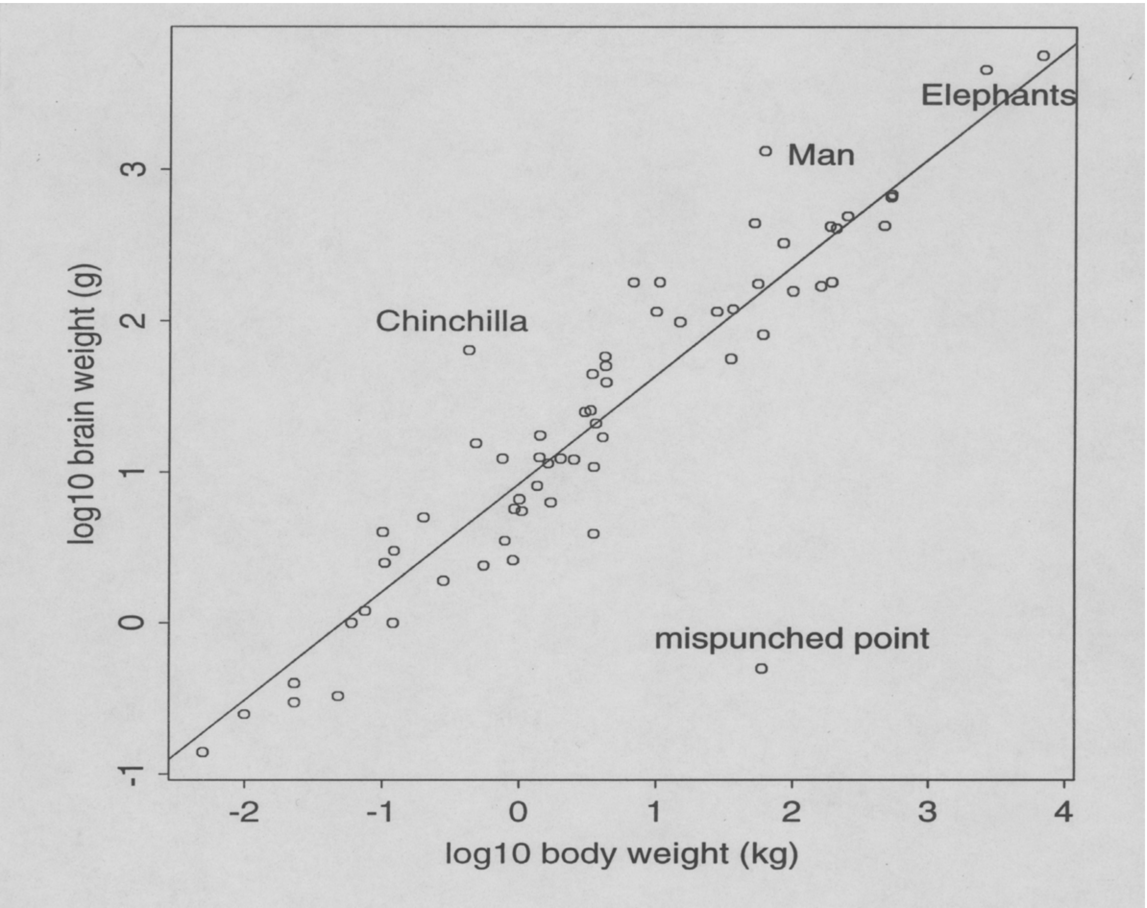
If the formal assumptions of the regression analysis hold, studentized residuals can be used to test for contamination, since each r_i follows a Student's t-distribution with (n-3) degrees of freedom under the hypothesis of no contamination. Hence, a two-sided test would assert contamination if $|r_i| >$ ta $_{/2,n-3}$, the upper-a /2 critical point from the t distribution with n-3 degrees of freedom. In this case, a is a *per-observation* error rate, and should again usually be set lower than .05. For example, in a perfectly "clean" data set containing 100 points, we expect 5 studentized residuals to exceed the a =.05 critical value, and 1 to exceed the a =.01 value, purely by accident. No guidelines have been suggested in the literature, but a @ 1/2n appeals to this author. For the data shown in Figure 2, using a = .01, the critical point is $t_{.005,59}$ = 2.657 and both of the elephants (*r*= 12.30 and -11.85) and also Man (*r*=3.95) flunk the outlier test.

These outlier tests are only valid if the assumptions of the regression hold, however. These assumptions, verbally stated, are:

- The values of the regressor X are known constants (measured with negligible error).
- At any fixed X, the long-run mean of many Y-values, say m(X), is a linear function of X.
- The regression "errors" (the deviations of repeated Y-values at a given X from their long-run mean m(X)) are Normally distributed, with constant variance, and are independent.

In the data of Figure 2, several of these assumptions are either questionable or difficult to assess. Linearity cannot be verified for body weights beyond 1000 kg, since there are so few points at these values. Constant error variance probably doesn't hold, with so many points packed into the lower left hand corner of the plot.

Figure 3. Log_{10}-transformed brain and body weights.



These data vary over several orders of magnitude in both variables, and no analysis of the raw data will distinguish between the lower orders of magnitude. As long as there are elephants in the data, the baboons, lemurs and field mice will all seem equal in size (will all seem to be 0, actually), unless the analysis is done on an order-of-magnitude scale: the log scale. Figure 3 shows a plot of this data in the log scale, i.e. Y*=$\log_{10}$(brain wt) versus X*=$\log_{10}$(body weight). When checked carefully, the formal assumptions of the regression appear to be reasonable, with the possible exception of some points whose Y* values do not fit the pattern (i.e. possible outliers). There are no leverage points now, but the point at lower right in Figure 3, labeled simply as "mispunched point", is a severe outlier since its studentized residual value is *r**= -7.56. The point was in fact artificially planted in this data for the purposes of demonstrating a point, but it is also present (but undetectable) in the raw data of Figure 2. It is also undetectable using univariate outlier tests such as Grubbs' test, since both its X and Y-values are separately well within the range of other values found in the data. This point is the promised example of a model-dependent outlier.

Upon removal of the mispunch and reanalysis, two other points in this data set emerge as possible outliers. Man (*r** = 2.670) barely signals using a =.01, but the Chinchilla's brain weight (*r** = 3.785) is highly unusual given its body weight.

CONCLUSIONS

Some discussion has been offered concerning the prevention and detection of contamination in samples and in regression. Grubbs' test can be adapted to the setting of repeated small samples, as would often be the case in water quality studies, by using a pooled variance estimator over several samples. There are also different versions of the test if one suspects more than one outlier in the sample. Also not discussed is the case of instrument miscalibration, which would result in a possibly large number of "outliers", which are actually shifted variable values, usually by an additive and/or multiplicative constant. Finally, no discussion of modern "robust" statistical methods such as Iteratively Reweighted Least Squares (IRLS) algorithms has been offered (see, e.g., Little 1990). These could, in some cases, be considered to be automatic outlier-detection algorithms; they are potentially very useful, but are still under development. Also, the danger of mindless dependency on automatic detection / elimination algorithms is worrisome.

LITERATURE CITED

Allison, T., and Cicchetti, D.V. 1976. Sleep in mammals: ecological and constitutional correlates. Science 194:732-734.

ASTM E 178-94 1994. Standard practice for dealing with outlying observations. American Society for Testing and Materials, Philadelphia, PA.

Belsley, D.A., E. Kuh, and R.E. Welsch. 1980. Regression diagnostics: identifying influential data and sources of collinearity. John Wiley and Sons, New York, NY.

Chambers, J.M., W.S. Cleveland, B. Kleiner, and P.A. Tukey. 1983. Graphical methods for data analysis. Duxbury Press, Boston, MA.

Deming, W., and D. Edwards. 1986. Out of the crisis. Massachusetts Institute of Technology, Center for Advanced Engineering Study, Cambridge, MA.

Flournoy, N., and L.B. Hearne. 1990. Quality control for a shared multidisciplinary database. Pages 19-23 in G.E. Liepins and V.R.R. Uppuluri, editors. Data quality control: theory and pragmatics. Marcel Dekker, New York, NY.

Grubbs, F.E., and G. Beck. 1972. Extension of sample sizes and percentage points for significance tests of outlying observations. Technometrics 14:847-854.

Lepage, N.J. 1990. Data quality control at United States Fidelity and Guaranty Company. Pages 25-41 in G.E. Liepins and V.R.R. Uppuluri, editors. Data quality control: theory and pragmatics. Marcel Dekker, New York, NY.

Little, R.J. 1990. Editing and imputation of multivariate data: issues and new approaches. Pages 145-166 in G.E. Liepins and V.R.R. Uppuluri, editors. Data quality control: theory and pragmatics. Marcel Dekker, New York, NY.

SAS Institute Inc. 1990. SAS/STAT User's Guide. SAS Institute, Inc., Cary, NC.

Simpson, J., A. Olsen, and J.C. Eden. 1975. A Bayesian analysis of a multiplicative treatment effect in weather modification. Technometrics 17:161-166.

Sokal, R.R., and F.J. Rohlf. 1981. Biometry. W. H. Freeman and Company, New York, NY.

Stolarski, R.S., A.J. Krueger, M.R. Schoeberl, R.D. McPeters, P.A. Newman, and J.C. Alpert. 1986. Nimbus 7 satellite measurements of the springtime Antarctic ozone decrease. Nature 322:808-811.

SCIENTIFIC DATABASES FOR ENVIRONMENTAL RESEARCH

John H. Porter

Department of Environmental Sciences, Clark Hall, University of Virginia,

Charlottesville, VA 22903

Abstract. The questions that scientists can answer are dependent upon the databases available to them. Modern genome research would not be possible without genome databases. Similarly, synthetic and integrative environmental research will be dependent on the quantity and quality of available databases. Examples of scientific databases include large "deep" databases such as Genbank and PDB, "wide" databases such as the National Geophysical Data Center and NASA Distributed Active Archive Centers (DAACs), and project-oriented databases such as those at Long-Term Ecological Research (LTER) sites. There are advantages and disadvantages for using database management systems that balance the capabilities gained against the costs of maintenance. The World Wide Web is a recommended interface for scientific databases. Such databases may be constructed on both UNIX and Windows NT workstations.

INTRODUCTION

There are several advantages to developing and using scientific databases (National Research Council 1997, Pfaltz 1990). First, databases lead to an overall improvement in data quality. Multiple users provide multiple opportunities for detecting and correcting problems in data. A second advantage is cost. Data costs less to save than to collect again. Often, environmental data cannot be collected again at any cost because of the complex of poorly controlled factors, such as weather, that influence population and ecosystem processes. However, the primary reason for developing scientific databases is the new types of scientific inquiry that they make possible. Such inquiries include: (1) long-term studies, which depend on databases to retain project history; (2) syntheses, which combine data for a purpose other than which they were originally collected; and (3) integrated multidisciplinary projects, which depend on databases to facilitate data sharing. Public decisions involving environmental policy and management frequently require data that are regional or national, but most ecological data are collected at finer scales. Databases make it possible to integrate diverse data resources in ways that support the decision-making process.

EXAMPLES OF SCIENTIFIC DATABASES

A useful analogy

A useful analogy in examining scientific databases is to consider individual data sets as "volumes" in a database "library." Libraries may have different sizes and different requirements for cataloging systems. For example, an individual might have a home "library" consisting of a relatively small number of books. The books would not be cataloged or organized, but simply placed on a shelf. An individual book would be located by browsing all the titles on the shelf. For an office library consisting of hundreds of books, a common model is to group books on the shelf by general subject so that only a subset of the library needs to be browsed. However, when the number of books in a library enters the thousands to millions, as for a public library, formal cataloging procedures are required.

This model also applies to scientific databases. If there are relatively few different data sets, a simple listing of the titles of the data sets may be sufficient to allow a researcher to locate data of interest. This is the prevailing model in single-investigator and small project databases. The databases are typically in the form of esoteric World Wide Web (WWW) pages that do not conform to metadata (information needed to use and interpret data) standards.

Examples of databases

Some databases specialize in a single or few types of data and implement sophisticated searching and analytical capabilities. Examples of this type of database are large databases such as Genbank which serves as a primary archive of genetic sequence data for the human genome project, with over one billion nucleotide bases in approximately 1.6 million sequences (National Center for Biotechnology Information 1997) and PDB, the protein structure database which contains over 6,000 atomic coordinate entries for protein structure (*http://www.pdb.bnl.gov/statistics.html*).

These are very large databases with funding in excess of one million dollars per year. In the library analogy, these databases are analogous to large, multi-volume reference works. They are highly "indexed," but focus on a restricted region of the data universe.

There are also various specialized types of databases that operate on a smaller scale. For example, MUSE is specialized software for managing herbarium specimens (Humphries 1997) and BIOTA is software for management of specimen-based biodiversity data (Colwell 1997). Like geographical information system software, these systems are commercially available and are used by a variety of institutions and investigators. In the library analogy, they would be books in a series that share format elements and address the same topic, but have different content. Like the large databases (Genbank, PDB), these databases are "deep" rather than "wide" (Table 1), providing in-depth services for a particular type of data.

Table 1. "Deep" vs. "Wide" databases.

"Deep" Databases	**"Wide" Databases**
• Specialize on one or a few types of data • Large numbers of observations of one (or few) type(s) of data • Provide sophisticated data query and analysis tools • Tools operate primarily on data content	• Contain many different kinds of data • Many different kinds of observations, but relatively few of each type • May provide tools for locating data, but typically do not have tools for analysis • Tools operate primarily on metadata content

"Wide" databases are data repositories that attempt to capture all data related to a specific field of science. For example, the National Geophysical Data Center (NGDC, *http://www.ngdc.noaa.gov/*) is operated by the National Oceanic and Atmospheric Administration (NOAA) and supports over 300 databases containing geophysical data (NGDC 1997). Such "data centers" use standardized forms of metadata (e.g., GILS, FGDC, DIF) for maintaining formal catalogs with controlled vocabularies for subjects and keywords. Similarly, the National Aeronautic and Space Administration (NASA) operates a series of Distributed Active Archive Centers (DAACs; see Olson and McCord, this volume) each of which specializes in supporting a particular area of earth or space science and have a varying number of different types of data sets. In the library analogy, these databases would be comparable to public libraries.

Additional "wide" databases are project-based databases. These are databases that support a particular multidisciplinary research project and may include a wide array of data focused on a particular site or research question. Examples of this type of database are the databases at individual Long-Term Ecological Research (LTER) sites (LTER Network Office 1997). These databases contain data relating to a wide array of scientific topics (e.g., weather and climate, primary productivity, nutrient movements, organic matter, trophic structure, biodiversity, and disturbance), along with information that supports site management (e.g., researcher directories, bibliographies and proposal texts). Management of the databases requires approximately 15% of the total site funding and they focus strongly on long-term data. Within the LTER network, there are diverse approaches to data management dictated by the locations of researchers (at some LTER sites, most researchers are at a single university; at others, they are at many different universities), and the types of data collected (studies of aquatic systems have different data needs than studies of terrestrial systems). Although the LTER network uses individual metadata standards at individual sites, there are network-wide standards for minimum metadata content. These databases are fairly "wide", but not particularly "deep" in the sense that they provide access to a wide variety of data, but do not provide specialized visualization or analysis tools for most types of data. In the library analogy, these databases would be comparable to a large individual or small departmental library.

Some databases, such as individual WWW pages created by individual researchers may be neither "wide" nor "deep." The level of development of such pages varies widely, as does the quality and quantity of the associated metadata. In the library analogy, the pages from a single researcher would be comparable to a very small personal library with little need for searching and cataloging capabilities. As an aggregate, across all researchers, these databases constitute a valuable resource, but one that is difficult to exploit because they can be hard to locate and metadata may be insufficient or difficult to translate into usable forms. Additionally, WWW pages are notoriously ephemeral, so they are a poor choice for long-term database administration.

A strategy for evolving a database

In making the myriad decisions needed to manage a database, a clear set of priorities is the developer's most valuable friend. Every database has some things that it does well (although no part is ever perfect) and some areas that need improvement. The process of database evolution is cyclical. A part of the database may be implemented using state-of-the-art software, but several years later the state-of-the-art has advanced to a degree that it makes sense to migrate the system to new software. Therefore, database systems should be based on current priorities, but with a clear migration path, or at least opportunities, to migrate toward future systems. When making decisions about the types of software to use in implementing the database and associated interfaces, it is critical to consider an "exit strategy." Software that stores data in proprietary formats and provides no "export" capabilities are to be avoided at all costs!

The need for foresight applies to more than just software. The priorities of users may change. A keyword search capability may be a top user priority, but once it exists a spatial search capability may be perceived as increasingly important. It is not possible to implement a database system *in toto*, so the strategy adopted for development must recognize that, although some capabilities are not currently implemented, the groundwork for those capabilities in future versions must be provided for. Thus, even though an initial system may not support spatial searching, collecting and storing spatial metadata in a structured (i.e., machine-readable) form is highly desirable.

An important form of foresight is seeking scaleable solutions. Scalability means that adding or accessing the 1,000th piece of data should be as easy (or easier) as adding the first. The genome databases faced a crisis when the flow of incoming data started to swamp the system (which depended on some level of manual curation of inputs). The subsequent adoption of completely automated techniques for submission and quality control allows the genome databases to handle the ever increasing flows of data. Every system has some bottlenecks and their identification and elimination before they become critical, is the hallmark of good planning and management.

Choosing Software

The choice of software for implementation of a database must be based on an understanding of the tasks you want the software to accomplish (e.g., input, query, sorting, analysis). Simplicity is the watchword as the world is full of sophisticated software that is expensive and difficult to operate, but that may provide little real improvement over simpler and less expensive software.

User interface

Although a variety of proprietary user interface options exist, it is hard to argue against using an interface based on World Wide Web (WWW) tools. Most potential users of a database will already have, or have access to, a WWW browser (e.g., Netscape and Microsoft Explorer) so there is no need to distribute specialized software. Most potential database users will already be familiar with a WWW browser, reducing the need for training. WWW tools continue to improve at a rapid pace. Important innovations have been the support of on-line forms and linking WWW servers to database engines so that WWW pages can be dynamically generated. The addition of programming languages (such as JAVA^{TM}) which allow secure operation of applications on the client-side greatly increases the types of operations that can be supported over the WWW. WWW tools can be used for input to a database, as well as for output. An advantage of this approach is that input of metadata and data can be made from many different locations, which can circumvent some potential bottlenecks.

Advantages and disadvantages of using a database management system (DBMS)

There are numerous advantages to using a DBMS. The first is that a DBMS has many useful built-in capabilities such as sorting, indexing, and query functions (Maroses and Weiss 1982). Additionally, large relational databases include extensive integrity and redundancy checks and support transaction processing with "rollback" capabilities, allowing one to recreate the database as it existed at a particular time. There has been substantial research into making relational DBMS as efficient as possible and many DBMS can operate either independently, or as part of a distributed network. This aids in scalability because if one computer starts to become overloaded, another can be added without having to substantially restructure the underlying system. Finally, most DBMS's include interfaces that allow DBMS linkage to user-written programs or other software, such as statistical packages. This is useful because it allows one to change the underlying structure of the data without having to alter programs that use the data.

Despite these advantages, most DBMS's are designed to meet the needs of business applications and these may be quite different from the needs of scientists (Maroses and Weiss 1982, Pfaltz 1990). For example, most commercial DBMS's have few graphical or statistical capabilities. DBMS's are typically designed to create standard reports that may be of little use to researchers. Additionally, DBMS's are typically designed to deal with large volumes of data of a few specific types. They are less useful when dealing with relatively small volumes of data of many different types. Similarly, they can be relatively inefficient in dealing with sequential data. There are some functions, such as highly optimized updating capabilities, that are not frequently used for scientific data because, barring detection of an error, data are seldom changed once in the database. Additionally, not all analysis tools can be easily interfaced with a DBMS and proprietary data formats used by a DBMS may limit archival quality of data. A final disadvantage of a DBMS is that it requires expertise and resources to administer. In some cases, the resources required may exceed the benefits accrued by using a DBMS.

Even if you decide not to use a DBMS for data, you may want to consider use of a DBMS for metadata (documentation). The structure of metadata is frequently more complex than that of data and conforms better to the model of business data (relatively few types of data, standard reports are useful). Most data are located based on searching metadata rather than the data itself so the query capabilities of a DBMS are useful. Similarly, metadata are changed more often than data, so that the updating capabilities of a DBMS are more useful for metadata.

Choosing a Computer System

At this time, there are two reasonable options for computer systems which support full-featured database creation: computers running UNIX and computers running Microsoft Windows NT. UNIX is a mature, full-function operating system. It has strong capabilities for multitasking and multi-user support. As a mature system, it is reliable and robust and there is a large body of WWW tools, many of which are free. On the down side, UNIX is difficult to learn and commercial software for UNIX is typically much more expensive than that for personal computer-based systems.

Microsoft Windows NT is a rapidly evolving operating system that has seen major improvements in operating system design that facilitate network access. Compared to UNIX, software and hardware are relatively inexpensive and most software is more "user friendly" than UNIX. The number of WWW tools for NT is growing rapidly. Limitations of databases on an NT are that they are more difficult to scale up than those on UNIX computers, and as a relatively new operating system, there can be problems with reliability.

The capabilities of these systems are similar enough that choice of a system may depend on the local computational environment. If UNIX computers are already in place and there is sufficient expertise to support them, UNIX may be the best choice. However, if those prerequisites are lacking, an NT system may be the better choice.

CONCLUSIONS

Development of scientific databases is an evolutionary process. Although databases evolve, they do not spontaneously generate! It takes the actions of an individual or group to bring them into being, often in a relatively simple form. It is not necessary that a new database try to incorporate all the features it will eventually encompass. Indeed, to do so is a prescription for disaster because it is extremely difficult to anticipate all the needs of the user community. Even if it starts in a simplified form, once in operation, a successful database generates its own momentum by coupling its user community into the development process.

LITERATURE CITED

Colwell, R.K. 1997. Biota: the biodiversity database manager. *http://viceroy.eeb.uconn.edu/biota*

Humphries, J. 1997. MUSE. *http://www.keil.ukans.edu/muse/*

LTER Network Office. 1997. Long-term Ecological Research. *http://LTERnet.edu*

Maroses, M. and S. Weiss. 1982. Computer and software systems. Pages 23-30 in G. Lauff and J. Gorentz, editors. Data Management at Biological Field Stations. A report to the National Science Foundation.

National Research Council. 1997. Bits of power: issues in global access to scientific data. *http://www.nap.edu/readingroom/books/BitsOfPower/* National Academy Press.

National Center for Biotechnology Information. 1997. GenBank Overview. *http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html*

Pfaltz, J. 1990. Differences between commercial and scientific data. Scientific database management, a report to the National Science Foundation. *gopher://lternet.washington.edu:70/00/newsletters/Reports/Miscell/uva_cs_90/cs_90-22*

Porter, J.H. and J.T. Callahan. 1994. Circumventing a dilemma: historical approaches to data sharing in ecological research. Pages 193-203 in W. K. Michener, S. G. Stafford and J. W. Brunt, editors. Environmental information management and analysis: ecosystem to global scales. Taylor and Francis, Bristol, PA.

DATA ARCHIVAL

Richard J. Olson and Raymond A. McCord

Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831-6407

Abstract. A data archive is a permanent collection of data sets with accompanying metadata such that secondary users can readily acquire, understand, and use the data. Although data archival for ecology is in its infancy and there are a limited number of permanent data archives for ecological data, ecologists can manage their data in ways that facilitate data sharing and prepare their data for eventual archival. In this chapter, incentives for archiving data are presented, components and functions of data archives are reviewed, and future directions for data archival are discussed.

INTRODUCTION

Traditionally, science involves making systematic observations that can be replicated. Over the past two decades, there has been a shift from traditional studies of isolated ecosystems toward more broad-scale modeling, synthesis, and assessment studies. Scientists are collecting data over the Internet in addition to doing field or laboratory work. As ecology moves toward regional and global multidisciplinary studies, mechanisms for sharing data with many disciplines (meteorology, hydrology, soil science, forestry, agriculture, botany, etc.) are needed. Consequently, submitting data to archives and acquiring data from archives are integral parts of today's scientific process. However, data archival is not yet given the attention, resources, or recognition required for it to become a routine part of the research and publication cycle.

A data archive is a permanent collection of data sets with accompanying metadata such that secondary users can readily acquire, understand, and use the data. An archive preserves data and metadata in an electronic form that will continue to be accessible as technologies change. An archive provides complete metadata so that secondary users with interests varying from watersheds to global change understand inherent limitations of the data and use the archived data properly in new applications. Whereas "archive" may imply simple preservation, the implicit goal is to facilitate data sharing to foster broader ecological discoveries. Archives are more than a long-term backup or an index or catalog with pointers to data sets stored elsewhere.

Data archival for ecology is in its infancy and there are a limited number of permanent data archives for ecological data. Often, different terms and functions are associated with data sharing and storage activities (Table 1). Informal sharing, repositories, or digital libraries may provide much or all of the functionality of an archive. The concepts presented here for data archives readily apply to other, less-formal data storage activities.

INCENTIVES FOR DATA ARCHIVAL

Although most ecologists may support the concept of data archival and even use data from archives in their research, they generally do not archive their own data. There is a trend for federally sponsored research announcements to require that the data generated by any proposed project be placed in an archive. In this section, we explore some of the factors that may contribute to avoidance of data archiving and some incentives to promote data archiving. In many ways, archiving data is equivalent to preparing a publication. The time and resources required to archive data can be equal to or exceed those for preparing, editing, and reviewing a publication. There may be uncertainty about the detail, format, and style of metadata (see Michener et al. 1997). In addition, there may be fears that providing immediate public access to data may result in others preempting the contributing investigator's opportunity to publish his or her findings

Table 1. Distinctions between data exchange and storage activities.

Venue	Data Exchange	Comments
Data custodianship	Data sharing by request, usually with colleague	Current expert, provides technical information, could authorize changes

	having technical expertise	to the data, may be primary compiler or inherited role
Data stewardship	Data sharing by request	Gatekeeper, minimal knowledge of the data, may inherit data from custodian
Data repository	May limit data access	Usually project-level support; limited functionality, may cease to exist after project ends
Digital library	Public access	Broad subject area, limited expertise and user support, includes tabular and graphic data
Data archive	Public access	May have thematic emphasis, search and order, long-term commitment, packaged metadata

first. Unfortunately, the scientific community, especially at administrative levels, currently does not acknowledge well-documented data as equivalent to a publication.

In order to facilitate free and open exchange of data among ecologists and to create the reward structure necessary to encourage ecologists to share and exchange data (Porter and Callahan 1994), the ESA Data Sharing and Archive Committee is proposing that the ESA create a venue in which the Society will electronically publish peer-reviewed data papers (Ellison, A. M., personal communication Oct 24, 1997). Data papers are envisioned to include extensive data and metadata (basically an expanded materials and methods section without results and discussion sections). If approved, the Committee will develop guidelines for publishing data and metadata and an appropriate peer review process. The data and metadata would be maintained in a long-term archive, and contributors and users would be able to cite the published data papers as they now cite papers in other ESA journals.

Project leaders, sponsors, and science managers can provide the following incentives for investigators to archive data:

- establish a citation policy to give credit to data contributors,
- establish a citation policy to give credit to multiple contributors to integrated data sets,
- provide adequate resources for data management to investigators,
- involve data personnel in the initial project planning,
- provide guidelines and training for metadata preparation,
- produce high-visibility data products (e.g., CD-ROMs or hardcopy data products), and
- give credit (i.e., include in promotion and salary actions) to those who produce well-documented data sets.

DATA ARCHIVE COMPONENTS

Archives consist of more than data and metadata. Other key components are the data storage system, information system, network connections, a security and backup system, data analysis tools (optional), archive staff, and, most importantly, a user services support staff. Although we will not fully discuss the computer technology component of data archives, the advances in this area, especially PCs, networks, and the World Wide Web (see Porter, this volume), contribute greatly to the growth of data archives. Typically, staff are a mix of systems specialists, database administrators, user interface specialists, information specialists, and scientists.

The Environmental Sciences Division of Oak Ridge National Laboratory (ORNL) has a 25-year history of managing and archiving ecological data, starting with the data from the International Biological Program (IBP) in the early 1970's. Currently ORNL has four data archive centers (Table 2). Each center uses a different technology and organization; however, all emphasize the combination of computer specialists and scientists, provide useful metadata, and supply citation information so that the original data contributors can be correctly acknowledged.

We have found that it is essential that scientists be involved in the archive operations. Scientists play a critical role in the organization and presentation of data, quality assurance/quality control review of data and metadata, and development of value-added products. In addition, an advisory group can effectively represent the interests of data contributors, secondary users, and sponsors.

DATA ARCHIVE FUNCTIONS

The flow of data from a contributor to a publicly accessible archive is a multi-step, potentially time-consuming process. Although there may be many variations on the overall process, an initial step is to connect a contributor to the appropriate data archive. Most archives continually work to identify user needs and data availability and to establish priorities to acquire new data. Often, a data archive is associated with a specific program or has a thematic orientation and will actively seek selected data sets. The data archive provides guidelines to the contributor for formatting and submitting data and metadata to the archive. The contributor prepares data and metadata as completely as possible and submits them following archive guidelines.

Archive staff review the data and metadata and may reformat them to achieve consistency and completeness, making sure that the metadata supply citation information so that the original data contributors can be correctly acknowledged. Staff also review the quality assurance that was performed on the data as documented in the metadata. They may also select keywords based on the metadata to be used in search and order functions. Archive staff and contributors work together to resolve questions and review changes. After consensus is reached, metadata and data are entered into the archive for public access and long-term storage.

In addition to maintaining a long-term, secure data archive, the archive staff also provide post-project support, such as answering user questions, informing users of updates and additions, and maintaining user statistics. Archives must also plan for the periodic upgrading of storage media. Archives can perpetuate the growth and value of their data holdings by including a strategy for incorporating data updates, value-added products (especially from synthesis and modeling applications), and user feedback. Staff can collaborate with scientists to determine useful enhancements to data sets (e.g., add common variables, aggregate to common units, or calculate uncertainty) on the basis of user needs. It is also crucial for the staff to promote the availability of the data by interacting with the user community, through attendance at professional meetings and use of Web marketing techniques.

Table 2. Data Archives at ORNL.

Data Archive / Web Address	Focus	Special Features
Carbon Dioxide Information Analysis Center (CDIAC) *http://cdiac.esd.ornl.gov*	Acquire, compile, quality-assure, document, archive, and distribute information on greenhouse gases and climate change in support of the US Department of Energy's (DOE) Global Change Research Program. Since 1992, CDIAC has hosted a component of the World Data Center-A for Atmospheric Trace Gases of the International Council of	Special emphasis on quality assurance, documentation, and derived, integrated products. User community includes many thousands of researchers, policymakers, educators, students, and corporate officials around the world. User services office.

	Scientific Unions to store and manage data on radiatively active trace gases and their concentrations.	
Atmospheric Radiation Measurement (ARM) Program Archive *http://www.archive.arm.gov*	Improve radiative transfer functions in General Circulation Models (GCMs) and the parameterization of cloud properties and formation in GCMs as part of the DOE's ARM Program.	Stores massive amounts of data: >2 million files (>1500 GB) and 70,000 new files (70 GB) per month. Provide large volumes of data to scientists: ~22,000 files (20 GB) per month, ~400 registered users.
Distributed Active Archive Center for Biogeochemical Dynamics (ORNL DAAC) *http://www~eosdis.ornl.gov*	Acquire, quality-assure, document, and archive multidisciplinary data for terrestrial ecosystems and provide access to the global change research community, policymakers, and educators, as part of the National Aeronautics and Space Administration Earth Observing System.	Web-based data search and order interface, browse/view data before ordering, multiple distribution media, free and ready access, user services office.
Oak Ridge Environmental Information System (OREIS) *http://www~oreis.cad.ornl.* *gov:8080/oreis/help/oreishome.html*	Develop a consolidated database to support environmental cleanup activities on the Oak Ridge Reservation for DOE.	Full relational database management system with links to statistical and geographic information system (GIS) tools, 5 million records added since 1994 (password-protected).

FUTURE DIRECTIONS FOR DATA ARCHIVAL

Ecological synthesis and assessment studies that address long-term regional and global ecological issues will continue to expand and use data from data archives. Sharing and archiving data can be more efficient if the following general principles are considered in the overall project planning and operations: (1) establish the flow of data from investigator to a long-term data archive as part of the work plan; (2) process data to achieve consistency and completeness of data and metadata; and (3) institute policies to give data producers adequate credit for their data archival efforts.

To more fully share data, we suggest the scientific research community embrace the following actions:

- provide incentives for sharing and archiving data,
- recognize data sets with metadata as valuable research products,
- establish a universal citation policy for data,
- establish guidelines for metadata,
- develop data distribution and archive centers, and
- ensure long-term financial and institutional support.

As ecologists, we have an opportunity to educate and lobby science administrators, program managers, and agencies about the data archival process, its intrinsic value, and required resources.

ACKNOWLEDGEMENTS

We would like to thank S. W. Christensen and B. T. Rhyne for their helpful reviews of this manuscript. This work was sponsored by the National Aeronautics and Space Administration (Distributive Active Archive Center project) under Interagency Agreement DOE No. 2013-F044-A1 and by the US Department of Energy (Atmospheric Radiation Measurement Archive project). Oak Ridge National Laboratory is operated by Lockheed Martin Energy Research Corp. under contract DE-AC05-96OR22464 with the US Department of Energy.

LITERATURE CITED

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(1): 330-342.

Porter, J. H. and J. T. Callahan. 1994. Circumventing a dilemma: historical approaches to data sharing in ecological research. Pages 193-202 in W. K. Michener, J. W. Brunt and S. G. Stafford editors. Environmental Information Management and Analysis: Ecosystem to global scales. Taylor & Francis, Bristol, PA.

THE WORLD WIDE WEB AS A TOOL

FOR ECOLOGICAL RESEARCH PROGRAMS

Barbara J. Benson

Center for Limnology, University of Wisconsin-Madison, 680 N. Park Street,

Madison, Wisconsin 53706

Abstract. The World Wide Web provides a new technology for the ecological research community to disseminate information and facilitate research programs. In this chapter, the history of the World Wide Web (WWW) and its implementation at Long-Term Ecological Research sites are reviewed. A summary of WWW use in the context of ecological research is presented along with selected examples.

INTRODUCTION

The rapid expansion of the World Wide Web (WWW) has created a new level of accessibility for ecological data and information. A review of the use of the WWW by sites within the Long Term Ecological Research Program (LTER) provides an overview of multiple ways in which the WWW can facilitate ecological research programs.

HISTORY OF THE WWW

Shipley and Fish (1996) observed that "the Web has exploded into an information revolution and a cultural phenomenon." Prior to the WWW, finding specific data or information on the Internet could be difficult. In 1989, Tim Berners-Lee at CERN (European Laboratory for Particle Physics) proposed a hypertext system that would provide simple and consistent access to documents from any source. He invented communication protocols that incorporated existing information systems (such as Gopher, and ftp) and browsing software capable of running on all platforms (Kennedy 1995). In 1993, the first graphical browser (Mosaic, NCSA) was released. The WWW has experienced dramatic growth since that time.

Table 1. Growth of the World Wide Web.

WWW Statistics

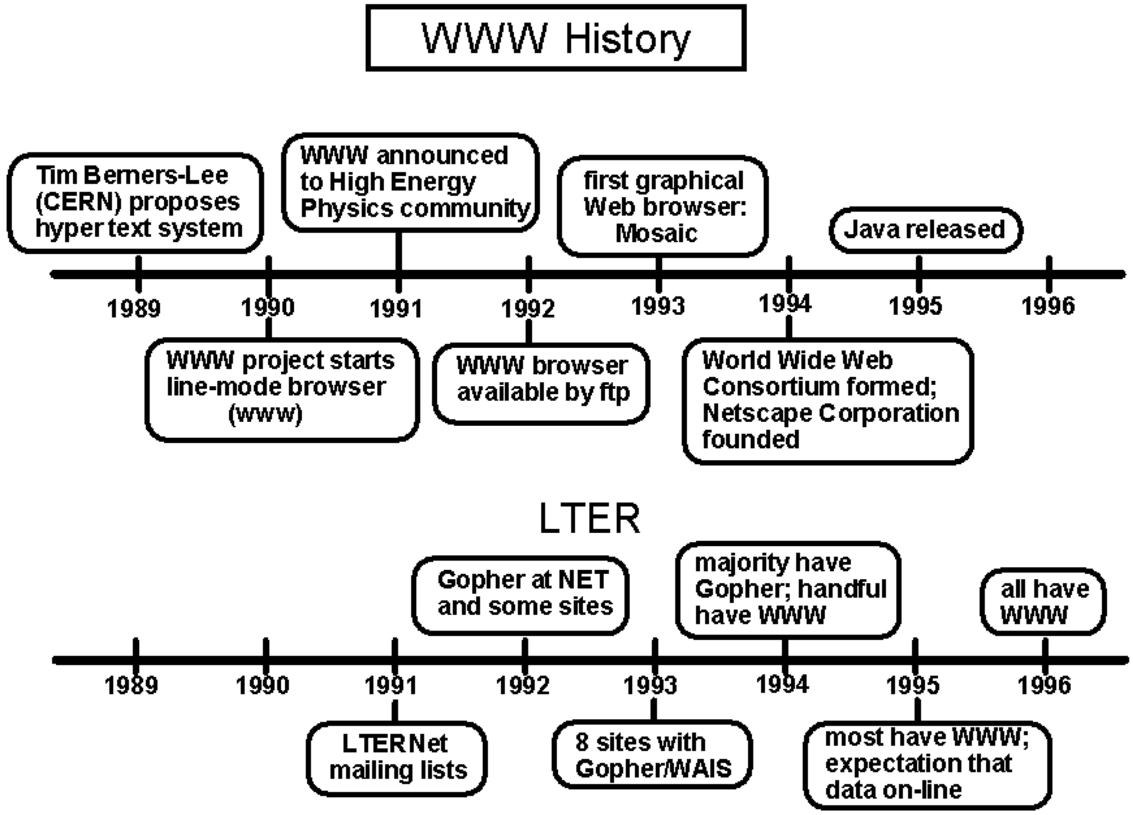
(from Matthew Gray of the Massachusetts Institute of Technology,

http://www.mit.edu:8001/people/mkgray/net/index.html)

Month	Number of Web Sites	Percent ".com" Sites
6/93	130	1.5
1/97	650,000 (est.)	62.6

With current WWW browsers, a vast network of information is available including multimedia documents with audio and video. These browsers also include email and conferencing functionality.

Figure 1. Highlights of the development of the World Wide Web (top) and its implementation at the LTER sites (bottom).


WWW History
Tim Berners-Lee (CERN) proposes hyper text system
WWW announced to High Energy Physics community
first graphical Web browser: Mosaic
Java released
1989
1990
1991
1992
1993
1994
1995
1996
WWW project starts line-mode browser (www)
WWW browser available by ftp
World Wide Web Consortium formed; Netscape Corporation founded
LTER
Gopher at NET and some sites
majority have Gopher; handful have WWW
all have WWW
1989
1990
1991
1992
1993
1994
1995
1996
LTERNet mailing lists
8 sites with Gopher/WAIS
most have WWW; expectation that data on-line

Data managers within the LTER research network quickly recognized the potential of the WWW (Ingersoll and Brunt 1995). By 1994, some of the eighteen LTER sites had home pages on the WWW, and by 1995 most sites had a presence on the WWW. At the same time, there was a growing expectation within and outside the LTER community that LTER data be accessible on the WWW.

USES OF THE WWW BY ECOLOGICAL RESEARCH PROGRAMS

From the perspective of an ecological research site, how can the WWW be used to enhance research programs? In presenting an overview of how sites within the LTER program are using the WWW, addresses (URL's) to specific locations on the WWW will be included as examples (Table 2). These examples provide only an introduction to what has been done. For a more thorough examination, all LTER site WWW servers can be accessed through the LTER Network home page (*http://lternet.edu*).

Research sites get many different types of inquiries for information, ranging from the K-12 student working on a class project, to prospective graduate students wanting to know more about the program, to colleagues looking for particular types of data for a research project, to members of the local research group looking for a copy of a document or the current calendar of events. A wide variety of information types can be provided on the site's home page to meet the needs of diverse users.

An LTER Site home page can provide access to general information about the research site, including site characteristics (e.g., Coweeta LTER Basin map, North Temperate Lakes LTER map to field station, North Temperate Lakes LTER lake characteristics). The research program can be described detailing major objectives, approaches to questions, research results and future directions. A personnel directory can be provided including address, phone, email address, fax, a biographical sketch, and links to other information. The personnel directory can be searchable and set up to provide on-line updates. Important site documents can also be accessed through the WWW. These documents might include a site history, a bibliography of publications, recent research proposals, and electronic publications.

Table 2. Locations of example documents from selected LTER site home pages on the WWW.

Research Site Information	
Coweeta LTER Basin map	*http://sparc.ecology.uga.edu/webdocs/gis/html/ maparchive/cwtbase.html*
North Temperate Lakes LTER map to field station	*http://limnosun.limnology.wisc.edu/tls/map/map.html*
North Temperate Lakes LTER lake characteristics	*http://limnosun.limnology.wisc.edu/lter_lake.html*
Sevilleta LTER personnel directory and links to home pages of individuals	*http://sevilleta.unm.edu/people*
Niwot Ridge LTER searchable bibliography	*http://culter.colorado.edu:1030/Niwot/Niwot_ Ridge_LTER_bibliography.html*
Data and Metadata	
Bonanza Creek LTER data catalog	*http://www.lter.alaska.edu/cgi-bin/w3- msql/dfd/dfd.html*
Short Grass Steppe LTER dynamic queries of database	*http://sgs.cnr.colostate.edu/data/data_cat/climateindex.html*
Virginia Coastal Reserve LTER spatial data archive	*http://www.vcrlter.Virginia.EDU/data/TMairAtlas.html*
Biodiversity information: Cedar Creek LTER catalog of flora	*http://www.lter.umn.edu/florfaun/flora/t1.html*
Project Management	
Niwot LTER data management policy	*Gopher://culter.colorado.edu/00/.NWTPOLICY.* *TXT*
Virginia Coastal Reserve LTER calendars	*http://atlantic.evsc.virginia.edu/calendar.html*

There has been considerable development on the WWW to provide access to data and metadata. Sites have data catalogs that provide an overview of data set availability and data set descriptions. LTER data sets are supposed to be available on the WWW within two years after collection with a minimum of restrictions (S. Collins, e-mail communication). Many sites make their data available as text files; however, some sites whose data reside in relational databases have developed programs that provide dynamic queries of the database. Through dynamic queries, subsets of a database in which the user has specified variables and the time period of interest can be provided. The results of such queries can be displayed graphically. Because databases are updated over time, it is desirable to maintain a log of database updates on the WWW so that users can determine whether the data which they downloaded at a given point in time is still consistent with the most current version of the data.

Spatial data such as GIS coverages, satellite images, and photo archives can also be provided through the WWW. Catalogs of spatial data may include thumbnail versions so that the user may have a look at the image without having to download a large file. Thumbnails may also provide information on satellite images that are not licensed to be distributed publicly.

Sites have published other types of data and products on the WWW including models and software developed at a site. Biodiversity information can be provided as species lists or more detailed descriptions of fauna and flora.

The WWW can be used to facilitate data management activities and to deal with issues related to data access. Researchers can provide the required metadata for submitted data sets through the use of forms for metadata entry. All sites provide a statement of their data access policy and some require those who download data to provide information through the use of forms. The collected information can then be passed to site researchers. Documents detailing data management policy and protocols have also been provided at most sites.

Project management can be facilitated through the WWW by communicating information of primary interest to the local research group. Publishing calendars and interactive forms for scheduling trips and equipment use has proved useful at some sites.

The amount of site information and data that are provided through the home page of an LTER site is extensive. At least one site is developing navigation tools (Andrews LTER; *http://www.fsl.orst.edu/lter/navigafr.htm*) to aid browsing through their large collection of linked documents.

The WWW page may also contain links to other sites. Some common links are weather information, professional societies (e.g., ESA), funding agencies (e.g., NSF), affiliated institutions, and sources of Internet information and help.

At a larger scale, the WWW can be used as a tool for an entire research network or even a network of networks. The home page of the LTER Network (*http://lternet.edu*) illustrates how the WWW can be used to create a unified point of entry to a distributed information system. Some network-level information products developed by the LTER data managers and network office data management staff include an all-site bibliography, personnel directory, and data catalog. Current projects include an integrated climate database across all sites and an expanded, updated data catalog.

In the future, the WWW will expand its contribution to ecological information systems. Development of documents for the WWW has become easier with the advent of HTML editors and the integration of HTML exporting capabilities in word processing software. The use of forms is allowing the flow of information to be two-way. Java ^{TM} (*http://java.sun.com*; Campione and Walrath 1996) scripts are being developed to automate documentation, quality control and data processing (see Jones, this volume). Research groups are exploring the WWW as an environment for collaboration.

ACKNOWLEDGMENTS

Financial support was provided by the National Science Foundation's Long Term Ecological Research Program, grant # DEB-96-32853.

LITERATURE CITED

Campione, M., and K. Walrath. 1996. The Java^{TM} tutorial: object-oriented programming for the Internet. Addison-Wesley. *http://java.sun.com/docs/books/tutorial*

Ingersoll, R., and J. Brunt., editors. 1995. Proceedings of the 1994 LTER data management workshop. *gopher://time.lternet.edu:70111/newsletters/Reports/DMworkshops/1994/94report.asc*

Kennedy, A. J. 1995. The Internet & World Wide Web: the rough guide. Rough Guides Ltd., New York, NY.

Shipley, C., and M. Fish. 1996. How the World Wide Web works. MacMillan Computer Publishing, New York, NY.

PROVIDING INFORMATION ON THE WORLD WIDE WEB

John H. Porter

Department of Environmental Sciences, Clark Hall, University of Virginia,

Charlottesville, VA 22903

Abstract. The capabilities provided by the World Wide Web (WWW) offer an opportunity for ecological researchers to share information resources. The hypertext markup language (HTML) is used to create documents for display on the WWW. HTML documents can be created using various tools from general text editors to more specialized programs. General principles of WWW page design that can be applied to improve content and usability include anticipating user needs and avoiding features that unnecessarily increase needs for network bandwidth. Search and indexing tools for WWW pages can be used to improve access to information. WWW pages can also be used to solicit information from users via on-line forms. Making HTML documents available on the WWW is accomplished by placing them on a server, which may be locally administered or available commercially.

GETTING STARTED - DEMYSTIFYING HTML

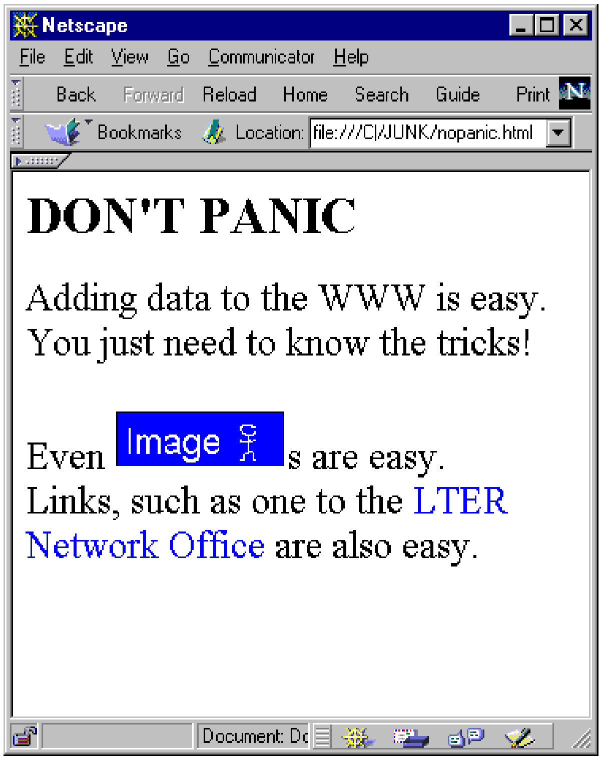
Information provided on the World Wide Web (WWW) comes in many multimedia forms. WWW software supports display of text, graphics, animation and sound. Despite its sophisticated and extensive capabilities, the underlying technologies are surprisingly simple. Web pages are written in the "HyperText Markup Language" (HTML). HTML files are simple ASCII text files that can be created by any software editor. They take the form of text enclosed in HTML "tags." The HTML "tags" are text enclosed in <>s. For example, <B> starts boldface and </B> turns off Boldface "a <B>test</B>" would appear as "a **test**."

A more extensive example that includes both an image and a linkage to a WWW server is:

```
<H1>DON'T PANIC</H1>

<P>Adding data to the WWW is easy. You just need to know the tricks! </P>

<P>Even <IMG SRC=image.gif>s are easy. <BR> Links, such as one to the <A
HREF="http://lternet.edu">LTER Network Office</A> are also easy.

</P>
```

Note that most tags occur as pairs. <P> and </P> are used to separate paragraphs. <H1> and </H1> delimit a heading, which is displayed in large type. There is an "anchor" <A HREF="http://lternet.edu"> </A> which outlines text ("LTER Network Office") that is displayed in blue. If clicked on in the browser, you would be transferred to the WWW site or document described in the HREF section of the anchor. Other tags occur as singletons. <BR> produces a line break. <IMG SRC="*name_of_graphics_file*"> inserts a graphics file. In the example the graphics file contains the word "image" in a blue box.

Figure 1. Illustration of page display (see text) from a WWW browser.


Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print
Bookmarks Location: file:///C|/JUNK/nopanic.html
Document: Dc

DON'T PANIC

Adding data to the WWW is easy. You just need to know the tricks!

Even Image s are easy. Links, such as one to the LTER Network Office are also easy.

An excellent "HTML Table of Contents," assembled by Ian Graham, is available at: *http://www.utoronto.ca/webdoes/HTMLdoes/NewHTML/htmlindex.html*. However, it is not necessary for a creator of HTML documents to know or understand HTML because of the increasingly diverse range of software tools that create HTML documents. Recent versions of Netscape, for example, include a built-in editor for HTML documents. Similarly, Microsoft WORD allows you to "*Save As*.." an HTML document. WordPerfect for Windows and many other software packages include an HTML converter. There are also specialized packages, such as Microsoft FrontPage, that focus on WWW publishing.

If you see something interesting about the structure of a WWW page provided by someone else, it is possible to see exactly how he or she did it. WWW servers send copies of HTML files to the browser for interpretation and display. Using the "view source document" command you can look directly at the HTML source used to generate the page and see the commands responsible for creating the interesting page display and apply those same "tricks" to creating your own pages.

WWW PAGE DESIGN

Design of WWW pages is an art rather than a science. Nonetheless, there are some principles that can be applied to the creation of effective and informative WWW pages. The first is to know your audience! What types of information will they want? What types of information will they need most? The second is to use that knowledge to structure information on the server. Frequently requested information should be easy to locate. Information should be grouped into understandable categories to keep the number of menu entries at roughly seven or fewer so that the entire menu can appear on a single page.

Use of graphics is an area of both great opportunity and great hazard. Appropriate graphics can add interest and clarity to a page. However, they need to be used with caution. Graphics files (typically .GIF or .JPG files) are frequently large and WWW page displays can be significantly slowed by the inclusion of several graphics files. This problem is especially acute for users with slow network connections. Although they have improved dramatically in recent years, modem connections via telephone lines are tens to hundreds of times slower than direct connections to the Internet via a local area such as an ethernet. Thus a graphic image that may take only seconds to display at 10 megabits per second, can take several minutes to display at 56 kilobits per second.

Animated graphics place an additional load on the user's computer, which can significantly impact other applications.

Graphical page backgrounds are an especially sensitive issue. They can be quite attractive on one system, but may display differently on other systems, which have different color capabilities. In some cases, the display of the background may be so poor as to make the overlaying page unusable. For this reason, it is a good idea to test any background on a variety of systems and software before employing it.

JAVA^{TM} Applets are another opportunity. These applets are small programs that run on the user's system. They can add interactive characteristics to a WWW site that go beyond those available with HTML alone. However, like graphics, they can substantially increase the load imposed on the user's system and should therefore be used with caution.

USING A WWW SERVER

Adding pages to a WWW server is easy. It is merely a question of uploading the relevant HTML and graphics files to a directory on the server where they are accessible. Most educational institutions have servers that are available for use by faculty and students. On a university UNIX system, the process is often as simple as creating a directory named "public_html" in your home directory. Files placed in the "public_html" directory are then accessible over the WWW at the address: *http://address.of.server/~your_id/filename.html* where *address.of.server* is the network address of the computer (such as *poe.acc.virginia.edu*), *your_id* is the user-ID for your account and *filename.html* is the name of the specific file stored in your "public_html" directory. Commercial Internet Service Providers (ISP's) can also provide access to existing WWW servers.

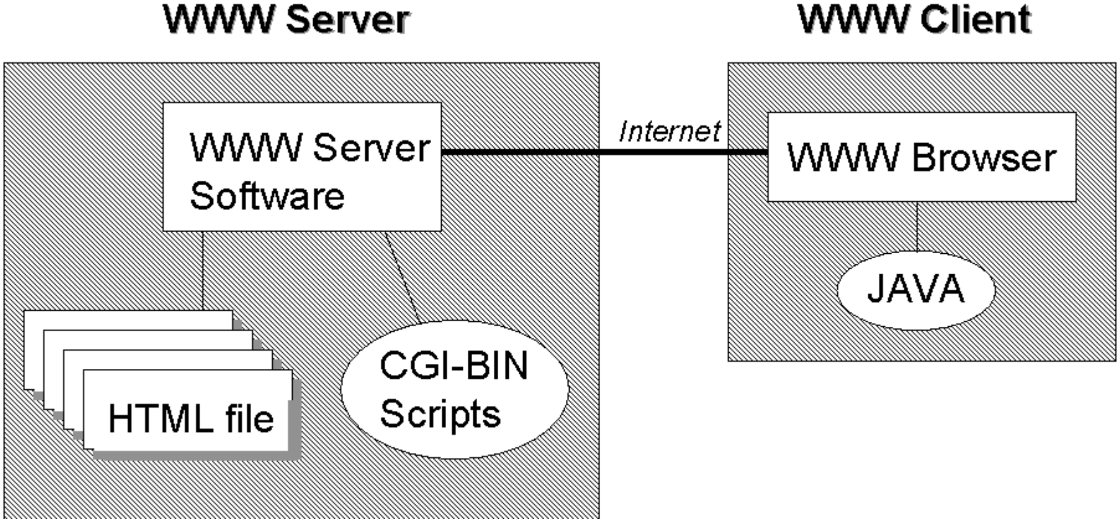
You can also develop your own server. Full-featured server software is available for Windows 95^{TM}, Windows NT^{TM}, Unix and Macintosh systems, both as commercial software and as shareware. More limited servers are available for MSDOS. The only limit on a server computer is that it needs to have a stable network address and be accessible 24 hours a day (since users are seldom in a single time zone).

Server software for the WWW is relatively simple. A browser program on the client computer sends a request to the server for a particular file (such as an HTML document or graphics file). The server then sends a copy of the requested file back to the client for display. Indeed, it is one of the paradoxes of the WWW that the client browser program needs to be much more sophisticated than the server! Whereas a basic server simply needs to respond to requests for files (similar to File Transfer Protocol - FTP), a client browser (such as Netscape) needs to know how to convert the HTML files and graphics files into forms suitable for display on the screen. More sophisticated servers support Common Gateway Interface (CGI) scripts which dynamically generate WWW pages and support passwords and other security features. Additionally, they may include extensive logging capabilities to track use of individual pages.

In addition to the display of "static" HTML documents, the WWW can also display the output from programs. Some of these programs run on the server. Typically these programs are stored in the /cgi-bin directory on the WWW server. When invoked, they generate output that is then returned across the network to the browser. These "cgi-bin" programs can be used to handle the input from WWW on-line forms and interface with database software.

Other programs run on the user's (client's) computer. Typically these are written in JAVA^{TM} or ActiveX^{TM} languages which incorporate special security features that limit what they are allowed to do on the client. This security is necessary because few of us would want to use the WWW if any link could download a program that would delete all our files! The relationship of the server and client (user) systems is displayed in Figure 2.

Figure 2. The relationship of the server and client (user) systems.



In deciding whether to use an existing server or create a new server, a critical issue is whether or not you need to use cgi-bin scripts. Most university and commercial servers restrict the use of server-side scripts because these place additional demands on the server computer and can create security holes.

SEARCHING AND FORMS ON THE WWW

The use of a WWW server is facilitated by tools that allow users to search for information without knowing where it is located on the server. There are several approaches that all involve using cgi-bin scripts (server-side programs) to access pre-computed indices. One approach uses full-text search engines. These create an index of all the words used in each WWW page. Searches yield links to all pages that contain a given search term. Examples of this type of search engine are WAIS (Wide Area Information Server), Glimpse/WebGlimpse^{TM} and Excite^{TM}. Other types of search engines permit the user to search only certain fields (e.g., search for term only in title) such as Z39.50-compliant search engines and relational databases.

The utility of the WWW is greatly enhanced by using web pages (i.e., web forms) to collect information as well as to distribute information. HTML includes ways to create on-line forms. These forms feature a variety of different options for input, including text fields (both single line and text blocks), radio-buttons, check boxes and selection bars. The output from a form is an encoded string that includes the name of each field and the value that the user assigned to that field. Form output can be decoded by cgi-bin programs to create new data files or to interact with database software. At the Virginia Coast Reserve LTER site (*http://www.vcrlter.virginia.edu*) WWW forms are used to allow researchers to add entries to the site research calendar, research abstracts, and annual reports. When coupled to a database, forms are used to update entries in the site personnel directory, biodiversity database and data catalogs. This greatly aids in developing scaleable solutions to common problems of site management by eliminating an input bottleneck. Any investigator with access to a WWW browser has the ability to update the databases.

SOURCES OF INFORMATION AND SOFTWARE

- HTML

http://www.utoronto.ca/webdoes/HTMLdoes/NewHTML/htmlindex.html - HTML table of contents

- Server Software

http://www.apache.org - Apache Server

http://hoohoo.nesa.uiue.edu - NCSA WWW Server Software

http://www.tucows.com - PC Networking Software *http://www.microsoft.com* - Microsoft WWW Server and Explorer Software

http://www.netscape.com - Netscape WWW Server and Navigator Software

- Indexing and Searching

http://www.goose.ycp.edu/~vkline/dragnet.html - Review on cataloging with lots of links

http://www.glimpse.cs.arizona.edu/webglimpse/index.html - Web Glimpse Search Engine

http://www.excite.com/navigate/ - Excite Web Site Search Engine

WEB-BASED DATA MANAGEMENT

Matthew B. Jones

National Center for Ecological Analysis and Synthesis, 735 State St., Suite 300,

Santa Barbara, CA 93101

Abstract. Data management techniques for integrating World Wide Web (web) publishing with data storage systems are important in furthering innovative and powerful insights in ecology, mainly through improved data exchange and collaboration. This chapter discusses the benefits and disadvantages of using the web for data management applications, and describes four categories of technological solutions that can be employed to integrate database systems with web distribution.

INTRODUCTION

Data management practices at the National Center for Ecological Analysis and Synthesis (NCEAS; *http://www.nceas.ucsb.edu*) are used to promote our mission of "advancing the state of ecological knowledge." Many projects at NCEAS involve the acquisition, synthesis, and analysis of data from multiple, distributed data sources, as well as remote collaboration on projects before and after events are convened physically at NCEAS. Several other ecological groups have verified the need for remote access to large and long-term ecological data sets (Gross et al. 1995, Michener et al. 1997). To promote the widespread sharing of ecological data, and to improve remote collaboration activities among our research scientists, NCEAS has researched, developed and is implementing a variety of techniques for managing data using the World Wide Web.

Objectives of this chapter are to: 1) present an overview of the costs and benefits of developing and using web-based tools for data management; 2) examine several common technical approaches to integrating databases with web access; and, 3) conclude with some implementation guidelines that have proved useful in systems developed at NCEAS.

BENEFITS AND DISADVANTAGES

The relative merit of using the web for data management applications can be distilled to a tradeoff between the portability and accessibility of Internet-deployed applications versus the potential costs of developing and maintaining web services at a site (Table 1). Web-based applications generally work across all computing platforms in an organization, either through a common web browser client, or through Java® applets. Both of these approaches can be deployed locally at a site, or they can be made accessible over the Internet for the broader community to access. In addition, using the web as the client interface for data management tools allows developers greater flexibility in their choice of scalable database solutions deployed on the server because interface considerations are divorced from storage and application logic needs. This separation of interface from application logic and data storage also permits application developers to migrate to new and improved technologies on the server as they become available without any change in the client-side interface that the user experiences. Finally, using the web for data management affords the obvious advantages of easing the process of sharing and distributing data with collaborators, the general public, and archive centers.

Although there are clear benefits to using the web for data management applications, several potential disadvantages also exist. First, the web is a relatively new and immature technology, and so the development tools available for creating web applications are, compared to other development areas, feature-poor and difficult to use. For example, the types of Rapid Application Development (RAD) tools available to C/C++ programmers are just beginning to emerge for web and Java® based development. Another facet of the technology's immaturity is reflected in the simplicity of the interfaces that one can build using HTML forms. Developers are limited to a small set of graphical widgets for use in presenting a user interface. Finally, like other complex technologies, web-based applications can require a high investment in software, as well as maintenance costs for the software and personnel for software administration.

Table 1. Costs and benefits of web-based data management.

Benefits	**Costs**
Cross-platform interfaces	Potentially high development and training time
Internet deployable	Potentially high software investment
Scalable database backend	Maintenance costs
Independent interface allows database migration	Interface simplicity / immaturity
Easier data sharing, interchange, and archive	

APPLICATIONS

The uses of web-based solutions in ecological data management are many and varied. One of the most obvious is to use the web as a mechanism for the distribution of existing data sets and their associated metadata. However, one can also use interactive web applications to create data entry forms for the collection of metadata and data, and to query and retrieve metadata. Data stored in a database can be converted to structured markup languages for interchange of data (e.g., XML) and for data presentation (e.g., HTML). Sophisticated query and visualization tools can be developed that give users a mechanism to remotely query data, find the subset that interests them, and then perform remote processing operations on those data. Remote processing functionality that is of interest to users includes quality control processing, data subsetting and aggregation, generation of descriptive and summary statistics, and generation of graphics for data visualization. Providing these simple analytic and visualization tools via a cross-platform, simple interface like the web empowers users to explore and use data that otherwise might be inaccessible.

WEB-DATABASE INTEGRATION TECHNIQUES

Overview

A wide variety of techniques exist for implementing the communication and data transfer mechanisms between web servers and data storage systems. The web is a client/server paradigm, so there is a tension between centralization of functionality at the server and distribution of functionality to clients. The most prevalent software solutions today do essentially all processing on the server side, and leave the clients with user interaction and display of information. However, recent advances in programming technologies (i.e., Java®) have blurred the roles of the client and server and promise to permit more advanced processing on client computers in a portable fashion. Although a number of platform or operating-system specific solutions exist, I have concentrated here on technologies that can be implemented cross-platform because the web was designed as a platform-neutral communication mechanism.

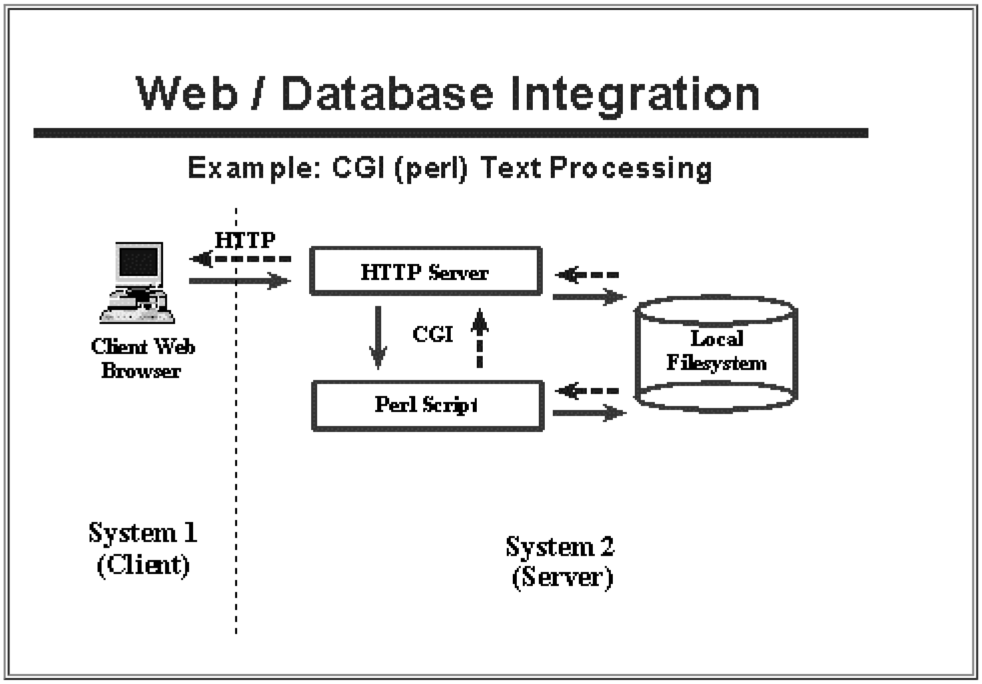
The techniques for web integration that are commonly employed can be broken into four general classes: 1) ASCII-oriented solutions; 2) Template parsing solutions; 3) Transaction monitor (middleware) solutions; and, 4) Java® applet solutions. The following sections will outline the basic features and benefits of each approach. Portability, scalability, ease of deployment, interface maturity and flexibility, cost, and client-processing capability are all considerations in evaluating the appropriateness of each technique for a particular application.

ASCII-oriented solutions

The widespread adoption of the ASCII standard as a universal character set has obvious advantages in terms of cross-platform portability and ease of deployment. The simplest case of web-based access to a data store, and one used as a foundation in many of the other techniques, is the delivery of static text documents from server to client over the Hypertext Transfer Protocol (HTTP). Web browsers (i.e., HTTP clients) are generally built to interpret and format the special type of ASCII documents known as HTML (Hypertext Markup Language) documents, but can in fact receive any type of data using this transport mechanism. To increase control over how data are delivered, many implementations add a processing script as an intermediary between data files and web server delivery of those files, making the process dynamic. The processing script can perform a number of tasks, including query processing, generation of formatting information like HTML code, data aggregation, etc. The script that executes often works by examining an ASCII text file that contains the data to be searched or processed. After determining which data are appropriate, the script formats the information and returns them to the web server. These processing scripts generally conform to the Common Gateway Interface (CGI), a standard that defines the mechanisms by which a web server can execute and communicate with processing scripts (Gundavaram 1996). CGI scripts can be written in most languages, including perl (common on UNIX), C^{TM}, C++^{TM}, Visual Basic^{TM}, and many others. CGI is simple to implement, inexpensive, and fairly easy to maintain, but generally does not scale well as the transaction load increases. In addition, the user interface elements available through the HTML "forms" specification are sometimes limiting, as is the lack of client side processing when CGI is used on the server.

Figure 1 illustrates the architecture of typical client/server transactions involved in delivering data via the web and CGI mechanisms. A web client requests, either via a URL or an HTML form, a set of data from the web server (solid arrow). The web server finds the requested file itself from the filesystem and returns it to the client (dashed arrow), or it executes a script, passing query information to the script via the CGI mechanism. The script executes and retrieves information from the filesystem according to the query parameters it received from CGI. When processing completes, the script sends the data (usually formatted in HTML) back to the webserver via CGI, and the webserver in turn sends the data back to the client that first made the request. This is a 2-tier client server solution, as the client and server generally reside on two different hosts.

Figure 1. Architecture of CGI text processing solutions

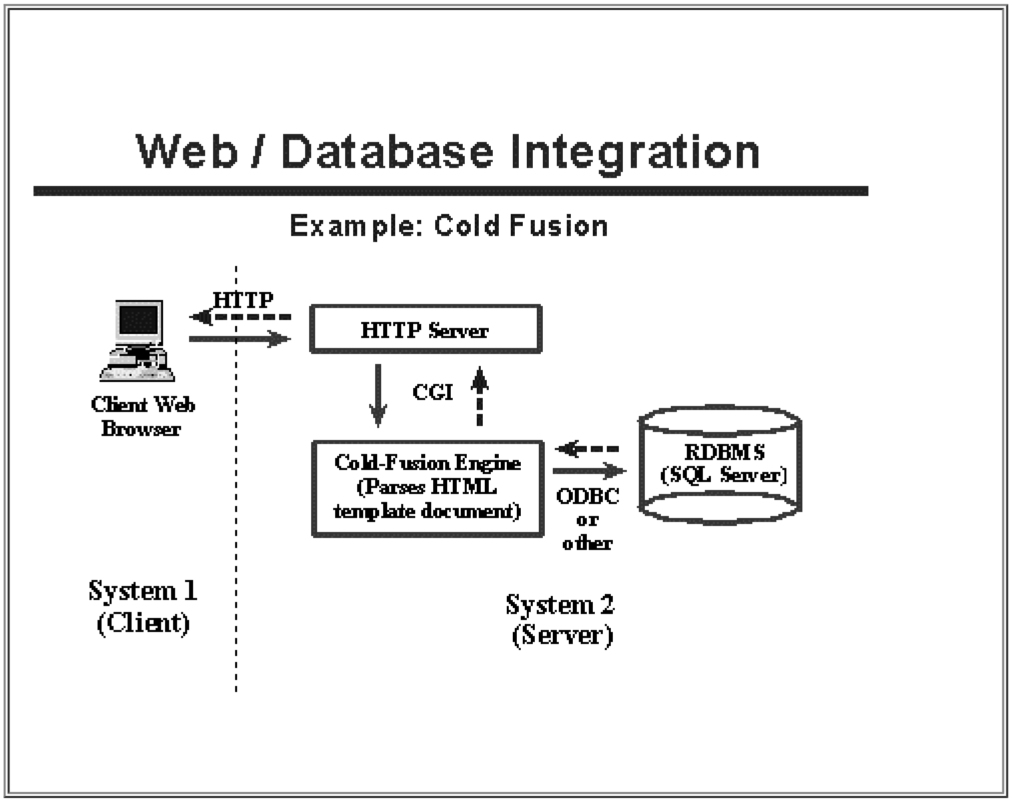


Template parsing systems

To improve access to database systems, several vendors have created systems for directly embedding database specific commands into HTML and other text files. A vendor-supplied program parses the HTML file and extracts the embedded commands, sending a database query to a database management system. The results that are returned from the query are interspersed in the HTML file according to formatting instructions, and the dynamically generated results are returned to the web server, which sends them to the client. This mechanism is similar to the ASCII database solutions described above, except that a proprietary language is used to embed commands in the HTML file that drives the query and formatting processor. Also, template parsing systems generally connect to relational database systems, and therefore they have the advantage of simplifying database integration. They are easy and powerful mechanisms for accessing a database, but generally lack scalability, don't contain the procedural functionality of more generic programming languages, and still are limited by HTML form interface elements and a lack of client-side processing. Examples of these systems include Allaire's Cold Fusion® and Microsoft's Internet Information Server® / SQL Server® combination; the Allaire product has the advantage of working with any Open Database Connectivity (ODBC®) compliant database -- a database interoperability standard -- and any CGI compliant web server, rather than being limited to specific products.

The architecture of template parsing systems (Figure 2) is similar to CGI / ASCII database systems. Again, a web client requests information and the web server passes the information via CGI to the template parsing program. The template parser retrieves the HTML template with embedded database commands, parses out the commands, and then makes a database connection (often ODBC) in order to execute those commands. The query results returned from the database are formatted by the parsing program and returned to the web server, which returns the dynamically generated document to the client.

Figure 2. Architecture of template parsing systems.


Web / Database Integration

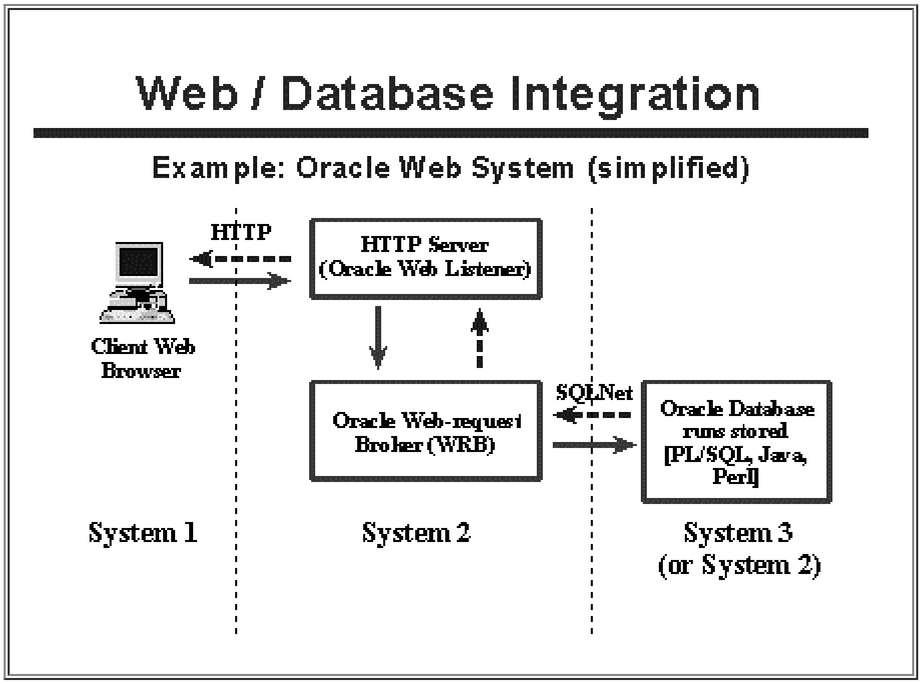
Example: Cold Fusion

Transaction monitor systems

A further extension of these concepts arises in the class of solutions called Transaction Monitor (TM) Systems (sometimes called "middleware"). Transaction Monitor software usually implements a 3-tier architecture where the client and database each reside on different systems than the transaction monitor, and the transaction monitor plays the role of mediating transactions between the requesting client and one or more data providers that can be distributed across multiple other hosts. This architecture is extremely flexible and scalable because it allows many backend database systems, each potentially running different database software, to participate in a transaction over the web. In addition, the transaction monitor can actively poll the available server systems and determine which has the most available processing resources, thereby increasing performance and distributing computational load across the server database systems. Examples of systems that can implement a transaction monitor system include Oracle's Web System®, and Microsoft's Transaction Server®.

An example transaction monitoring architecture is illustrated in Figure 3. As usual, a web client makes a connection using HTTP to a server, which then launches the transaction monitoring (TM) software. The gateway between these systems can be CGI, but more often it is a proprietary interface that maximizes performance. The TM system then distributes query requests to one or more relational database systems on the same or different hosts (n-tier). Again, the gateway between the TM system and database systems are generally high-performance, proprietary drivers provided by the TM system. In addition, the database systems themselves often store the application logic and formatting instructions in stored procedures, rather than having to parse text transmitted via the web server gateway (e.g., CGI).

Figure 3. Architecture of transaction monitor systems.

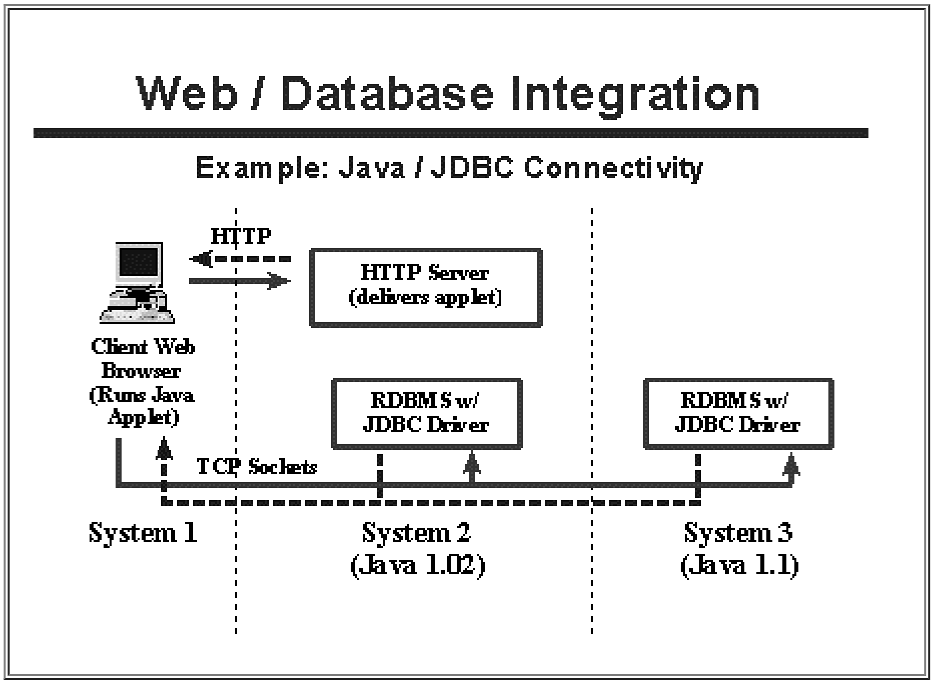

Web / Database Integration
Example: Oracle Web System (simplified)
HTTP
HTTP Server (Oracle Web Listener)
Client Web Browser
SQLNet
Oracle Web-request Broker (WRB)
Oracle Database runs stored [PL/SQL, Java, Perl]
System 1
System 2
System 3 (or System 2)

Integration using Java® and JDBC®

The Java® programming language has a library called "Java Database Connectivity®" (JDBC), which provides a platform and database independent programming interface to access multiple distributed databases of varying types. In using this system for integrating databases with web sites, one develops Java® applets that are delivered over a web connection (HTTP), and then the applets execute on the client machines. This mechanism is by far the most flexible because it allows the programmer to design an n-tier database system with connections to many database systems, all without specialized, expensive middleware software. Because the applet runs on the client machine, it allows full freedom in client side processing for field validation and interface fine-tuning. There are two principal disadvantages: Java® is a lower level language than others described here and therefore is substantially more complicated to use for interface development; and Java's performance is still much lower than many natively compiled interface-building systems. However, for most interface activities, performance is not particularly demanding and Java® will usually allow more responsive interfaces than HTML does. Other systems like Microsoft's ActiveX can be used to implement similar systems, but they lack the basic advantage of all of the systems described here: interoperability across virtually any operating system. Java® applets are employed in this way in several commercial systems, including SAS' Intr*Net® product and many middleware systems such as Symantec's dbAnywhere®. JDBC® drivers exist for most major database systems, including Oracle®, Sybase®, SQL Server®, and others.

A typical architecture for using Java® and JDBC starts with a web client requesting an HTML page that has an embedded Java® applet (Figure 4). The web server delivers the applet to the requesting client (a potentially time-consuming process), and then the client executes the Java® applet, ending the interaction with the web server. The Java® applet uses JDBC calls to open up separate TCP connections to one or more relational database systems, independent of the web server. It then communicates with these database systems using JDBC calls to query and update data, while displaying the results in a custom developed user interface. This type of mechanism allows substantially more flexibility in implementation than any of the other systems, at the cost of development time. The complexity of designing application logic for a Java® applet to manage one or more database connections and an easy-to-use interface should not be underestimated, but neither should its potential power.

Figure 4. Schematic of a Java®/JDBC architecture.


Web / Database Integration
Example: Java / JDBC Connectivity
HTTP
HTTP Server (delivers applet)
Client Web Browser (Runs Java Applet)
RDBMS w/ JDBC Driver
RDBMS w/ JDBC Driver
TCP Sockets
System 1
System 2 (Java 1.02)
System 3 (Java 1.1)

CONCLUSIONS

The variety of mechanisms described here allow everything from simple, easily implemented web-database communication to high end, scalable solutions for critical applications. The categorization that I developed was a means of simplifying a continuum of overlapping, non-exclusive technological solutions, and should be interpreted as such. For example, many transaction monitor systems may use CGI gateways, and Java® solutions may make more HTTP connections than indicated. Nevertheless, the basic features of those systems are used as indicated.

When designing a mechanism for web-database integration, one must weigh the relative strengths and weaknesses of the different approaches outlined above for a particular application. If the application is relatively local in scope and small in scale, it will probably be simplest to use the CGI-based ASCII approaches. For more complex applications, and for applications where scale and performance are critical, some of the more complex approaches outlined here, such as transaction monitor systems, may be appropriate. Finally, where substantial control of processing on the client computer is needed, and where portability across computing platforms is important, custom-designed Java® applications and applets become beneficial choices.

In implementing and researching these solutions, I have found a number of general guidelines useful to keep in mind across all of the systems. First, as soon as one attaches a computer to a network, and especially when one provides access to data over the Internet, security concerns arise. Writing both CGI scripts (in any language) and Java® programs has inherent risks; one must carefully examine the mechanisms by which user input is validated and checked before it is used to execute programs on the server system, or you may inadvertently grant full access to a database or operating system (see Garfinkel 1997). Second, although some aspects of web-database integration seem simple, full scale integration is much more difficult to design and implement; conservatism in estimates of development time help to make projects successful. Designing a modular system in which each module has utility before the entire system is completed can help in this regard, as well as making it easier to upgrade modules as new technologies arise. Third, all of these mechanisms for integration allow a clean separation of user-interface from data storage; by designing your applications this way you can upgrade backend storage systems when the need arises without impacting the user's method of interacting with data.

In the end, these technologies are only useful to ecological data managers when they improve the quality of science in the discipline or open up new areas for research. At NCEAS we hope that the integration and synthesis of data will allow new insights into the structure and function of ecological and evolutionary systems. Our development of data management technologies is guided by our need to synthesize data from multiple sites, or data arriving in many formats, as well as a desire to exchange data with colleagues. This paper represents a synthesis of technological solutions that ecological data managers may find useful in their own efforts.

ACKNOWLEDGMENTS

The ideas in this paper were substantially improved through discussions with Mark Schildhauer. This work was funded by the National Center for Ecological Analysis and Synthesis, a Center funded by the National Science Foundation (Grant #DEB-94-21535), the University of California - Santa Barbara, and the State of California.

LITERATURE CITED

Garfinkel, S. and G. Spafford. 1997. Web Security & Commerce. 1st Edition. O'Reilly and Associates, Cambridge, MA.

Gross, K.L., C.E. Pake, E. Allen, C. Bledsoe, R. Colwell, P. Dayton, M. Dethier, J. Helly, R. Holt, N. Morin, W. Michener, S.T.A. Pickett, and S. Stafford. 1995. Final report of the Ecological Society of America Committee on the Future of Long-term Ecological Data (FLED). Volume I: Text of the report.

Gundavaram, S. 1996. CGI Programming on the World Wide Web. 1st Edition. O'Reilly and Associates.
Michener, W.K., J.W. Brunt, J. Helly, T.B. Kirchner, and S.G. Stafford. Non-geospatial metadata for the ecological sciences. Ecological Applications 7:330-342.

VIRTUAL WORKING GROUPS AT NCEAS: USING THE WEB

TO FACILITATE SCIENTIFIC COLLABORATION

Mark P. Schildhauer

National Center for Ecological Analysis and Synthesis, 735 State St., Suite 300,

Santa Barbara, California, 93101

Abstract. Advances in computing and networking are creating new ways for scientists to engage in long-distance collaborations. This chapter describes how the National Center for Ecological Analysis and Synthesis (NCEAS) is using the World Wide Web and other network services to enable *ad hoc* teams of ecologists to share ideas and data. It also discusses how technological developments in the near-future will further increase ecologists' abilities to interact using the Internet.

INTRODUCTION

Ecologists are rapidly becoming aware of the advantages of using internetwork technologies in the service of their science. In less than ten years, email has risen from being an exotic high-end application to being almost universally adopted for routine communications among scientists. Within the last five years, the World Wide Web (WWW, web) has become an extremely popular and convenient 'place' to present scientific resources, including data, preprints, and references. Currently, there is intense interest in using the Internet for easy querying and access of ecological data. Ecologists can further benefit by broadening their perspective on internetwork technologies, especially regarding the possibilities for doing effective collaborative science.

Software that supports collaboration is categorized as *groupware* or *CSCW*-- for 'computer-supported cooperative work'. CSCW researchers are addressing issues that facilitate computer-based collaboration, including location independence, synchronous (real-time) and asynchronous communication, coordination within and among groups, and information workflow.

CSCW is a rapidly developing area, and many of the specific solutions discussed in this paper will be outmoded within several years by more robust, easy-to-use, and comprehensive alternatives. Nevertheless, there are already immediate advantages that can be attained by adopting groupware approaches. Specific groupware solutions currently in use at NCEAS are the focus of this chapter. These are largely workaround solutions that involve imaginative deployment of standard services. The intent of this chapter is to familiarize ecologists with the potential and basic concepts of groupware, as well as the technological underpinnings for these services. The tone is non-technical, so that the contents will be understandable to any ecologist with only minimal background and interest in Internet technologies.

PROBLEM

A primary mission of NCEAS (*http://www.nceas.ucsb.edu*) is to facilitate integrative and synthetic research in the ecological and environmental sciences. NCEAS supports collaboration among groups of scientists with complementary and cross-disciplinary expertise. NCEAS-sponsored groups may be comprised of a few to several dozen individuals, who are associated with a diversity of institutions. These scientists might not normally engage in close collaborative efforts. Since NCEAS does not support the collection of new raw data, these ventures also frequently require collection and collation of relevant information from many sources. These newly aggregated data then serve as the basis for analyses that will hopefully lead to novel, integrative insights.

A challenge to the NCEAS' computing team was how to enable dozens of *ad hoc* working groups of scientists to effectively collaborate--prior to, during, and following any workshops or conferences at our physical facilities in Santa Barbara, California. We had to consider that our potential clients represent the entire national (and in many cases, international) community of ecological research scientists. These individuals have access to highly varying levels of computing, technical support, and network bandwidth at their home institutions. We also had to accommodate the need for scientists to work comfortably with familiar technological tools at their home sites, while providing a centralized service linking them together as a collaborative effort through NCEAS.

SOLUTIONS, 1996-1997

Development of a 'shared virtual working environment' was approached through deployment of readily available software, in conjunction with special configurations of computing servers. For example, these services allow anyone with a standard WWW browser, coupled with access to email and ftp, to participate in a collaborative networked research environment, that allows for efficient dissemination, updating, and browsing of data files, routine communications, analytical approaches, formatted works-in-progress, and supportive graphics. Before describing these mechanisms in greater detail, we provide some background as to our design requirements, and approaches to solving these issues for a large, diverse, distributed user base.

Prerequisites: network bandwidth and computational power

We started with what was feasible for the end-user, given the current level of Internet access and computer expertise of research ecologists circa January 1996, when NCEAS started hosting groups of collaborating scientists. We had to make assumptions as to computational power as well as the network bandwidth available to a typical individual in the academic or governmental sectors, since most of our clients are research ecologists working at universities, field stations and research laboratories, and federal or state agencies.

It was assumed that individuals have reasonable connectivity with the Internet-meaning that most have 24-hour access from their research offices, with faster than modem-speed (approx. 28.8 kilobits/sec) bandwidth. In fact, we were aware that a few individuals would only have intermittent modem-level access--especially those working in remote field sites, but the great majority of scientists coming through NCEAS are from institutions with at least T1-level connections to the Internet (Table 1).

Table 1. Suitability of different network bandwidths for varying types of applications, assuming that networks are not congested. The relative speed of '1' for the modem is based on the currently popular 28.8kbps speed. Each successive row encompasses all the functionality of lower speed access types.

Access type	**Speed**	**Task suitability**	**Typical location**
Modem	1	email, simple WWW	Home phone
T1	54	graphical WWW, small data transfer	Remote office
Ethernet	347	Limited videoconferencing, shared file systems	Existing LAN
Fast Ethernet	3,470	streaming video, large data access	Emerging LAN
Gigabit Ethernet	34,700	multi-channel multimedia, live data feed	LAN in 3-5 years
ATM/SONET: OC3-OC48	5,382-86,111	Integrated voice/video/data; guaranteed Quality of Service (QoS)	Next generation Internet; LAN

			in 5-10 years?

Our problem is complicated by the dispersed distribution of our clientele. This means that although individuals might have good connectivity within their local area network (LAN), severe bottlenecks can occur anywhere between their desktop and the ultimate destination of NCEAS' servers. High-speed access to the Internet, however, is a necessary prerequisite for advancing groupware approaches among *distributed* (remotely located) individuals. Luckily, current interest in the Internet among the US federal government (*http://www.ngi.gov*), academicians, and the commercial sector, enables us to forecast that next generation internetworking initiatives, currently exemplified by 'experimental' networks such as Internet-2 (*http://www.internet2.edu*) or the vBNS (*http://www.vbns.net*), will provide vastly increased bandwidth to most academically-based ecological researchers within the next several years (Table 1).

We also had to make assumptions about the level of computational power routinely available to NCEAS' clients at their home institutions. It was decided that most individuals were familiar with email software and a graphics-based WWW browser (e.g., Netscape® as opposed to Lynx), and that this was an increasing trend. Given this supposition, we developed solutions requiring minimal configurations roughly equivalent to an Intel 486/DX2-66 running Windows 3.1, or PowerMac at 66MHZ running MacOS System 7 or higher (Table 2). The machines would need a minimum of 8MB of RAM, and ideally 16MB or more. These machines would also need a TCP/IP networking stack installed on them, to work through an Ethernet card or via SLIP or PPP through a modem.

Our minimal system configuration represents technology that would have been common on a scientist's desktop circa 1994, i.e., about two years prior to our launching of NCEAS' collaborative work areas. We expect a large part of our client base has since upgraded their desktops to more powerful systems.

Table 2. Computational power available to typical ecological researcher, using PC's as an example, and assuming cost of approximately $2500 at time of purchase. Intel power comparisons are estimated from iComp 2.0 measurements (*http://www.ideasinternational.com/benchmark/bench.html*).

Computer	**CPU and RAM**	**Relative Power**	**'Purchase' Date**
PC	Intel 486- 66 MHz, 8MB	1	Early 1994
PC	Intel Pentium- 133 MHz, 16MB	3.7	Early 1996
PC	Intel Pentium II- 300 MHz, 64MB	11.2	Early 1998

Client/Server model

The client/server model forms the basis for all currently popular Internet services, and was the model we turned to for providing services to NCEAS' distributed user-base. *Servers* answer the 'requests' for service from a potentially large number of client systems at any given time. If these are critical services, the more powerful hardware at a site will be allocated for these purposes, with the expectation of accurate and reliable service backed by a trained professional staff. Server software usually requires considerable expertise in order to be properly configured. This unfortunately prevents most individual ecologists from setting up systems that would enable them to easily 'publish' on the WWW, or even facilitate simple file transfer among remote colleagues without compromising their own system's security. *Client* systems, on the other hand, run software that is relatively easy to install and learn, but only useful if capable of connecting to a server. So, e.g., while many scientists are finding that it is quite simple to upgrade and configure their

Web browser, they are still unable to implement some very effective server-side functions unless provided by local systems administrators. In essence, the client/server model enables an organization to leverage the capabilities of a few high-powered, secure, and well-tuned servers to deliver information to a broad user base.

Any computer can function as a server if the appropriate software is installed. But on today's Internet, most of the large, powerful servers are UNIX systems, with multiple CPUs, true multitasking and multi-user capabilities, and highly optimized throughput. The main services provided by today's Internet servers include email, such as the SMTP-based 'sendmail' included with virtually every UNIX system; Web serving via programs such as Apache or Netscape FastTrack^{TM}; and less obvious but critical functions such as the Internet Domain Name Service (DNS), which enables clients to use computer names (hostnames) rather than numerical IP addresses to locate other machines on the Internet.

The rapid and continuing success of the Internet over the past decade arises as a consequence of the still growing number of inter-connected servers running standard services for local groups of users (e.g., at departmental or campus levels), while at the same time easy-to-use clients for these services enable networked desktop systems to run useful 'proto-groupware' such as email or the WWW. It is also an indication of the infancy of the Internet's mainstream acceptance that servers are still relatively difficult to install and maintain. We expect that server-based services will increasingly become easier to configure on individual desktops--as computing power increases and market forces drive software design towards more simplified and automated installation. Indeed, personal WWW servers are already becoming quite common on consumer operating systems such as Windows 95 and the MacOS, and many X server software packages for the PC contain applications that enable one to set up their machine as a WWW server, telnet server, ftp server, etc.

Implementation

Estimates of access to network bandwidth and computing power typically available to ecological scientists constrained our potential solutions, as did the current immaturity of available groupware solutions. Some of the more sophisticated groupware packages required major investments of time and money to use effectively, as well as installation of specialized software on each potential client machine-thereby disqualifying them as solutions for NCEAS' needs. Furthermore, some groupware packages require that all client systems exist within a LAN, and are not yet effective for integration over a wide-area network (WAN), especially large, public-access networks like today's Internet.

Design goals

Instead of using proprietary groupware approaches, we chose to adhere to open standards, and turned to the Web as our mechanism for delivering groupware solutions. The use of standard Internet services provided us with maximal interoperability--enabling anyone with Web access to use our services, regardless of whether they are on UNIX, Macintosh or PC systems. We identified several achievable features that we believed would provide substantial benefits to distributed workgroups of ecologists. These included: assurance of privacy for the groups' materials; simplification of storage and access to the materials; exchange of richly formatted items, such as graphics and proprietary data formats; and facilitation of rapid exchange for upload and download. We accomplished all our services using freely available server software, and require only email clients and a graphical Web browser (ideally one supporting frames) for client access.

Privacy

We provided privacy to each working group via access control mechanisms that are available on many Web servers. The specific mechanisms vary from one server package to another, but these all essentially involve some systems-level configuration of an authorization file allowing computer or user-level access to specific areas on the Web server (see section 5 of the WWW Security FAQ-- *http://www.w3.org/Security/faq*, or NCSA's user authentication tutorial-- *http://hoohoo.ncsa.uiuc.edu/docs/tutorials/user.html*). We use single accounts for each working group, with everyone in the group sharing a password. This is not ironclad security (e.g., a firewall), but it provides a sufficient measure of confidentiality for scientific working groups. The user authentication mechanism also leads to a sense of closeness among the collaborators due to the focused content of the prescribed area to which all the participants are contributing.

Single, authoritative archive

We identified an archival function as critical for the groups' virtual work areas. One of the difficulties for any scientist is filing email and their attachments, shared working papers, intermediate analyses and data sets, etc., and tracking these with respect to updates and revisions. We provide a single repository for all these items, so that group members can login to their private Web area whenever they want to review information or download the latest update to the data. The onerous task of individually filing group work is eliminated. Instead, one individual from each group usually becomes the 'steward' for information placed within the private areas, and coordinates the group with respect to versioning of preprints, intermediate data sets and analyses, etc.

Closed email list

An important component of the virtual working groups is the creation of a private email list. Each collaborative group at NCEAS has its own email list. These lists only accept postings from list members, protecting the group from unwanted mailings. The list is centrally managed, so that modifications are immediately in place, and individuals don't need to maintain their own separate email aliases. Also, rather than expecting each scientist to individually file these messages, we maintain a complete archive of the mailings within each group's private work area. These archives are presented in a threaded format on the Web, so that mailings related to the same subject can be grouped together. In this way one can follow a discussion rather than browsing through the entire archive holdings for information about a topic.

We chose the procmail package with SmartList extensions (*ftp://sunsite.informatik.rwth-aachen.de/pub/packages/procmail*) for email list creation and management on our systems. It is freely available software that will run on any UNIX platform, but it can require some significant configuration of your email server. We use the procmail/SmartList combination because it is more flexible and scalable than another popular and free mailing list management package called 'Majordomo' (*http://www.greatcircle.com*).

We used the software package hypermail in order to provide threaded, HTML-formatted messages within each groups' private Web area. This free software runs on UNIX machines (http://www.eit.com/packages/hypermail/hypermail.html), and translates UNIX mail-formatted messages into HTML documents cross-referenced by subject, author, and date for threading. The threading is a convenient feature, commonly found on network news readers, that is also quite handy for grouping email, particularly by subject or sender. Ideally, correspondents pay attention to the subject line of their email, so that messages will file properly within a subject thread, or spin off on a new topic.

Rapid exchange of graphics, data, and text

Email software has grown in functionality over the past several years to include easy attachment and automatic decoding of any type of file, including non-textual materials such as graphics and executable software. Still, it is not yet efficient to send data files via email, since correspondents frequently find that they cannot decode one another's attachments. Also, email is typically delivered to an area of the mail server that may not be equipped to handle potentially large files, such as raw data or detailed image files. Both of these issues should soon resolve towards easier exchange of files via email, as email software complies more closely with the Internet MIME standard specifying how attachments are to be embedded in messages, and servers are configured to deal with large transient mail spools. But as of 1998, email is not entirely satisfactory for non-text file exchange.

Web browsers are also deficient for file exchange, due to their current lack of flexibility in uploading files. Capability to upload is built-in to the http specification that Web browsers use (PUT or POST commands), but current browser software tends to limit implementation of these functions to Web forms, and not allow unrestricted upload of an arbitrary file. For this reason, we resort to the old Internet standard *ftp*, or file transfer protocol, for file upload. NCEAS creates special areas that allow working groups to accomplish ftp uploads of data to our system. We then move these files over to the private areas following the instructions of the individual groups' data stewards. We find that Web browsers serve quite well for downloading files. NCEAS' staff provides some minor html-embellishment to make presentation of the files more informative and attractive for prospective downloading by group members.

Use of ftp also enables us to avoid the difficulties with email exchange of attached materials. For example, users of Eudora® software frequently employ (often unwittingly) an encoding format called 'binhex' to their attachments. Unfortunately, 'binhex' is virtually unsupported outside the Macintosh and Eudora realms, so non-Eudora users are often unable to decode these types of attachments. By using ftp, our scientists can upload any type of file, and then use their Web browser to download any type of file. No special encoding--binhex, base-64 or other--is necessary, so no decoding difficulties arise.

We also sought a way in which scientists could exchange fully-formatted documents including text and graphics, without having to worry about everyone having the same package or version of software in order to read the file. We chose the portable document format, PDF® from Adobe® (*http://www.adobe.com/prodindex/acrobat/adobepdf.html*) to provide us with a cost-free, platform-independent way to accomplish this. Adobe's Acrobat® Reader is freely available for all the major UNIX, PC, and Macintosh operating systems. NCEAS' staff currently facilitate the translation of Word®, WordPerfect®, etc. documents into PDF by using Adobe's Acrobat software-which is not cost-free as opposed to the Acrobat Reader. The derived PDF files go directly onto the Web, where they are available for viewing by anyone with the Acrobat Reader software. One drawback to PDF is that the files are only suitable for viewing or printing- editing is not possible without additional costly software add-ins to Acrobat.

FUTURE FUNCTIONALITY

Several previously described standard Internet services were used to create 'virtual working groups' on behalf of the scientists collaborating through NCEAS. These methods had the use of the Web at their core--as a centralized and authoritative repository for email transactions, and a restricted-access location for the exchange of rich text, graphics and raw data. There are a number of other groupware functions, however, which we were unable to effectively implement given the constraints of today's networked computing environment. I will discuss these briefly here, as a preview to what ecologists can expect to become commonplace services within the next several years.

Real-time application sharing

All the capabilities described above for NCEAS' virtual working groups involve asynchronous interactions: individuals upload or download data, send messages, etc., and other participants access this information through the Web when it is convenient. But perhaps the most requested capability that we cannot yet support is that of real-time application sharing. This would enable multiple individuals to jointly view and control a program, with envisioned usage primarily that of collaboratively working through analyses using packages such as SAS® or MATLAB®.

Luckily, most robust scientific and analytical packages run optimally on computers with UNIX operating systems, where the X Window System, or X11, is the standard for graphical display (*http://www.opengroup.org/tech/desktop/x*). X Window is built on a platform-independent, client/server model, so one is able to run X-based graphical applications (clients) on remote computers as long as one's local computer has network connectivity and X-server software. X and the Motif interface comprise the standard GUI for most UNIX systems, and are available as add-on software for PC's and Macintoshes. X-based applications running on NCEAS' big UNIX servers enable remote scientists to login to our systems and run applications as if they were on-site.

The primary complication to running X applications over today's networks is bandwidth: 'best effort' service of current TCP/IP networks frequently leads to unsatisfactory performance when running over a long-haul network, such as the Internet. As described in Table 1, however, network bandwidth to scientists' desktop computers is likely to increase substantially over the next several years. This will create a situation in which remote logins to powerful computers running X Window applications will provide extremely good performance and potential for interaction. Nevertheless, X Window has already proven quite convenient to our remote collaborators in running less graphically intense jobs on NCEAS systems. Since UNIX is a robust multiprocessing and multi-user operating system, NCEAS' servers can support numbers of remote individuals simultaneously accessing and running jobs, each with graphical interfaces to applications via X Window.

There are several groupware implementations of X, which enable multiple, remote users to share a graphical session.

Notable among these is the freeware package, xmx (*http://www.cs.brown.edu/software/xmx*), which runs on UNIX systems. There are also some professionally supported X multiplexors which we did not test. Xmx allows multiple accounts to access the same X Window session, and to pass control of the mouse and keyboard among the participants sharing the application. Unfortunately, the program does not currently support connections to PC's running X server software. PC support is a planned feature, and may be available by the time this article is in print.

Multimedia Teleconferencing over the Internet

Multimedia teleconferencing comprises several valuable services that will greatly facilitate scientific collaboration over the Internet. Of these functions, NCEAS' scientists could benefit greatly from a shared whiteboard: 'live' document-sharing, in which multiple participants can draw, modify, and annotate shared graphics and text in real time. Ecologists would ideally like to accomplish this with audio contact over the network, to easily converse about the shared display (and without incurring long-distance telephone charges!). It would also be nice if this were a multi-point service, rather than 'point-to-point' (only two participants). The concept of document sharing can be extended to that of application sharing--allowing for collaborative work in popular desktop applications which often are not compliant with the X Window protocol described in the preceding section. Ultimately, we expect these features will be integrated with full multi-point live video contact and bi-directional file transfer.

These are all services for which recently defined standards promise a number of interoperable products in the near future. The relevant specifications are the International Telecommunication Union's (ITU) T.120 standard for real-time data conferencing, and the H.320 series for video conferencing (for updates on the status of these still-developing standards, see: *http://www.imtc.org/imtc*). The T.120 standard provides a common specification for application sharing, and future applications built in compliance with the standard will be a potential alternative to the shared X solutions described above. T.120 also describes how multi-point data conferences can be achieved in a vendor-neutral way. Similarly, H.320 provides standards for how audio and video are to be compressed and delivered over varying bandwidths in a vendor-neutral way. At the present time, however, there is lack of interoperability among most vendors' offerings, so that, e.g., PC users running a Window's conferencing application might be unable to communicate with Macintosh or UNIX users.

We have tested several conferencing solutions at NCEAS, but these all have some drawbacks--relative to application instability (essentially, *beta* software that crashes or performs erratically), unsatisfactory performance over a long-distance network connection due to insufficient and/or unreliable bandwidth, and lack of interoperability prohibiting a broad user base running on PC's, Mac's and UNIX boxes. Among recent releases, Netscape's Conference comes closest to providing a cross-platform solution, with support for point-to-point shared whiteboarding and audio communication for PC/Mac/UNIX. Microsoft's NetMeeting currently only runs on PC's with Windows 95 or NT.

The emergence of the ITU standards coupled with growing network bandwidth to the desktop will catalyze unprecedented growth in network-based multimedia teleconferencing over the next several years.

CONCLUSIONS

At NCEAS, we currently use the WWW to enable groups of scientists to collaborate on ecological research projects. By providing a single, private area for each group to post and discuss data and results, we relieve scientists of the task of individually tracking and filing these items. The private email list enables individuals to quickly communicate with a potentially changing group membership, again with the convenience of knowing that all messages will be archived in threaded format and available through the Web at any time for easy reference. Use of the PDF format enables scientists to share works-in-progress with full formatting and graphics, regardless of what specific software programs were originally used to generate the materials.

NCEAS' powerful UNIX servers support multiple users running multiple applications--capabilities not yet available on common desktop computer operating systems. By using X Window, remotely located scientists are able to accomplish analyses on our systems while working with a full graphical interface. Current unpredictability of network bandwidth, however, prevents remote users from running applications over the network on our systems as effectively as they might from a workstation on their own desktop. This annoyance will disappear when the next generation Internet provides ample bandwidth to academic and research communities.

Internet technologies themselves are still rapidly evolving. As network bandwidth increases and desktop computers continue to grow more powerful, more complex and effective services become possible. Standards groups continue to define and extend how interoperable solutions can be promulgated over the Internet. Groupware is one area that is likely to benefit greatly from these developments, since these solutions can be very demanding in terms of bandwidth and computational power, and become far more effective if they are capable of running on multiple platforms.

Two technology trends stand out in particular with regards to possibilities for enhancing scientific collaboration. First, the client-server model for delivery of most Internet services will continue, but individual desktops will become more capable of delivering server-side solutions as the relevant software becomes simpler to install and maintain on hardware that can deliver satisfactory performance while handling multiple clients. This will be especially true for smaller working groups seeking modest levels of service to facilitate close interaction. Second, the trend towards full multimedia teleconferencing will continue over the next several years, as standards solidify and guaranteed adequate bandwidth reaches the individual desktop.

ACKNOWLEDGEMENTS

I would like to thank Matt Jones for hours of stimulating discussion on how to provide networked computing services to ecological scientists. I'd also like to thank the conveners of this workshop--Bill Michener, Jim Gosz, Art McKee, and John Porter--for giving me the opportunity to present some of the solutions we've adopted at NCEAS. Finally, I want to express my appreciation to the scientists coming through NCEAS who were willing to discuss and test new approaches to doing collaborative work in the ecological and environmental sciences.

This work was funded by the National Science Foundation (Grant#DEB-94-21535), the University of California - Santa Barbara, and the State of California.

VISUALIZATION OF ECOLOGICAL AND ENVIRONMENTAL DATA

John J. Helly

San Diego Supercomputer Center, MS 0505, University of California, San Diego

La Jolla, CA 92093

Abstract. Scientific visualization encompasses a wide range of image generation methods, from open-ended, general-purpose software packages (e.g., AVS^{TM}, IBM Data Explorer^{TM}), to domain-specific geographic information systems (GIS). This paper provides a synoptic view of what it takes to develop meaningful, quantitatively reliable and presentable thematic images appropriate to the unique requirements of ecologists and their environmental and ecological data. It presents an overview of processing methods and resource requirements, and is intended to enable individual researchers to anticipate and plan for visualizing their research data.

INTRODUCTION

Ecological and environmental data have a variety of distinctive features making them both valuable and challenging to visualize (Helly et al. 1996, Gross et al. 1995). Chief among these is the fact that these data are irregularly and sparsely distributed in space and time. This is due to the difficulties inherent in field sampling large geographical areas over long periods of time at frequent intervals and numerous locations. These limitations are amplified by the cost associated with related laboratory analyses, and the difficulty in replicating experimental units. The development of useful quantitative images in a meaningful context is made more challenging by the need to correlate and integrate survey data with ancillary data covering widely ranging spatial and temporal measurement scales. The tools to accomplish this fall into the category of visualization and, more specifically, scientific visualization software, as a consequence of the quantitative nature of the resultant images.

To represent the range and diversity of ecological and environmental data, this paper presents three visualization projects undertaken in recent years at the San Diego Supercomputer Center (SDSC). These projects are distinguished from each other by the kind of data used to produce the images. The reason for choosing these three examples is that they span the range of strictly observational field data to strictly computer-generated data. The first example (Plate 1), bird abundance data, represents data that are sampled irregularly in space and time and contain missing values (San Diego Bay Project, *http://sdbay.sdsc.edu*). The second example (Plate 2), solar radiation data, possesses aspects of each of the other two since it contains data that are sampled regularly in time, but irregularly in space (The Solar and Meteorological Surface Observational Network (SAMSON), *http://www4.ncdc.noaa.gov/cgi-win/wwcgi.dll?WWNolos ~Product~CD-006*). The third example (Plate 3), landscape erosion, is typical of data sampled regularly in space and time with no missing values. By regular we mean that data points occur in a systematic pattern. This may be realized as a rectilinear pattern like a rectangular grid (Plate 4), or a curvilinear pattern like an ellipse. A description of the production of each image is followed by a discussion of methods for obtaining presentable hardcopy and softcopy output and the resource requirements.

THE PROBLEM OF IRREGULAR DATA

The principal problem with irregular data in visualization is the need to interpolate it onto a regular grid so that it can be displayed on a two-, or sometimes three-dimensional output device. While there are many methods for doing this, they are generally cryptic and require considerable knowledge of the underlying numerical methods to use them effectively. Recently, some software systems have been offered which greatly reduce the burden on the novice programmer (Fortner Software, *http://www.fortner.com/*) and there are extensive compilations of public domain software for the more experienced (Netlib Repository, *http://www.netlib.org*). Ultimately, numerical values must be mapped to pixel values on a screen or hardcopy. Significant distortions and inaccuracies can be inadvertently introduced into images by inappropriate use of interpolation techniques. It is important to recognize that steps involving interpolation are not always obvious to the uninitiated. Typically, some interpolation occurs explicitly under user control, however, additional interpolation may also occur implicitly within the visualization tools during the definition of object geometries, and especially during the 'rendering' process. Consider, for example, the common problem of aliasing or 'stair-stepping'. These are common interpolation artifacts interfering with the production of continuous-tone images. Proper use of visualization methods requires an understanding of how and where interpolation is used and an understanding of the type and limitations of the sampling methods used to collect the data being interpolated. Both types of knowledge are needed to evaluate the effect of interpolation on data presentation.

THE NEED FOR EXPLICIT DEFINITION OF SCENE COMPOSITION

Modern visualization packages tend to organize themselves around the type of object to render and the type of data structure needed to perform the rendering. For example, an isosurface will generally require a different type of underlying data representation, or data model, than will a volume. Most of the time spent pre-processing data for use in visualization is associated with 'shaping' the data for a particular data model. Therefore, to minimize wasted effort and false starts, it is useful to clearly define the information content of the desired scene. One should consider, for example:

- Are you interested in developing maps or displaying process dynamics? Maps are often multivariate and can be approached using scientific visualization tools or GIS (geographic information systems). Currently, process dynamics are best visualized using scientific visualization tools since they generally provide greater control over the way in which data objects are formulated and rendered, and provide functions to semi-automatically generate a sequence of related images required for an animation.
- What are you trying to show; what is your theme? Multiple themes generally require multiple color maps and legends. The depiction of discrete or categorical data will usually have different requirements than will continuous data. If you are using a map as a background it will be important to consider issues such as vertical exaggeration of relief, viewpoints, direction of lighting, and scale and resolution of thematic data relative to the underlying 'basemap'.
- Do you want to be able to measure things from your image or use it simply for illustration? Quantitatively comparing thematic values across images requires consideration of issues such as controlling data ranges for color maps between images to ensure comparability as well as image size (i.e., number of pixels in rows and columns). For maps, some projections are better suited for linear and areal measurements than others (e.g., universal transverse mercator (UTM)), and some projections are better suited for some parts of the world than others (Alpha et al. 1982, Bugayevskiy and Snyder 1995, Robinson and Snyder 1991, Snyder 1987).
- Do you want to depict a time series or a cumulative result? Time series animations usually require some type of clock indicator to inform the viewer of the location of any given scene in the series. Cumulative results may require encoding the displayed thematic data in both space and time.

GENERALIZED PRODUCTION STEPS

Regardless of the specific research goals, there are three major steps in the production of images using scientific data: acquisition, transformation (or pre-processing), and visualization. These steps are largely defined by the interfaces and processing required to obtain data from multiple sources and convert them to a suitable form consistent with common spatial, temporal and quantitative scales and the input requirements of the visualization data models.

1. The time spent acquiring, transforming and integrating data for a given scene can grow exponentially as the number of data files increases. The acquisition of basemap data (e.g., Plate 2), can be accomplished through the World Wide Web for certain types of publicly available data such as the 1-degree quad data available from the USGS web site (National Mapping Information, *http://mapping.usgs.gov/*). A great deal of other important data such as precipitation and winds can be very difficult to find for any given location due to the sparseness of the sampling stations and difficulties involved in finding the creators and maintainers of the data. The advent of digital libraries and data repositories will help to reduce some of these difficulties but these are still in developmental stages (CEED: Caveat Emptor Ecological Data Repository, *http://ecodata.sdsc.edu*; ACM Digital Library, *http://www.acm.org/dl/*).

   2. Transformation of the raw data into a form suitable for ingestion by the visualization software is an *ad hoc* process involving the use of ASCII editors (e.g., vi or emacs) and general-purpose data processing software (e.g., SAS^{TM} (Statistical Analysis System, *http://www.sas.com*), S-Plus^{TM} (S-Plus, *http://www.mathsoft.com*)).

Much of this effort is spent in quality assurance/quality control to determine data ranges, sorting, and statistical summarization into regular spatial and temporal patterns. Finally, the data are written out to files to be used for input by the visualization software.

3. How visualization is accomplished depends largely on the software and hardware available since the cost of these tools is usually quite high. Most researchers are required to make do with the resources at hand. At the time of this writing, modern visualization methods are generally executed on UNIX workstations with significant processor and storage capabilities. Visualization is highly memory-intensive since the object geometries must be largely held in memory as the image is rendered. At present, the dominant visualization tools at SDSC are IBM Data Explorer^{TM} (IBM Data Explorer, *http://www-i.almaden.ibm.com/dx/*) and AVS^{TM} (Advanced Visual Systems, *http://www.avs.com/*) for scientific applications and ARCINFO^{TM} (ARCInfo, *http://www.esri.com*) for GIS applications. There are many other tools available with their respective pros and cons, and opinions on these will vary widely.

EXAMPLE 1: MAP OF BIRD ABUNDANCE FROM FIELD SURVEY DATA

The basemap in Plate 1 is composed of bathymetry data on a rectangular grid with 50-meter spacing originally in state-plane coordinates. The coastline is described by vector data stored as ordered pairs of state-plane coordinates; both provided by US Navy. The geographic (or geodetic) coordinates and names of the sample stations were taken from reports produced by the Unified Port District of San Diego, as were the thematic bird abundance data. These data were then georeferenced in the following way. The gridded bathymetry, and non-gridded coastline data were converted to geographic coordinates using SAS^{TM}. Station names were plotted on a navigational chart to obtain their latitude and longitude pairs. The thematic bird abundance data were merged with station names. These were organized into three separate data streams from four individual flat, ASCII input files. These georeferenced data were then converted to a UTM projection to preserve the spatial accuracy of the map. Map projections were accomplished using the GCTP^{TM} (General Cartographic Transformation Package) available from the USGS (National Mapping Information, *http://mapping.usgs.gov/*). The processing flow within IBM Data Explorer^{TM} is shown in Plate 5.

Special consideration was given to the problem of color map assignment. These must be chosen such that the bathymetry data do not obscure the thematic data. The coastline was included to sharpen the land-sea boundary. The opaque circles for the abundance data were used to mark station locations as well as to emphasize the discrete nature of the observations. The translucent squares were used to emphasize continuity of habitat while emphasizing the interpolation used to obtain estimate it. Interpolation was accomplished by regridding the abundance data using a nearest-neighbor method in which the radial distance to neighbors was explicitly controlled. Regridding is a colloquialism for the more generic term of resampling used commonly in remote sensing. Detailed discussions of this and related topics can be found in *Remote Sensing and Image Interpretation* (Lillesand 1989). A plan view was chosen since the area depicted is relatively small; only a few kilometers on a side. The image was finally written out as a tiff image file.

EXAMPLE 2: GROWING SEASON DYNAMICS ON A MAP

The basemap for the images in Figures 2 and 6 was developed from USGS 1-degree DEM (Digital Elevation Model) data with the thematic solar radiation data taken from a NOAA CD-ROM [The Solar and Meteorological Surface Observational Network (SAMSON), *http://www4.ncdc.noaa.gov/cgi-win/wwcgi.dll?WWNolos ~Product~CD-006*]. The radiation data were spatially distributed on an irregular spacing at the county level for 20 geographical locations. These data were then combined into regional data set using SAS. This combining of county data into a regional data file was particularly challenging due to the large space requirements. Each county file was 33 megabytes in size due to both the number of observations and the large number of variables in each county file. These data were combined by first dropping all extraneous parameters before merging the daily latitude, longitude, and radiation values for monthly averaging.

The resulting data were converted to a common UTM (NAD83) projection after resolving a spatial registration problem resulting from the use of NAD27 for the DEM and NAD83 for the county stations. As in the first example, the data were converted to flat ASCII input files for input to IBM Data Explorer. The color map assignment was chosen to ensure an intuitive understanding of high versus low radiation values. The images (Plate 6) are not precisely comparable because the mapping of color to data values is not constant between the images. The elevation values were scaled upward by approximately 150% to emphasize vertical relief and provide convenient landmarks without imposing a grid that would interfere with the continuity of the animation.

EXAMPLE 3: LANDSCAPE EROSION DYNAMICS

Since the data in Plate 3 were all computer-generated for five dimensions (x, y, z, time, water-depth) no basemap was required. Three separate model runs were done, each with one file per time step (126, 312 and 96 files; each approximately 1 megabyte in size) and different values for model parameters resulting in different rates of erosion and topography. Similarly, georeferencing and map projections were not required. SAS was used to produce both input data files from the raw data as well as the auxiliary files needed to control the semi-automated generation of such a large number of image frames.

Color maps were chosen to emphasize dry versus wet and to aid in the perception of ridges versus valleys. The view was assigned to emphasize valleys versus ridges and to resolve figure-ground perception difference between two of the investigators. A particularly thorny problem emerged in that as erosion progressed from frame to frame, an obscure parameter in the software was causing the image to be rescaled, which in turn caused the scene to translate vertically on the screen as time progressed due to misregistration between frames. The logarithm of water-depth was chosen to accommodate a range of water depth values of approximately 14 orders of magnitude. Each image frame was successively written to the file system for importing into the video production process used for the animation.

IMAGE OUTPUT

Modern image processing methods have evolved to the point where numerous image file formats can be inter-converted (San Diego Supercomputer Center Image Tools. San Diego Supercomputer Center, (Available via anonymous ftp from ftp.sdsc.edu (132.249.20.22), 1998)). However, there are important differences within image file standards that can be quite puzzling and problematic. There is still the fundamental difference between raster (e.g., tiff, gif, jpg, png) and non-raster (e.g., postscript, hpgl) formats which typically make it possible to convert from raster to non-raster formats effectively, but not routinely in the other direction. As a rule-of-thumb, it is generally safe to rely on the uncompressed, RGB tiff file format as your default choice. There is also a CMYK (i.e., Cyan-Magenta-Yellow-blacK) tiff format used chiefly for offset printing. From this type of image file virtually any desired hardcopy output type can be obtained. This will also be the largest in size so it is sometimes awkward to move it around and process it. At present the second choice is generally jpg which is, however, a compressed format. This is acceptable for many applications, but does not generally contain all of the original image information since it is a 'lossy' compression method. This means that it loses bits through an encoding scheme to save space. There are also 'loss-less' compression methods. There are many graphics service bureaus that can provide hardcopy output as 35mm slides or other professional quality output media beyond what the commonly available printers can provide. Animation file formats are also semi-standardized on MPEG and AVI formats.

SUMMARY

In this paper, examples of visualizing ecological and environmental data have been presented. Each represents an approach to developing meaningful, quantitatively reliable and presentable thematic images. Moderately technical descriptions of processing methods and resource requirements, and information necessary to *plan* data visualizations were provided. Detailed implementation mechanics and, perhaps most significantly, problems associated with the introduction and propagation of map errors were not discussed but can be further investigated in *Environmental Modeling with GIS* (Goodchild 1993).

LITERATURE CITED

Alpha, T.R. and J.P. Snyder. 1982. The properties and uses of selected map projections. U.S. Geological Survey, Reston, VA.

Bugayevskiy, L.M. and J.P. Snyder. 1995. Map projections: a reference manual. Taylor & Francis, London, UK.

Goodchild, M. 1993. Environmental Modeling with GIS. Oxford University Press, New York, NY.

Gross, K.L., C.E. Pake, E. Allen, C. Bledsoe, R. Colwell, P. Dayton, M. Dethier, J. Helly, R. Holt, N. Morin, W. Michener, S.T.A. Pickett, and S. Stafford. 1995. Final report of the Ecological Society of America Committee on the future of long-term ecological data (FLED). Volume I: Text of the report. (*http://www.sdsc.edu/~ESA/FLED/FLED.html*).

Helly, J., S. Levin, W.K. Michener, F. Davis and T. Case. 1996. The State of Computational Ecology. San Diego Supercomputer Center, San Diego, CA.

Lillesand, T.M. 1989. Remote Sensing and Image Interpretation. John Wiley and Sons, New York, NY.

Robinson, A.H. and J.P. Snyder. 1991. Matching the map projection to the need. American Congress on Surveying and Mapping, Bethesda, MD.

Snyder, J.P. 1987. Map projections--a working manual. U.S. G.P.O., Washington, DC.

INFORMATION ACCESS AND DATABASE INTEGRITY AT THE NORTH TEMPERATE LAKES LONG-TERM ECOLOGICAL RESEARCH PROJECT

Barbara J. Benson and Maryan Stubbs

Center for Limnology, University of Wisconsin-Madison,

680 N. Park Street, Madison, Wisconsin 53706

Abstract. The North Temperate Lakes LTER data and information system is designed to facilitate ecological research. The primary information management goals focus on information access and database integrity. The use of relational database software is discussed in the context of these two goals.

INTRODUCTION

The North Temperate Lakes Long-Term Ecological Research (NTL-LTER) program (Magnuson et al. 1991) was established in 1981 by the National Science Foundation as part of the LTER network of research sites (Callahan 1984, Swanson and Franklin 1988).The NTL-LTER project has two study areas in Wisconsin where patterns, processes, and interactions of lakes and their surroundings are examined at a nested set of spatial and temporal scales.

Data management is an integral component of the NTL-LTER project (Benson 1996).The design of an information system must be based upon the research agenda. NTL-LTER has research goals (Table 1) which require diverse data sets, linkages among data sets, and multiple spatial scales. Expansion of our research agenda has included regional-scale investigations and the study of human interactions with lake ecosystems.

Table 1. NTL-LTER research goals.

1) Perceive long-term changes in the physical, chemical, and biological properties of lake ecosystems.

2) Understand interactions among physical, chemical, and biological processes within lakes and their influences on lake characteristics and long-term dynamics.

3) Develop a regional understanding of lake ecosystems through an analysis of the patterns and processes organizing lake districts.

4) Develop a regional understanding of lake ecosystems through integration of atmospheric, hydrologic, and biotic processes.

5) Understand the way human, hydrologic, and biogeochemical processes interact within the terrestrial landscape to affect lakes and the way lakes, in turn, influence these interactions.

The NTL-LTER data and information system has been designed to facilitate interdisciplinary research. Our primary goals have been 1) to create a powerful and accessible environment for the retrieval of information that facilitates linkages among diverse data sets and 2) to maintain database integrity.

INFORMATION ACCESS

To provide an optimal environment for information access, we required the following criteria be met: 1) the data structures facilitate queries, 2) the client interfaces are easy to use, and 3) adequate metadata are available to permit data interpretation. We implemented our information system using Oracle^{TM} database software on a Sun Ultrasparc^{TM} 2. Other LTER sites have built successful information systems without using relational databases; however, there were some strong reasons to use a relational database for our data. The relational structure is well-suited for queries and facilitates linking data sets. Relational databases can handle large, complex data sets. Database structures can be easily expanded or changed. Normalizing data tables can eliminate data redundancy. Built-in security, recovery and export capabilities create a more secure environment for access, updates, and backup. Through the use of SQL (Celco 1995, Date 1997, Ladanyi 1997) scripts, procedures can be saved, therefore, documented and reused. Our system, Oracle^{TM} is multi-user and multi-tasking.

Industrial-strength relational databases such as Oracle^{TM} can be rather expensive to purchase unless your organization has special terms with the vendor. These products are very large and complex, with numerous configuration options, and there tends to be a significant learning curve to utilize database features fully. However, benefits include full concurrency control, recoverability, high performance, and stable vendor presence in the field.

Researchers at NTL-LTER can use an end-user query tool (Oracle Discoverer 2000^{TM}, formerly called Oracle Data Browser^{TM}) to retrieve data from the database. This point-and-click interface is being used routinely by researchers to obtain exactly the data sets of interest. Joining of tables, aggregation, and sorting can be performed with this tool.

An alternative access method is through the World Wide Web (WWW). On-line data sets can be accessed through the data catalog (*http://limnosun.limnology.wisc.edu/datacat.html*). These data sets are text files retrieved from the Oracle database with metadata at the top of each file. However, maintaining the data then requires maintaining both the database and the retrieved text files. We are now developing dynamic query capability from the WWW to the Oracle^{TM} database (Stubbs and Benson 1996). In addition to avoiding maintaining text files, these dynamic queries also permit more powerful information retrieval for the user.

Currently, we have implemented dynamic queries for meteorological data (*http://limnosun.limnology.wisc.edu/climate.html*). The user can select parameters to retrieve and specify the time period of interest. It is also possible to generate summaries over a specified time period (e.g., total precipitation or mean air temperature) or to graph a parameter over time.

DATABASE INTEGRITY

Maintaining the integrity of a database requires controlling the access for writing to the database. In addition, the database needs to be protected by an adequate backup system and be recoverable. The data in the database must have been subjected to quality control/quality assurance protocols. File format and storage media need to be addressed to guarantee useable long-term archiving.

The Oracle^{TM} database software provides considerable functionality for database integrity issues. Setting up passwords, privileges, and roles controls read and write access. Oracle export utilities can be used to backup the database and protect against accidental deletion or incorrect updating of a table or, if necessary, be used to restore the entire database.

Quality control mechanisms have been established including random blind samples and replicate analyses. The data entry software has some built-in error checking, and a two-person team proofreads entered data. Finally, researchers review summary tables and further error checks are performed, such as ion balances and calculation of critical parameters, from a redundant data set.

FUTURE DIRECTIONS AND CHALLENGES

The NTL-LTER project like many other ecological research programs is being challenged by new types and expanded volumes of data as the scope of research expands and new technology affects measurement. The expansion of the research program to include human interactions with lake ecosystems is generating the need to incorporate new types of data sets into the database. The data management staff is interacting with social scientists to design database tables and provide metadata for a growing collection of new data sets including land ownership, census, and attitude survey data. The increased volume of spatial data, especially from satellite-based sensors, requires that the ecological science community be prepared to use these data and that appropriate data management be in place.

The use of the WWW to distribute data will expand as we continue to construct query functionality from the WWW to the Oracle database. We also plan to use the WWW interface for data entry directly into the database. This interactive data entry will be designed to provide immediate feedback on entry errors.

ACKNOWLEDGMENTS

Financial support was provided by the National Science Foundation's Long Term Ecological Research Program, grant # DEB-96-32853.

LITERATURE CITED

Benson, B. J. 1996. The North Temperate Lakes LTER research information management system. Proceedings of the Eco-Informa Workshop, Global Networks for Environmental Information, November 4-7 1996, Lake Buena Vista, Florida, 11: 719-724.

Callahan, J. T. 1984. Long-term ecological research. BioScience 34:363-367.

Celko, J. 1995. SQL for smarties: advanced SQL programming. Morgan Kaufmann Publishers, Inc. San Francisco, CA.

Date, C.J. 1997. A guide to the SQL standard fourth edition. Addison-Wesley Longman.

Ladanyi, H. 1997. SQL unleashed. Sams Publishing. Indianapolis, IN.

Magnuson, J. J., T. K. Kratz, T. M. Frost, C. J. Bowser, B. J. Benson, and R. Nero. 1991. Expanding the temporal and spatial scales of ecological research and comparison of divergent ecosystems: roles for LTER in the United States. Pages 45-70 in P.G. Risser, editor. Long-term ecological research. John Wiley & Sons Ltd. New York, NY.

Stubbs, M. and B. J. Benson. 1996. Query access to relational databases via the World Wide Web. Proceedings of the Eco-Informa Workshop, Global Networks for Environmental Information, November 4-7 1996, Lake Buena Vista, Florida, 11:105-109.

Swanson F. J. and J. F. Franklin. 1988. The long-term ecological research program. Eos 69(3):34, 36, 46.

EVOLUTION OF THE KONZA PRAIRIE LTER INFORMATION MANAGEMENT SYSTEM

John M. Briggs

Division of Biology, Ackert Hall, Kansas State University, Manhattan, KS 66506-4901

Abstract. The overall objectives of the Konza Prairie LTER Information Management System are to assure data integrity (correctness, at all times, of all items in the research database), provide security for the database (protection against any loss of data), and facilitate use of data by the original investigator(s) as well as by future investigators. This program has expanded considerably from serving only a localized research group (in its original version in 1981), to its present capabilities of responding to requests for data from investigators across the globe. During this time, protocols for the development of this database have evolved with experiences gained from the growth of the research program on Konza Prairie and knowledge gained from other multi-disciplinary research efforts (especially other LTER sites). It is vital for the continued growth of a research program that the information management system be responsive to growth and adapts accordingly. This is especially true with ever-changing computer technology, and as the scientific use of the data changes over time.

INTRODUCTION

Ecological research is maturing from small-scale studies involving one or a few investigators in a single discipline for a short time period, to multi-disciplinary investigators examining regional and global patterns and processes for possible decade or longer studies. It is essential that a parallel growth in proper scientific information management also occurs (Stafford et al. 1994). This is especially true with the rapid change that has occurred in computer and network technology. The purpose of this chapter is to examine how the information management system of a field research site (the Konza Prairie Research Natural Area (KPRNA)) has developed from serving a localized, small number of independent researchers to its present capabilities of responding to requests for data from investigators across the globe.

KONZA PRAIRIE RESEARCH NATURAL AREA HISTORY

Konza Prairie was established as a research facility in 1972, primarily as a result of the efforts of the late Dr. Lloyd C. Hulbert. Initially, KPRNA included only 371 ha, but additional purchases in the early 1980's expanded the site to its present size of 3,487 ha. It was primarily established to examine the importance of fire, grazing, and climate in maintaining tallgrass prairie. The area is owned by the Nature Conservancy and is leased to the Division of Biology at Kansas State University for long-term research purposes. A watershed-level (catchment unit) fire frequency experimental design that includes replicated long-term unburned (20 yr) and annual, two-, four-, and ten-year frequencies of prescribed spring, summer, fall, and winter fires. Overlaid on this design is a grazing experiment with blocks of watersheds designated as ungrazed, grazed by native ungulates (*Bos bison*) and watersheds grazed by domestic cattle (*Bos taurus*) (Knapp and Seastedt 1998).

The Konza Prairie LTER Program

The Long-Term Ecological Research (LTER) Program of the National Science Foundation began funding research in 1980 (Callahan 1984, Franklin et al. 1990). Konza Prairie was one of the six original LTER sites selected by NSF in 1981 and is now in its fourth funding cycle (*http://climate.konza.ksu.edu/general/lter4/lter4.html*). Since it was a relatively young site in terms of ecological research compared to other LTER sites when the LTER program started, it did not have a large number of old historical data sets as did the other LTER sites. There are only two biological data sets from KPRNA that date prior to 1981 and only seven publications from 1971 to 1980. Thus, at a very early stage of the KPRNA LTER program, it was possible to incorporate sound data ecological management practices. This was based upon the desire of LTER sites not to repeat the IBP's mistake of not having an adequate data management program in support of the research program and also due to the efforts of the PIs during this early stage of development. In addition, meetings of the LTER community and NSF stressed the importance of information management at each site.

The Konza Prairie LTER information management system

During the early 1980's, considerable effort was made by the Konza LTER staff to implement a base-level research data management plan. Its primary goal was to have all interested researchers locate, interpret, and utilize data. This plan was designed using guidelines established by Gorentz et al. (1983) and is documented in Gurtz (1986). The overall objectives of that plan were to: 1) assure data integrity (correctness, at all times, of all items in the database), 2) provide security (protection against loss of data), and 3) facilitate use of data by the original investigator(s) as well as by future researchers. These simple but ambitious objectives are still being followed today (Briggs and Su 1994) and even though computers and perhaps more importantly, network technology are rapidly changing the way ecologists use and share scientific data, these guidelines ensured that the research program at KRPNA and information management system matured together over time.

The Konza LTER investigators are committed to the documentation and archival of data collected at this site. It is considered one of the most important tasks that each investigator performs as part of their effort in the Konza LTER program. The LTER program at KPRNA (KNZ) is dedicated to having all long-term data sets and key short-term data on-line and available to the scientific community and general public as soon as possible. The ultimate goal is to have all data on-line within two years of collection, processing and the completion of appropriate quality control procedures. KNZ LTER researchers have an obligation to make available all LTER-funded data to the KNZ LTER database and to publish those data in a timely fashion. They also recognized that investigators must have a reasonable opportunity for first use of data they have collected. KNZ LTER data are defined and processed for on-line access according to the protocols outlined below.

Type I. Core, long-term data sets (with associated metadata) that address Konza LTER objectives and hypotheses as outlined in LTER proposals I-IV and that are supported primarily by LTER funds. These data will be available on-line two years after data are generated and quality control is completed. We recognized that some data sets would take longer to get on-line than others due to the time required for adequate quality control or due to the demand for certain data by others.

Type II. Short-term data sets supported primarily by LTER funds, key short- or long-term data sets supported by other funding, graduate student data sets, discontinued long-term data sets, one-time surveys, etc. Data sets in the above categories that are supported primarily by LTER funds must be made available to the data manager/PIs and may be placed on-line within 2 years of completing quality control -- but only at the discretion of the data manager/PIs. Data not supported by LTER funds will be placed on-line only if mutually agreed upon by the investigator(s) and the data manager/PIs.

Our present KNZ information management system includes archived LTER data with an electronic data catalog that allows any person with internet access to browse, examine, or download any data set on-line without any restriction (*http://climate.konza.ksu.edu/toc.html*). The only requirement for individuals who use the data is to acknowledge the source of the data using the following simple format:

"Data for XXX was supported by the NSF Long-Term Ecological Research Program at Konza Prairie Research Natural Area"; where XXX is the list of data set(s) used in the publications, reports, or proposals. (Both the data access policy and the suggested formats for using KNZ data are at: *http://climate.konza.ksu.edu/intro.html*).

To reduce time and errors associated with data entry, and to maintain data integrity, over the past decade, specialized data entry programs and data checking protocols have been developed during the past decade (See Briggs and Su 1994, Briggs et al. this volume for additional information). The design of the current Konza Prairie LTER database is straightforward. All data sets are in ASCII format (with the exception of GIS coverages and satellite images; these are not on-line). Having data in only ASCII format and not in a relational database, does not permit complicated searches or subsetting of the data. Monitoring the use of data sets by researchers over the past ten years has revealed that most scientists simply want information about data sets (metadata) or simple access to the data sets. In the past, KNZ invested in developing a remote user access to our database using an ORACLE interface (Briggs and Su 1994). This was done prior to the huge explosion in WWW access. After six months of use and numerous complaints about structured query language, we simply installed a WWW server (*http://climate.konza.ksu.edu; http://climate.konza.ksu.edu*) and started using simple file structure as our database. Use of the site has grown from about two accesses a week using the ORACLE^{TM} database, to over fifty accesses a day during the month of July 1997. As with other LTER sites (see other papers by Baker, Benson, Porter, this volume), we are not only using the WWW as a place to distribute data but as a tool to inform scientists, as well as the general public, about KNZ. However, based upon recent developments in ORACLE and the WWW (Henshaw et al., this volume; Benson, this volume), we are again exploring the possibility of using a relational database such as ORACLE, but this time with an easier interface. Thus, scientists could not only access the data over the web, but as Henshaw et al.(this volume) has demonstrated, the power and utility of a relational database can be fully utilized without scientists needing to know SQL.

It has only been possible to get the KNZ data on-line because of the past commitment that KNZ had to information management. Thus, any site that is interested in information management should begin with the task of identifying the objectives of their information management system and determine how their scientists want to access the data. At present, using the WWW as an outlet for their data sets can be a simple tool, that can, over time, develop as a powerful and vital research tool. However, if the basic premise for information management is not built into a site's system at the beginning, and if the senior scientists are not supportive of the information management system, even the most powerful system is doomed to fail (Strebel et al. 1994). An information management system must involve and be endorsed by the user community it was designed to serve (Stafford et al. 1994).

For long-term security, in addition to our WWW server, we store all archived files (data files that have been entered and verified as correct by the investigator(s)) on a variety of electronic media from 1/2" magnetic tapes, 8mm tapes, hard disks, to re-writeable optical disks. Our goal is to have at least three copies of our database stored in different physical places at all times to reduce the possibility of losing data due to hardware failures, changes in computer technology or disasters. Researchers should plan on computer technology to change and try to build systems that are hardware- and software-independent.

One of the more important and useful products of our information management system has been a "Methods Manual" that details procedures for ongoing and prior studies. KNZ staff has maintained this Methods Manual since 1981 that details how each LTER data set is collected. It includes items such as precise maps of the vegetation survey, sample data sheets, and very detailed procedures on instrument installation and use. The Manual provides the necessary details to interpret the more extensive data documentation files maintained for each data set. This document is updated yearly and a completely revised manual is produced every 5 years. We have found this document to be one of the most valuable items that our research group produces.

SHORTCOMINGS

One of the most important decisions a site has to make when developing an information management system is to decide which data sets are not going to be archived. KNZ has struggled with that decision and for many years tried to document everything (from short-term experiments, graduate student work to other funded projects on KPRNA). However, due to our limited resources, we have been forced to focus only on those data sets that are funded from the core LTER grant. (See *http://climate.konza.ksu.edu/general/lter4/lter4.html* for a complete list). While results from most of these non-documented data sets are published, the resulting publications typically do not include adequate detail to be considered properly documented data sets. Properly documenting and providing access to all data sets for the entire research community is beyond the scope of KNZ staff. We recognize, though, that short-term studies, if properly documented, could be examined in the future, and study sites possibly re-sampled to address new ecological questions. Thus, any short-term data set that is properly documented and archived in reality, becomes a long-term data set. Consequently, we encourage all scientists who work on Konza Prairie to properly document their studies.

SUMMARY

KNZ has learned many lessons in developing and refining their information management system over the past 15 years. Most of these lessons parallel Michener's (this volume) "rules of thumb" for metadata and other data management recommendations (Strebel et al. 1994), but warrant mentioning again.

- Incorporate interactions between scientists and data managers at the beginning of the project. Data managers and scientists need to work and, most importantly, talk to each other on a regular basis, not at the end of development of a project. If the senior scientists are not supportive of the information management system, it will fail.

- It is essential that the data manager(s) listen and respond to the user community. If users don't like the system the data manager is using, it should be changed.

- Plan on computer and network technology to change!

ACKNOWLEDGEMENTS

Konza Prairie Research Natural Area is a preserve of The Nature Conservancy and is managed by the Division of Biology, Kansas State University. This paper was supported by the NSF Long Term Ecological Research Program at Konza Prairie Research Natural Area.

LITERATURE CITED

Briggs, J.M., and H. Su. 1994. Development and refinement of the Konza Prairie LTER Research Information Management Program. Pages 87-100 in W. K. Michener, J. W. Brunt and S. G. Stafford, editors. Environmental information management and analysis: ecosystem to global scales. Taylor and Francis, Bristol, PA.

Callahan, J.T. 1984. Long-term ecological research. BioScience 34:363-367.

Franklin, J.E., C.S. Bledsoe and J.T. Callahan. 1990. Contributions of the Long-Term Ecological Research Program. BioScience 40:509-523.

Gorentz, J., G. Koerper, M. Marozas, S. Weiss, P. Alaback, M. Farrell, M. Dyer and G.R. Marzolf. 1983. Data management at biological field stations. Report of a workshop at W. K. Kellogg Biological Station, Michigan State University, May 17-20, 1982. Prepared for the National Science Foundation.

Gurtz, M.E. 1986. Development of a research data management system. Pages 23-38 in W.K. Michener, editor. Research data management in the ecological sciences. The Belle W. Baruch Library in Marine Science Number 16. University of South Carolina Press, Columbia, SC.

Knapp A.K., and T.R. Seastedt. 1998. Grasslands, Konza Prairie and long-term ecological research. In A.K. Knapp, J.M. Briggs, D.C. Hartnett and S.L. Collins, editors. Grassland dynamics: Long-term ecological research in tallgrass prairie. Oxford University Press, New York, NY.

Stafford, S.G., J.W. Brunt and W.K. Michener. 1994. Integration of scientific information management system and environmental research. Pages 3-20 in W.K. Michener, J.W. Brunt and S.G. Stafford, editors. Environmental information management and analysis: ecosystem to global scales. Taylor and Francis, Bristol, PA.

Strebel, D.E., B.W. Meeson and A.K. Nelson. 1994. Scientific information systems: a conceptual framework. Pages 59-86 in W.K. Michener, J.W. Brunt and S.G. Stafford, editors. Environmental information management and analysis: ecosystem to global scales. Taylor and Francis, Bristol, PA.

PALMER LTER INFORMATION MANAGEMENT

Karen S. Baker

Scripps Institution of Oceanography, University of California at San Diego,

La Jolla, CA 92093-0218

Abstract. Information management for the Palmer Long-Term Ecological Research (LTER) project is based upon a simple, but functional system which fulfills long-term study site requirements that data be accessible, recorded consistently, and archived digitally. Historical functions of the system include both an online bibliography as well as milestone timelines. With the advent of the common gateway interface internet tool, a dynamic data catalog now provides access to online data.

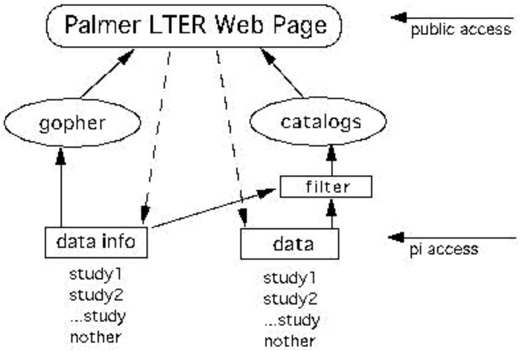
INTRODUCTION

The initial vision of the Long-Term Ecological Research (LTER) network emphasized data management (Michener 1986, Michener et al. 1994). Individual sites have met data management challenges with a variety of approaches (Baker 1996, Benson 1996, Briggs and Su 1994, Ingersoll et al. 1997, Porter 1996, Spycher et al. 1996, Veen et al. 1994). The Palmer LTER, established in the fall of 1990, initiated site data management coordination with the appointment of a data manager in 1992. The Palmer LTER fulfills the data requirements of being a long-term study site by requiring that information be recorded consistently, quality assured, archived digitally, and accessible online (Baker 1996). The advent of the Internet and reliable software plays a critical role since the Palmer participants reside at different home institutions across the country and conduct research in the Antarctic either on station or aboard ship.

Funding limitations have dictated the need for a simple system that takes advantage of technology but minimizes high technology costs. To avoid the need for a central librarian, the responsibility for documentation and data maintenance has been assigned to the data originator. Further, files are kept in unformatted text, avoiding the need for individual users or a data librarian to learn special (mark up) languages. World Wide Web (WWW) use at the network level (lternet.edu) is extended by the Palmer LTER at the local level to coordinate and highlight topics and to preserve a hierarchical data and documentation (metadata) structure initially implemented for a gopher presentation (Figure 1).

Data sets are organized by studies where each study consists of a research cruise or a field season and each data set may consist of one or more data files. Each investigator has an account on the central computer and may access either of the parallel, hierarchical metadata and data directories. It is the responsibility of the individual investigator to document methods and transfer online data sets. Both study and metadata template forms are maintained online. Once files are uploaded and privileges turned over to the data manager, the files are added to the list of those served through a web browser. Files are maintained and updated by the individual principal investigators.

Figure 1. The hierarchical structure for the database showing both the data sets organized into studies as well as the public and the principal investigator (PI) entry points.



HISTORICAL DEVELOPMENTS

The data manager provides some group historian functions through creation and maintenance of the bibliography, a vision timeline, and a milestone timeline. The bibliography is maintained with UNIX software (Bibix) which works in conjunction with the nroff/troff text editing system. Key categories include manuscripts in refereed journals, unrefereed papers, reports, abstracts, and talks by LTER investigators or other individuals associated with LTER efforts. Starting in 1995, bibliographic entries included an abstract and were associated with a larger LTER effort (Chinn 1997). Further, a list of works related to the LTER Palmer field site is maintained. A timeline with lists of changes by date was initiated to document shifts in philosophy or methods that affect core research at the site. Originally, the primary objective of the milestone chart was to document the annual data management developments at the site level in context with the network level (Table 1), but major research developments were also included.

As shown in the milestone table (Table 1), an online information system for browsing metadata was implemented in 1992. This structure was suitable for network distribution in 1993 when gopher software became available. Although data were stored in a hierarchical structure parallel to that of the metadata, they remained accessible to all PIs who had individual user computer accounts rather than through the public browse system. An online World Wide Web implementation of this browse system was completed in 1994, first with Mosaic and, subsequently, with Netscape.

Table 1. A data management subset of the milestone history for the Palmer Long-Term Ecological Research site.

Management Milestones

1990 lter: new LTER site added (PAL)

1991 lterdm: DM meeting San Antonio, Texas with ESA

lterdm: GIS Working Group report

lterdm: GPS Workshop

paldm: bibliography started

paldm: historical files, weather

paldm: cruise eventlog initiated

1992 lter: new LTER site added (MCM)

lterdm: DM meeting Honolulu, Hawaii with ESA lterdm: outreach/China

paldm: datamanager designated (K.Baker)

paldm: 600mbyte disk online

paldm: develop dataforms

paldm: online browse of metadata

paldm: first mac ip tunnel

1993 lterdm: DM meeting Madison, Wisconsin with ESA

lterdm: Environmental Information Management & Analysis Symposium

paldm: datamanager becomes PI

paldm: gopher network browse of metadata

paldm: lter data policy drafted

paldm: field documentation facilitated

paldm: field file transfer via satellite explored

1994 lterdm: DM meeting Seattle, Washington OUTREACH

lterdm: coordinated online data table

paldm: McMurdo User Working Group Member 1994-1997

paldm: www implementation (1990 www began; 1992 mosaic; 1994 netscape)

paldm: 2GB disk online

paldm: historical documentation

paldm: station eventlog initiated

1995 lter: televideo cc meeting

lterdm: DM meeting Snowbird, Utah STRATEGIC VISION, NIS

paldm: PAL/MCM: GIS in Limnology and Oceanography, ASLO, Reno, NV

paldm: skeleton notebook/files online, in field

1996 lterdm: ECOINFORMA participation

lterdm: DM meeting Archbold Biological Station, Florida

paldm: PAL/BAS data manager meeting

paldm: data policy rewritten to OPP/NSF criteria

paldm: data catalog created using dynamic web page paldm: 9GB disk online

paldm: ZIP disk drives used in the field

paldm: field data transfer via LES9 satellite weekly

paldm: member ASA Palmer Working Group on weather

1997 lterdm: DM meeting Univ. of New Mexico

lterdm: ESA/LTER/OBFS Information Workshop

paldm: Member of Data Manager Climate Committee

paldm: Web based forms developed

Dynamic data catalog

A data catalog presentation of the online data sets was implemented in 1996. A dynamic web page (one that runs a program to produce desired content) produces the Palmer catalog upon request from the existing directories of study data sets. Development of dynamic web pages holds many advantages (Wasser 1996), including low cost as well as maintaining the hierarchical study or data set structure for the Palmer LTER. A non-interactive retrieval of documents is performed by a common gateway interface (cgi) script via WWW, gopher, and file transfer protocol (ftp) servers. The study catalog highlights study documentation, metadata, and data (Table 2). Most files are column-delimited ASCII text in order to facilitate transfer, and most graphics are in graphics interchange format (gif). The catalog script gathers tagged lines of the documentation forms for summary as a data catalog organized either by data set or by study.

Several documents for each study are standard: (1) an overview, (2) site maps, (3) a participant list, and (4) an event log chronologically listing the type and location of measurements made during the study. The event log provides an initial cross-index of all component participation for the duration of each study.

Table 2. Subset overview of the dynamic web catalog page.

Palmer LTER Catalog by Study

Palmer LTER Cruise List --

91nov PD91-09: Annual cruise [participants] [table] [eventlog] [map]

* bioacoustics info (acoCalib.dat acoEvent.dat acoMatchMulti.dat

swarmHdr.dat)

* bops info (bopscast.list opticsnoon ts)

* chl info (pd9109.chl)

* chn info (pd9109.chn)

* krillgrowth info (IGR.details IGR.sum adultIGR larvalIGR)

* nutrientsdi info (pd9109.nuthplclog pd9109.nuts)

* pigmhplc info (pd9109.hplc pd9109.nuthplclog)

* prprodpi info (pd9109.pidata pd9109.pifit)

* trawlgen info (krill.tl.wwt trawl.catch trawl.list)

* trawl2m info (altscatter krill.lfhist krill.raw)

92nov PD92-09: Marine Carbon Cycling [participants] [table] [eventlog] [map]

93aug PD93-07: Spring cruise [participants] [table] [eventlog] [map]

93jan PD93-01: Annual cruise [participants] [table] [eventlog] [map]

93mar NBP93-02: Fall cruise [participants] [table] [eventlog] [map]

94dec PD94-12: SantaClaus cruise [participants] [table] [eventlog] [map]

94jan PD94-01: Annual cruise [participants] [table] [eventlog] [map]

95jan PD95-01: Annual cruise [participants] [table] [eventlog] [map]

96jan PD96-01: Annual cruise [participants] [table] [eventlog] [map]

97jan PD97-01: Annual cruise [participants] [table] [eventlog] [map]

Palmer LTER Season List [map1] [map2] ------------------------------------

9192pal Palmer Station Season [participants] [eventlog]

* adbreed info (broods91 ckcnts91 flwts91 humpop91 repro91)

* addemog info (bands91 census91 seen91)

* adforage info (diet91 fish91 header91 krill91 prey91)

* adtelem info (telem91 telhdr91)

* bioacoustics info (acoBiomass.dat acoEvent.dat)

* chl info (9192pal.chl)

* chn info (CHN_Diel.txt CHN_STA_B.txt CHN_STA_E.txt CHN_Sample_Log.txt)

* krillgrowth info (adultIGR adultIGR.details adultIGR.sum larvalIGR

larvalIGR.details larvalIGR.sum)

* nutrientsdi info (9192pal.nuts)

* pigmhplc info (9192pal.hplc)

* prprodpi info (9192pal.pidata 9192pal.pievents 9192pal.pifit)

* zodtrawl info (trawl.lis zodtrawl.rec zodtrawlI.dat)

9293pal Palmer Station Season [participants] [eventlog]

9394pal Palmer Station Season [participants] [eventlog]

9495pal Palmer Station Season [participants] [eventlog]

9596pal Palmer Station Season [participants] [eventlog]

9697pal Palmer Station Season [participants] [eventlog]

DISCUSSION

The existing Palmer LTER data structure creates a simple and functional system. A forms-driven summary of multiple file transfers would be a useful extension of the currently established catalog system, although browser developments may progress to address such issues. A network catalog across all LTER sites is under development. Under consideration is the interface of existing metadata to national standards as well as the question of archiving a long-term online repository of data. Efforts will continue to facilitate connectivity whether at the local, network or national level.

ACKNOWLEDGEMENTS

This work was funded by NSF Grant OPP90-11927. Thanks to S. R. Baker for help with figures. This is Palmer LTER Contribution #149.

LITERATURE CITED

Baker, K.S. 1996. Development of Palmer long-term ecological research information management. Proceedings of the Eco-Informa Workshop, Global Networks for Environmental Information, November 4-7 1996, Lake Buena Vista, FL, 11:725-730.

Benson, B.J. 1996. The North Temperate Lakes research information management system. Proceedings of the Eco-Informa Workshop, Global Networks for Environmental Information, 4-7 November 1996, Lake Buena Vista, FL, 11:719-724.

Briggs, J.M., and H. Su. 1994. Development and refinement of the Konza Prairie LTER research information management program. Pages 97-100 in W.K. Michener, J.W. Brunt, and S.G. Stafford, editors. Environmental information management and analysis: ecosystem to global scales. Taylor and Francis, Bristol, PA.

Chinn, H., and C. Bledsoe. 1997. Internet access to ecological information-the US LTER All-Site Bibliography Project. BioScience 47:50-57.

Ingersoll, R. C., T. R. Seastedt, and M. Hartman. 1997. A model information management system for ecological research. BioScience 47:310-316.

Michener, W.K., J.W. Brunt, and S.G. Stafford. 1994. Environmental information management and analysis: ecosystem to global scales. Taylor and Francis, Bristol, PA.

Michener, W.K. 1986. Research data management in the ecological sciences. University of South Carolina Press, Columbia, SC.

Porter, J.H., B.P. Hayden, and D.L. Richardson. 1996. Data and information management at the Virginia Coast Reserve Long-Term Ecological Research Site. Proceedings of the Eco-Informa Workshop, Global Networks for Environmental Information, November 4-7 1996, Lake Buena Vista, FL, 11: 731-736.

Spycher, G., J.B. Cushing, D.L. Henshaw, S.G. Stafford, and N.Nadkarm. 1996. Solving problems for validation, federation, and migration of ecological databases. Proceedings of the Eco-Informa Workshop, Global Networks for Environmental Information, November 4-7 1996, Lake Buena Vista, FL, 11:695-700.

Veen, C., C. Federer, D. Buso and T. Siccama. 1994. Structure and function of the Hubbard Brook data management system. Bulletin of Ecological Society of America 75:45-48.

Wasser, C. 1996. Dynamic data transfer via the world wide web: Increasing your visitors' understanding of ecological data. Proceedings of the Eco-Informa Workshop, Global Networks for Environmental Information, November 4-7 1996, Lake Buena Vista, FL, 11:737-742.

MANAGEMENT OF A LONG-TERM WATER QUALITY DATABASE:

FLATDAT FOR THE FLATHEAD LAKE BIOLOGICAL STATION

Melissa E. Holmes and Geoffrey C. Poole

The University of Montana, Flathead Lake Biological Station,

311 Bio Station Lane, Polson, MT 59860-9659.

Abstract. Long-term monitoring databases present data management challenges that are unique. The development of an information management system must be carefully planned to determine expectations of the system in terms of use, output and longevity. Data and metadata must be adequate for accurate future analyses. A system must evolve to address an organization's changing needs and take advantage of new technology.

INTRODUCTION

Long-term monitoring databases present unique data management challenges. First, the personnel who collect and manage monitoring data may change over time, often resulting in inconsistencies in the ways data are collected, analyzed, and stored. Second, techniques used to collect monitoring data may change over time due to improvements in data collection methodologies. Third, archiving and documenting data sets that result from ongoing long-term monitoring are difficult because there is often no "final product." Instead, the database is continually growing and represents the current and ever-changing status of the monitoring program.

For more than two decades, the Flathead Lake Biological Station (FLBS) has been monitoring water quality in Flathead Lake and its catchment. In order to address the data management challenges presented by this monitoring program, we began to develop a digital information management system in 1992 called *FlatDat*. By providing a central repository for the FLBS monitoring data, FlatDat helps to ensure that: a) data are collected, entered, and archived in a consistent manner; b) any changes in standard procedures in the field or laboratory are documented; and c) the current status of each project is accurately represented in a location where FLBS researchers can access the data and track progress.

FlatDat provides a total data management solution for the acquisition, calculation, retrieval, and archival of data generated by the analysis of water samples at the station. It tracks the status of each water sample brought into the lab, automates all calculations in our analytical lab by generating different types of electronic worksheets for each lab methodology, archives data in a form that is accessible to researchers, and generates billing reports for accounting purposes.

DEVELOPMENT OF FLATDAT

The development of a data management system requires a significant investment of resources. For the Flathead Lake Biological Station, the largest portion of this investment is in personnel assigned to the project. Approximately half of one full-time Data Manager's time is spent on evaluation, development, and maintenance of the FlatDat system. Several thousand dollars are spent each year on purchasing and upgrading computer equipment and software, and on the personnel required to evaluate and maintain FlatDat. Thus far, the development of the system has been funded entirely by the Flathead Lake Biological Station. The current implementation of FlatDat was written using Microsoft® Foxpro® for Macintosh®.

The management of ecological data has received well-deserved scrutiny in recent years (e.g., Michener et al. 1994). FlatDat was designed based on four premises that arose from such scrutiny: 1) electronic data are most flexible and powerful when stored in the rawest form possible; 2) data must be secure, yet accessible; 3) computerized databases should work the way people work; and 4) data management should be inexorably linked to existing tasks and jobs. Each of these premises is discussed below.

Electronic data are most flexible and powerful when stored in the rawest form possible.

In many ecological databases, only the final data values are entered. These data are often the result of a variety of calculations, manipulations, and transformations that are not evident in the numbers themselves. This can lead to several problems. First, the limitations inherent in a data collection methodology are apt to be ignored and the data are more apt to be misapplied or misinterpreted. Since only the final product (i.e., the final value) is available, people generally *assume* a high level of accuracy in the methodology. Even if the database user questions a particular datum, the only choices are "take it" or "leave it" and, frequently, no means are available to check the datum by tracking its genesis.

On the other hand, if raw values are entered into the database and all calculations are programmed and automated, the database can be used to investigate the genesis of any particular datum. Readings from field or lab instruments are available for scrutiny as are the calculations used to convert raw data into finalized values. Additionally, in the event an error is discovered in the calculations, or if an investigator wishes to calculate final values using a different method (for instance, to compare numbers to another study that used an alternative method), then the new calculations can be programmed into the computer and the entire database recalculated automatically. Programming calculations into the database ensures consistency in calculations over time and generates an accurate record of any changes in methodology.

In FlatDat, raw data (e.g., readings from analytical equipment) are entered into electronic laboratory worksheets. These worksheets are built automatically by FlatDat based on the status of samples that have been collected and logged into the computer by field personnel. If necessary, electronic worksheets can be altered easily by the chemical analyst to delay analysis of some samples or include quality control samples if necessary. The analyst then prints out the blank worksheet, writes down the raw readings on the hard copy as the analyses are run (thereby providing a hard-copy for future reference), and enters the values from the hard copy back into the electronic worksheet. The computer performs necessary calculations as the data are entered and prints a final copy of the worksheet when saved. The hand-written hard copy and final hard copy are compared to check for errors, stapled together and filed for future reference. Through the use of a "quality control number" ("qc number") assigned to each value resulting from the worksheet, any value in the database can be traced back to a specific worksheet generated in the analytical lab.

This provides the ability to track down errors when unlikely values are discovered during data analysis and ensures consistency over time. However, consistency over time only extends as far back as raw data are entered. In order to provide this level of accuracy, consistency, and scrutiny to our entire period of record, all *raw* lab readings from 1979 through the first implementation of FlatDat in 1992 have also been entered into FlatDat. During the ~20 year period of record, several methodological changes were made in the lab, including changes in detection limits and purchases of new instruments. FlatDat is able to accommodate such changes because each analysis for each sample is linked, via the electronic worksheet and qc number, to the methodology used at the time it was run. To date, FlatDat contains the results of over 200,000 individual water quality analyses from tens of thousands of water samples collected in the Flathead Basin.

Data must be secure, yet accessible.

In order to be most useful, long-term monitoring data must be easily accessible to researchers in formats that are flexible and useful. However, unrestricted access to the database risks corruption or loss of the data. FlatDat provides a simple query window that allows even novice users to formulate complex queries. The results of the query can be saved to external files that can be imported into statistical software or spreadsheets. FlatDat employs several levels of security as well. First, users must log into the FlatDat program to use the database. Users are granted privileges that are appropriate to the level of access required. Most users can only view and query data. Field technicians can log samples into the computer and edit samples that have not yet been analyzed in the lab. Chemical analysts can edit electronic worksheets. Some especially powerful tasks (such as altering the pre-programmed methodologies or recalculating the database) are only available to the database administrator. Any time a change is made to a sample description or electronic worksheet, FlatDat records the date of the change and the person who made the change.

Assigning appropriate access privileges to the networked disk drive where the data are stored provides additional security. Additionally, we designed a database "maintenance" utility that checks the database for logical errors and rebuilds indexes associated with each database table. Finally, nightly incremental backup of the database ensures that we can restore the database to the state in which it existed at any particular date in the past. On a monthly basis, a backup copy of the database is stored off-site to protect against catastrophic loss.

Computerized databases should work the way people work.

A computer program should always save time. It should never make the user's job more difficult. When an information management system is being designed, the number of users and their skill levels need to be considered. FlatDat utilizes a graphical user interface with windows and forms making the program easy to use and preventing users from inappropriately seeing or changing the underlying data files. Controls in these windows and forms limit the information that can be entered into each field. For instance, some controls only allow users to enter a data element from a list of available elements, force the user to enter data in a particular format, or automatically default values such as dates and user names. Restrictions like these help to ensure ease of use and accuracy, but reduce flexibility for more advanced users. In developing FlatDat, we worked closely with the users to design a system that strikes a balance between data security, ease of use, and flexibility. During development, users were encouraged to provide feedback regarding what they liked and disliked about the database and whether or not solutions were efficient or inefficient. To the extent possible, the program was modified to incorporate this information while ensuring data integrity.

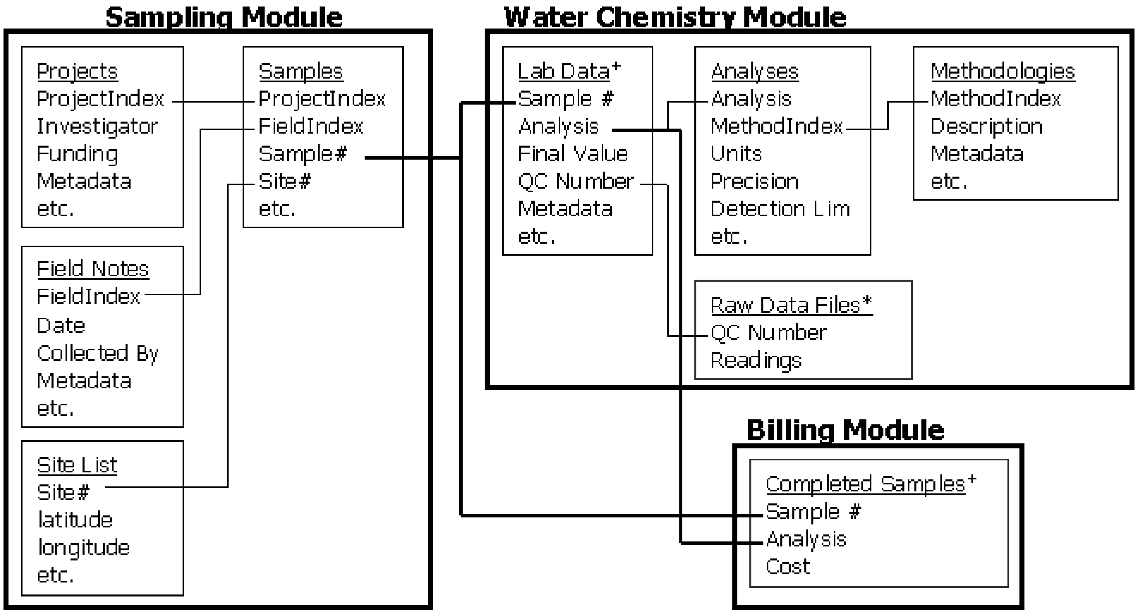
Data management should be inexorably linked to existing tasks and jobs.

A primary goal of FlatDat was to allow individual employees to manage the data they need to manage in order to do their job. Additionally, we intended to reduce duplication of effort, improve communications regarding data resources, and increase productivity (*sensu* Michener and Haddad 1992). This was sometimes difficult, since some employees are inexperienced with or even resistant to using a computerized database. However, since field personnel most accurately know what happens in the field and lab personnel know what occurs in the lab, the database will be most accurate if both groups are empowered to manage the data. To accomplish this, data entry screens were designed to look like the field and laboratory sheets already in use at the time when FlatDat was implemented. This provided a comfortable and familiar electronic "environment" for personnel who were not necessarily skilled computer users. The system is entirely menu- and mouse-driven to accommodate the novice, yet liberal use of "short cut" keystroke and "hot-buttons" allows rapid navigation by experienced users. Again, the use of forms and windows that limit what a person is allowed to enter or change is critical. However, ample opportunity to enter field and lab notes also provides a means of recording critical information in a more flexible format.

DATA FILES AND RELATIONSHIPS

FlatDat consists of three modules: the Sampling Module, the Water Chemistry Module, and the Billing Module (Figure 1). The Sampling Module contains information about water samples that have been collected. Each sample belongs to a project that has been developed by a particular investigator or group of investigators. Sample collection sites for all projects are stored in the same "site list", thereby encouraging cross-compatibility between projects over the long term. The "field notes" table contains data and metadata that describe the field sampling trip and the conditions under which the samples were collected.

Figure 1. Relationships diagram for the FlatDat system. Fine lines indicate linked data fields within a module. Bold lines indicate linked data fields between modules.



*Each methodology has its own raw data file; the specific raw data file associated with a particular analysis is determined by the methodology used.
+Each sample can have multiple records in this database; one record for each chemical analyses performed

The Water Chemistry Module contains data and metadata describing laboratory analyses of water chemistry. Any number of analyses may be run on each sample. The "lab data" table stores the results from each analysis, along with quality control information and lab notes (metadata). The raw data used to calculate the final values are stored in the Raw DB Files. Since each methodology run in the lab generates different raw data, each methodology has a separate and unique raw data file.

A unique sample number is assigned to each sample as it is collected. This sample number is stored in both the "samples" table and the "lab data" table and is used to link data in the Sampling Module to data in the Water Chemistry Module.

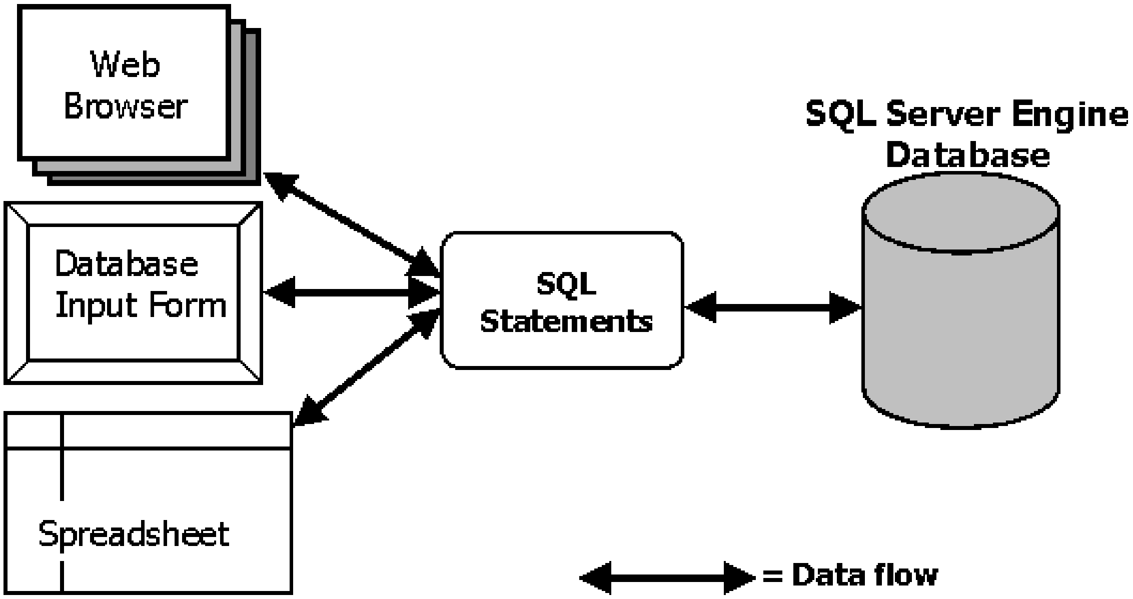
The Billing Module uses data from the "samples" and "lab data" tables to determine which water samples are complete (i.e., all requested analyses are run and entered into FlatDat). Information about completed samples is pulled into the Billing Module, where an invoice report is generated and the sample is marked as having been billed.

FUTURE IMPROVEMENT FOR FLATDAT

The advent of a mixed-platform environment at the FLBS, along with the needs for Internet access, better database tools, and better integration of metadata (*sensu* Ingersoll et al. 1997) have influenced the design of the next version of FlatDat. Although calculations and methodology have remained much the same, software and storage needs have changed entirely.

In the next version of FlatDat, the data will be stored in a centralized Structured Query Language (SQL) server that will be able to provide query results to a variety of applications including web browsers (Figure 2). This will allow more flexibility in matching the capabilities of existing software to requisite tasks, thereby reducing the amount of custom programming required. For instance, database forms can be used to log samples, spreadsheets can be used to create electronic worksheets, and the database can be queried via a web browser allowing cross-platform access to the data. In a manner similar to Ingersoll et al. (1997) we plan to make specific portions of the data available to the public on the World Wide Web. We are currently in the process of evaluating specific software tools for use in the next version of FlatDat.

Figure 2. Data flow in the FlatDat system. The data will be stored in a SQL server that will provide query result to a variety of applications, including web browsers and spreadsheets.



Recently, issues surrounding the importance of managing and incorporating metadata have received overdue attention (e.g., ESA 1995, NRC 1995). Having finished compiling and entering our historic water chemistry data (i.e., data from the period 1979-1992), the need for organization and integration of our metadata has become evident. While long-term FLBS employees may know where to access metadata in old field-logs and journals, newer employees do not know what metadata are available or where to find it, thereby reducing the utility of our consistent and meticulously organized data set. The new version of FlatDat will include fields for tighter integration of metadata and, as we did for analytical data, all existing field notes and other types of metadata will be entered for the entire period of record. The new version will also feature a more robust billing module, including the capability to easily bill multiple funding sources for a particular project.

CONCLUSIONS

When preparing to build an information management system, system requirements and design are extremely important and should be completed before implementation begins. The developer should work with management and key users to decide what information needs to be tracked and what is expected of the system in terms of use, output, and longevity. As the needs of an organization change and computers and software become more advanced, an information management system must evolve to address those needs and take advantage of new technology.

The expenses associated with the maintenance and evolution of information systems must be an *a priori* acknowledgement and planned for accordingly over the long term. With only occasional exception, biologists and ecologists can no more be expected to manage data effectively than information specialists could be expected to design and conduct ecological research. While research staff, students, and scientists must be involved in the data management process in order for it to be effective, there is no substitute for hiring and retaining staff who are trained in information system management to oversee the data management process.

The development of FlatDat has been a learning process for everyone involved. Our progress to date has increased the utility of our data set and the productivity of investigators at the FLBS. We expect that the next implementation of FlatDat, based on the needs and principles outlined above, will fulfill our data management needs well into the next millennium.

LITERATURE CITED

Gross, K.L., C.E. Pake, E. Allen, C. Bledsoe, R. Colwell, P. Dayton, M. Dethier, J. Helly, R. Holt, N. Morin, W. Michener, S.T.A. Pickett, and S. Stafford. 1995. Final report of the Ecological Society of America Committee on the future of long-term ecological data (FLED). Volume I: Text of the report. (*http://www.sdsc.edu/~ESA/FLED/FLED.html*).

Ingersoll, R.C., T.R. Seastedt, and M. Hartman. 1997. A model information management system for ecological research. BioScience 47: 310-316.

Michener W. K., J. W. Brunt, S. G. Stafford, editors. 1994. Environmental information management and analysis: ecosystem to global scales. Taylor and Francis, London, UK.

Michener, W.K., K. Hadadd. 1992. Chapter 1-Database Administration. Pages 4-14 in J.B. Gorentz, editor. Data management at biological field stations and coastal marine laboratories. Report of an invitational workshop, April 22-26 1990, W.K. Kellogg Biological Station, Michigan State University, East Lansing, MI.

[NRC] National Research Council. 1995. Finding the forest in the trees: the challenge of combining diverse environmental data. National Academy Press, Washington, DC.

THE H. J. ANDREWS CLIMATOLOGICAL FIELD MEASUREMENT PROGRAM

Donald L. Henshaw

U.S. Forest Service Pacific Northwest Research Station,

3200 SW Jefferson, Corvallis, OR 97331

Frederick A. Bierlmaier

Forest Science Department, Oregon State University, Corvallis, OR 97331

Hazel E. Hammond

U.S. Forest Service Pacific Northwest Research Station,

3200 SW Jefferson, Corvallis, OR 97331

Abstract. Research modeling in ecology and hydrology demands the collection of short time-step, spatially distributed climatic measurements. Standardized sets of measured parameters with standard collection methods are critical for comparability of measurements. Standard data archival formats, quality assurance procedures, and sound mechanisms for method documentation are essential for efficient handling of large quantities of electronic data. Web access to near real-time climatological data as well as long-term archives has proven invaluable to researchers.

INTRODUCTION

Research modeling and other spatial studies in hydrology and ecology are placing increasing demands on long-term measurement collection systems. The advancement of remote sensing and Geographical Information System (GIS) technology has spurred the development of spatially distributed ecosystem models requiring short time-step, spatially-distributed measurements to understand processes and test hypotheses at different scales. Coordinated climatological field measurement programs with standardized collection methods, quality assurance procedures, and data archival formats are necessary to meet the needs of multidisciplinary programs. Long-term measurements of climatological variables at various temporal and spatial scales provide a necessary foundation for understanding ecosystem processes and documenting environmental changes (Greenland 1993).

Technical advances in climatological instrumentation such as meteorological sensors, data loggers, and telecommunications allow for routine measurement of critical climatological parameters at fine spatial and temporal scales. Collection of vast amounts of data with improved accuracy is now possible, but quality assurance of these data is not guaranteed without significant changes in field measurement programs. Field collection techniques, quality assurance procedures, and data processing and archival systems must evolve to meet the demands of rapidly growing data collection efforts. Documentation of the methodologies and archival formats are necessary for the long-term survival and utility of this information.

Researchers also wish to take advantage of new web and telemetry technologies to gain near real-time access to critical stream and weather conditions. Access to climatic summary data is also requested shortly after significant weather events. Computer systems employing telemetry and web technology provide researchers remote access to raw weather data and a basis to determine the necessity of a site visit (e.g., for purposes of storm sampling). Telemetry also provides technicians with a way to monitor the operation and accuracy of these remote stations.

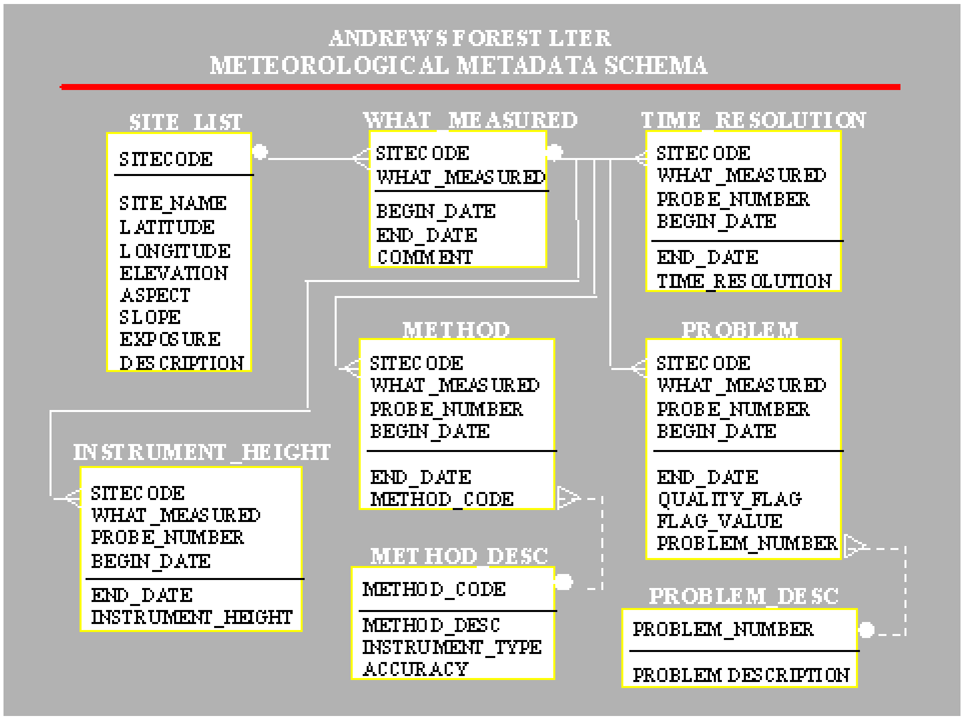
HISTORICAL PERSPECTIVE/BACKGROUND

The hydroclimatological measurement program of the H. J. Andrews Experimental Forest LTER site started in 1952. The existing system evolved as an accumulation of measurement programs with various designs and objectives in response to individual projects and scientist interests (Rothacher et al. 1967, Emmingham and Lundburg 1977, Waring et al. 1978). Until recetly, most meteorological data collection had been driven by very site-specific questions.

Climatic stations as well as precipitation and temperature networks were established independently. Field techniques, collection methods, data handling procedures, and the parameters measured were highly variable among measurement sites. In the early 1990's, it became clear that new objectives for the hydroclimatological measurement program had superseded many of the original objectives and would require a more comprehensive network design and new instrumentation (Henshaw et al. 1995).

Much of the Andrews Forest (*http://www.fsl.orst.edu/lter*) occupies rugged terrain with steep mountain slopes and extensive conifer stands. The elevation range of 420 to 1630 m makes most of the forest accessible only by snowcat during most winters. This has traditionally limited the ability to find suitable, accessible sites for meteorological measurement, and especially precipitation measurement. To overcome the limited environmental representativeness of our measurement system, two new additional weather stations were built to give the forest four comprehensive "benchmark" meteorological stations (See site map). Additionally, a system of secondary precipitation and temperature measurement stations has been established to augment the benchmark stations. Secondary stations are strategically located to examine spatial variability of precipitation, air, soil, and stream temperature along environmental gradients. Radiotelemetry was established at the "benchmark" stations and selected stream gaging stations to create "virtual" accessibility to these remote sites.

Figure 1. Andrews Forest LTER meteorological metadata schema.



Modifications were made to both new and existing stations to standardize the variables measured, temporal resolution, collection methodology, sensor heights, and instrumentation. Quality assurance/quality control procedures as well as data archival formats and procedures were also modified and standardized. To track changes in instrumentation and methodology, collection time resolution, and sensor heights, and to fully document problems encountered with meteorological measurement, a database schema for metadata was established (See Figure 1).

METEOROLOGICAL DATABASES AND METADATA

The Forest Science Data Bank (FSDB) has been actively involved with managing hydrological, meteorological, and other related data for the Andrews LTER for over twenty years (Stafford et al. 1986). Due to the piecemeal evolution of the measurement program, separate databases and data formats were established for these independent efforts. For example, climatic stations collected different sets of variables and were structured in separate databases. The precipitation network and thermograph networks also were independent databases. To prepare for the increasing influx of electronically collected data, databases were standardized for all meteorological data collected on the Andrews. This standardization has greatly simplified processing methods by allowing the use of standard programs for quality assurance checking of all meteorological data. An added benefit is the ability to automatically create web pages for access to these data.

Within the Forest Science Data Bank (FSDB) every data set table has an associated metadata file that: 1) describes the table's variables (attributes) including each variable name, definition, units, and domain (data type and format); 2) indicates whether the variable is coded; and 3) describes attribute information critical for generic quality assurance checking, such as whether a variable a) can be null, b) is a primary key, and c) shows the common variable range (Spycher et al. 1996). When attributes are listed as coded, a second metadata file lists all coded variables, all valid codes, and code definitions. The meteorological data take advantage of these metadata files to document the common database structures and properly define the measurement variables. Table 1 lists the independent tables that are established for the measurement parameters. Separate tables are maintained for daily and higher temporal resolution data.

All documentation or metadata regarding the hydroclimatological measurement system is stored and maintained by the FSDB. Records of all changes in field instrumentation, sensors, locations, sensor heights, and temporal resolutions are maintained for every measurement variable in a relational database schema (See Figure 1). Tables within the schema include specific measurement site descriptions and the parameters measured at each station with spanning dates. Histories of changes in collection resolution, sensor heights, instrumentation, and encountered problems are maintained. To avoid redundant descriptions, independent tables are established to describe methods (with instrumentation) and problems that were encountered.

QUALITY ASSURANCE (QA)

Electronic data collection provides a mechanism to collect immense quantities of data with very sophisticated measurement sensors and data loggers. Meteorological data are commonly collected at 15-minute or hourly time intervals and sometimes with multiple sensors for each measurement parameter. This is a tremendous change from the historical data collection efforts which consisted primarily of digitized and manually read records from various recording charts and punch tapes, and hand-recorded field notes. However, the transition to electronic data collection does not assure that the collected records are reliable and error-free, and the data require extensive quality assurance (QA) checking. Also lost is the instantaneous feedback the field technician had when examining a chart in the field. This bird's-eye view of seeing the overall graphical picture is replaced with a worm's eye view of the data logger's single discrete digital values.

In October 1991, the LTER Data Managers conducted a survey on the use of electronic data collection instrumentation at the LTER sites (Ingersoll and Chapal 1992). It was discovered that most meteorological data are collected, stored, and transferred by electronic means at LTER sites. The resultant survey responses provided an assessment of common problems and solutions encountered in electronic data collection and describe the QA procedures at the LTER sites. It is clear that while new electronic technologies have greatly reduced the need for manual data entry and digitizing, field techniques and QA protocols need to evolve to reduce the increasing chances of undetected problems residing in final data archives.

QA procedures for electronically collected data must involve field technicians, the site data manager, and the principle investigators to be most effective. Additionally, the ability to detect problems quickly provides more timely feedback to field technicians. Field procedures used at the Andrews and other LTER sites to help provide quality assurance include:

- Routine equipment inspection: Pre-designed check sheets that can lead the technician through standard protocols are very useful.

- Routine instrument calibration: Local site or manufacturer recalibration or instrument replacement is critical.

- Independent sensor checks: Make independent measurements side-by-side with the field instrument to ascertain field sensor reliability. The reliable lifetime for some electronic sensors can be relatively short (~1 year).

- Redundant measurements: Provide an alternate sensor as a backup. For example, a non-electronic recording chart might be used alongside an electronic data logger.

- Telemetry: Radio or satellite telemetry is an invaluable tool in knowing whether a station is functioning properly and can quickly alert technicians to measurement problems.

- Mechanisms for noting problems or making comments: Field recorders, field notebooks or check sheets, and tape recorders provide a means to describe observations.

- Communication: Strong communication between the field technicians, data manager and principle investigators.

QA continues once the data reach the laboratory. The Andrews LTER performs three types of programmatic quality assurance checks of electronically collected meteorological data.

Generic checking is performed on every data set in the FSDB by referencing the metadata to find and report potential problems in the data. Problems might include data values out of normal range, data values missing when "not null" is specified, use of improper codes (e.g., for coded variables such as data quality flags), and improper data structure or data format of ASCII files.

Special rule checks specific to climate data include: 1) checks for date and time validity and proper sequencing; 2) checks that minimum<=mean<=maximum when applicable; and 3) checks for proper placement of missing flags in the quality code fields. Computer code for implementing these special rules is maintained with the climatic metadata. Visual checking of data values is performed by: 1) graphing a single parameter over time; 2) graphing redundant or multiple sensors for comparison, and 3) graphing related or associated parameters together. Examples include: 1) displaying a station's mean, maximum, and minimum values over time in a moving time graph; 2) creating scatter plots of the data of one site versus another; and 3) displaying wind data in polar graph form. Models have also provided valuable insight into discovering inconsistencies in measurement sensors, e.g., snow hydrology models (Duan 1996).

Once the quality assurance checking has been completed, there remain substantial questions in determining how problem data should be documented. The Andrews LTER data structures employ data quality flags for every measured value. Determining whether a value is tagged "questionable" or "missing" is very subjective. Inaccurate values showing a consistent bias may still have redeeming value and are typically coded "questionable" and retained. Very questionable data may be ultimately coded "missing" and discarded if no interpretation is possible. Since researchers and modelers often require that no values be "missing" over key time periods, these missing values will frequently be estimated. Most estimation techniques and equations are saved in the metadata, and values are tagged "estimated." Field technician notebooks or logs with comments can be invaluable in evaluating problem data. A "Comments" database table is maintained to describe periods of flagged data and provides a useful log of problems and solutions.

DATA ACCESS

Historical climatic and stream flow data were collected on strip charts, circular charts, punch tapes and hand-written check sheets. In the early 1980's, weather station technology allowed capture of our primary station's data on a nearby computer at the Andrews headquarters (Bierlmaier and McKee 1989). However, all data were hand-delivered to the Corvallis research facility, data processing took weeks and in some cases years, and researchers had no access to any real-time weather condition data. In 1993, a dedicated 56Kb direct network link was established, greatly improving the connectivity of Andrews computers with the Corvallis research group's local area network.

In 1995, radio telemetry capability was added to the remote benchmark stations and two stream gauging stations providing instantaneous access to remote weather stations (See Plate 7). During a storm in February of 1996 when access to remote sites was impossible due to washed out roads, fallen trees, and flooding, radio telemetry provided data to researchers that would not otherwise have been available.

In 1996, a suite of Microsoft™ software, including Foxpro™ were used in conjunction with Campbell Scientific™ software and data loggers to develop programs for automatically accessing raw telemetry weather station data and processing the data into web pages accessible over the World Wide Web (*http://fredb2.fsl.orst.edu/).* This near real-time weather web page is updated hourly in the winter and daily in the summer. All of the Andrews Forest LTER historical and current climate databases can be found on the WWW at *http://www.fsl.orst.edu/lter/data/studies/ms01/ms01fmt.htm*. All metadata files associated with this as well as the data sets themselves are written to HTML and ASCII files for web access.

CONCLUSION

As research modeling in ecology and hydrology demands the collection of short time- step, spatially distributed climatic measurements, coordination of climatological field measurement programs becomes very important. Standardized sets of measured parameters with standard collection methods are critical for comparability of measurements. Standard data archival formats, quality assurance procedures, and good mechanisms for method documentation are essential for efficient handling of large quantities of electronic data and their long-term usefulness. Web access to near real-time data has also proven invaluable to researchers in assessing remote site conditions.

ACKNOWLEDGMENTS

Support for the H. J. Andrews climatological field measurement program is provided by the Pacific Northwest Research Station and the National Science Foundation grant DEB9632921. This paper benefited from discussions with Al Levno, Craig Creel, Arthur McKee, John Morean, and Greg Downing.

LITERATURE CITED

Bierlmaier, F. A., and W. A. McKee. 1989. Climatic summaries and documentation for the primary meteorological station, H.J. Andrews Experimental Forest, 1972 to 1984. Gen. Tech. Rep. PNW-242. US Department of Agriculture, Forest Service, Pacific Northwest Research Station, Portland, OR.

Duan, J. 1996. A coupled hydrologic-geomorphic model for evaluating effects of vegetation change on watersheds. Dissertation. Oregon State University. Corvallis, OR

Emmingham, W.H., and G.A. Lundburg. 1977. Climatic and physiological data summaries for the H.J. Andrews reference stand network. Coniferous forest biome internal report 166. University of Washington, Seattle, WA.

Greenland, D. 1993. The climate of the H.J. Andrews Experimental Forest, Oregon, and its regional synthesis. USDA Forest Service, Pacific Northwest Research Station. Cooperative Agreement No. PNW 92-0221.

Henshaw, D.L., W.A. McKee, and A. Sikka. 1995. Program for hydroclimatological measurement at the H.J. Andrews Experimental Forest. *http://www.fsl.orst.edu/lter/research/compplns/clima/hjamet.htm*

Ingersoll, R., and S. Chapal. 1992. Management of electronically collected data within LTER. *http://www.fsl.orst.edu/lter/im/ingersol.htm*

Rothacher, J., C.T. Dyrness, and R.L. Fredriksen. 1967. Hydrologic and related characteristics of three small watersheds in the Oregon Cascades. U.S. Department of Agriculture, Forest Service, Pacific Northwest Forest and Range Experiment Station. Portland, OR.

Spycher, G., J.B. Cushing, D.L. Henshaw, S.G. Stafford, and N. Nadkarni. 1996. Solving problems for validation, federation, and migration of ecological databases. Proceedings of the Eco-Informa Workshop, Global Networks for Environmental Information, November 4-7 1996, Lake Buena Vista, FL, 11:695-700. Environmental Research Institute of Michigan(ERIM), Ann Arbor, MI.

Stafford, S.G., P.B. Alaback, K.L. Waddell, and R.L. Slagle. 1986. Data management procedures in ecological research. Pages 93-113 in W. K. Michener, editor. Research data management in the ecological sciences. The

Belle W. Baruch Library in Marine Science No. 16. University of South Carolina Press, Columbia, SC.

Waring, R.H., R.H. Holbo, R.P. Bueb, and R.L. Fredriksen. 1978. Documentation of meteorological data from the coniferous forest biome primary station in Oregon. Gen. Tech. Rep. PNW-73. US Department of Agriculture, Forest Service, Pacific Northwest Forest and Range Experiment Station Portland, OR.

CLIMATE DATABASE PROJECT: A STRATEGY FOR IMPROVING INFORMATION ACCESS ACROSS RESEARCH SITES

Donald L. Henshaw

U.S. Forest Service Pacific Northwest Research Station, 3200 SW Jefferson,

Corvallis, OR 97331

Maryan Stubbs and Barbara J. Benson

Center for Limnology, University of Wisconsin-Madison, Madison, WI 53706

Karen Baker

Scripps Institution of Oceanography, University of California-San Diego, La Jolla, CA 92093

Darrell Blodgett

Forest Soils Laboratory, University of Alaska, Fairbanks, AK 99775

John H. Porter

Department of Environmental Sciences, University of Virginia, Charlottesville, VA 22903

Abstract. To facilitate intersite research among the network of Long-Term Ecological Research sites, information managers are exploring strategies for linking individual site information systems. A prototype to provide climatic summaries dynamically has been developed and serves as one model for improving access to data across sites. Individual sites maintain local climate data in local information systems while a centralized site continually updates and provides access to all sites' data through a common database. Common distribution report formats have been established to meet specific needs of climate data users.

INTRODUCTION

Information managers associated with the Long-Term Ecological Research (LTER) program have developed a basic foundation for a Network Information System (NIS) with a primary goal of facilitating intersite research (Stafford et al. 1994). To accommodate the needs of various intersite studies and synthesis efforts within the LTER, it is critical to develop dynamic systems for providing comparable data from multiple LTER sites. Improving access and adding query capability to intersite data using network information servers is a major component of current NIS development (Brunt 1996). With each site operating its own information management system, the LTER NIS will employ a variety of strategies in linking these individual systems (Porter et al. 1997).

Climate meteorological data are collected at all LTER sites and are the most frequently requested data. Synthesis groups need ready-access to climatic summaries from multiple sites. A NIS prototype to provide climatic summaries dynamically has been developed and serves as one model for improving access to data across sites. This approach allows individual sites to maintain the local climate data in local information systems while a centralized site continually updates and provides access to all sites' data through a common database.

BACKGROUND

A standards document developed by the LTER Climate Committee (Greenland 1986) established baseline meteorological measurements to characterize each LTER site. Standardized measurements provide a basis for coordinating meteorological measurements at two or more sites and enable intersite comparisons. More recently, a project to conduct climatic analyses of the LTER sites (CLIMDES; *http://lternet.edu/im/climate/climdes/*) gathered individual site temperature and precipitation data (1960-1990) and created on-line monthly summaries for each site

(Greenland et al. 1997). While the CLIMDES project satisfied an immediate need for access to monthly site climate data, the structure provided no method for maintaining and updating these summaries or satisfying frequent requests for daily climate data. Most of the LTER sites had their climate data available on the World Wide Web (WWW), but the data sets were sometimes difficult to find and were formatted and aggregated differently at each site.

The NSF-funded XROOTS (*http://lternet.edu/im/xroots/aclim.htm*) project requires intersite climate data to synthesize belowground productivity using root biomass data from multiple sites. The idea that distribution of data in report formats amenable to users independent of the data storage format was explored in an XROOTS climate workshop (Bledsoe et al. 1996). Two monthly distribution report formats were recommended to accommodate both spreadsheet (V-One, i.e., twelve monthly values for one variable per record) and database (V-Many, i.e., one monthly value for many variables per record) users (See Table 1).

OVERVIEW

As part of the LTER Information Managers' NIS development, the LTER climate database project (ClimDB) has developed a prototype for harvesting daily climate data in a standardized exchange format using the WWW from a subgroup of LTER sites. The harvested data are stored in a centralized relational database. Climate variables include daily minimum, maximum, and mean air temperature and daily precipitation. Applications have been developed initially to generate the two XROOT monthly distribution formats using this centralized database of daily values. Additionally, a webpage (*http://www.limnology.wise.edu/climdb.html*) has been created to provide access to the daily and monthly climate data as well as to permit query by LTER site, weather station, and date.

SPECIFIC EXCHANGE AND DISTRIBUTION FORMATS

Each of the five sites participating in the prototype development process provided climate data files in a standardized daily exchange format at an Internet address (URL). For this model, the site files could be either static or produced by a dynamic script. A comma-delimited format was agreed upon after discussions revealed the diversity of approaches, opinions, and needs among sites. For instance, date can be stored as a single 8-character field, comma separated, or Julian day designated. It is important to note there is not one "right" exchange format. The primary criteria require that individual sites "filter" local site data into the exchange format. The standardized daily exchange format agreed upon is as follows:

Site, station, date, value1, flag1, value2, flag2, value3, flag3, value4, flag4

where,

site the three-letter LTER site code

station that site's name for the weather station

date 8-character field, yyyymmdd

value1, flag1 mean air temperature and corresponding flag

value2, flag2 maximum air temperature and corresponding flag

value3, flag3 minimum air temperature and corresponding flag

value4, flag4 precipitation and corresponding flag

All temperature values are reported in degrees Celsius and precipitation in millimeters. Each value has a corresponding data quality flag where flags are coded as follows:

G or blank value is a good value

E value is estimated

Q value is questionable

M value is missing

T trace value (for precipitation only)

Here is a brief example of the daily format from the Andrews Forest (AND) site's Primary Meteorological Station (PRIMET) aligned for readability:

```
AND,PRIMET,19960101,6.8, ,10.8,Q,4.5, , 0.0,T AND,PRIMET,19960102,5.3,
,10.6,Q,0.8, , 4.3, AND,PRIMET,19960103,7.7, , 9.7, ,4.1, ,20.6,
AND,PRIMET,19960104,4.2, , 6.7, ,2.4, ,11.4, AND,PRIMET,19960105,4.8,E,
7.4,E,2.7,E, ,M AND,PRIMET,19960106,5.7,E, 9.7,E,1.3,E, ,M
```

Daily climate data from all sites are harvested automatically from the local sites using a simple script calling the WWW line mode browser. An example of the harvest command line for the Andrew's Forest climate data is:

```
www -n -source http://www.fsl.orst.edu/lter/webmast/and_clim.txt >and.dat
```

Data are stored in a relational database at the centralized site. [Note: Currently, the prototype is using the Oracle^{TM} database management system and the centralized site is the North Temperate Lakes LTER Site. Eventually the ClimDB project will move to the LTER Network Office, and the relational database software may change.] Application programs produce two monthly distribution tables (See Table 1).

A webpage allows the user to query for daily data in addition to providing the two monthly tables. Monthly summary values are displayed along with the number of valid daily values included in the summary. Missing and questionable values are excluded from summary values. Listing the number of valid data values used in calculating a monthly value gives the user some assurance about the value's accuracy and represents a valuable addition to any distribution format.

METADATA

Every meteorological station will be described in a central metadata database. An entity-relationship diagram (See Figure 1) shows the proposed schema for the metadata database. The metadata database is currently being developed in Oracle^{TM}. LTER-site-level information, individual station descriptions, and specific measurement documentation form the three major entities. Standardized web forms will be used to collect this information from participating sites. Metadata term definitions will be made available on the central webpage. Metadata will be critical for intersite studies in evaluating key differences in site descriptions and methodology.

Table 1. Examples of the two monthly distribution tables (V-One and V-Many) are shown for the Andrews Forest (AND) site's Primary Meteorological Station (PRIMET). The "#" indicates the number of valid daily values (including estimated values) that were used in calculating the monthly summary value.

V-One. V-One displays one variable per table and is primarily intended for use in spreadsheets. These two abbreviated examples show mean monthly air temperature and total precipitation.

AND PRIMET Avg_mean_air_temp_c

Year	Jan	#	Feb	#	Mar	#	Apr	#	May	#		Nov	#	Dec	#
1991	0.1	31	5.8	28	4.5	31	6.9	30	10.0	31	...	6.5	30	3.2	31
1992	3.3	29	5.8	29	8.1	30	10.0	30	15.0	31	...	5.0	30	1.0	31
1993	-0.6	31	0.6	28	6.0	31	7.7	30	13.2	31	...	-0.8	30	-0.2	30

AND PRIMET Totl_precip_mm

Year	Jan	#	Feb	#	Mar	#	Apr	#	May	#		Nov	#	Dec	#

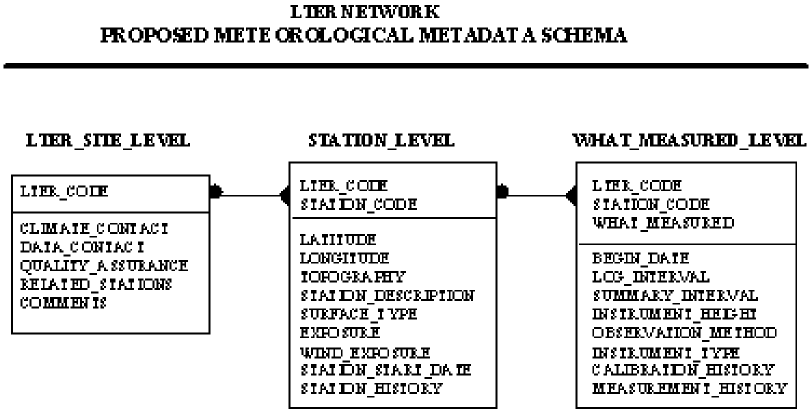
1991	232	31	208	28	221	31	242	30	195	31	...	451	30	214	31
1992	160	31	201	29	40	31	290	30	20	31	...	337	30	419	31
1993	242	31	95	28	354	31	394	30	237	31	...	103	30	278	31

V-Many. V-Many displays many variables per table and is primarily intended for use in relational databases. This example includes all four prototype variables of monthly mean, maximum, and minimum air temperature and total monthly precipitation.

AND PRIMET

Year	Month	Mean	#	Max	#	Min	#	Ppn	#
1991	Jan	0.1	31	5.3	31	-3.0	31	232	31
1991	Feb	5.8	28	12.4	28	2.1	28	208	28
1991	Mar	4.5	31	11.2	31	0.3	31	221	31
1991	Apr	6.9	30	13.3	30	2.6	30	242	30

Figure 1. Proposed schema for the metadata database



CONCLUSIONS

With an increasing focus on intersite activities within the LTER program, the LTER Information Managers are developing a Network Information System to facilitate intersite research. This LTER NIS prototype for climate data will serve as a model for other intersite data set integration efforts. The approach allows for the diversity in information management systems across the LTER network. Data sets are distributed across multiple sites, but are accessible in common distribution formats from a central site. Specially formatted distribution reports have been established to meet specific needs of climate data users, but the design is extensible in that it permits update with additional formats as the need arises.

ACKNOWLEDGMENTS

The authors would like to acknowledge contributions from the North Temperate Lakes LTER site (DEB 9632853) for participating in the development of this prototype and for supporting the centralized database and web pages. Contributions from the H. J. Andrews Experimental Forest (NSF grant DEB 9632921), the Bonanza Creek Experimental Forest (DEB 9211769), Palmer Station (OPP 9632763), and the Virginia Coast Reserve LTER (DEB 9411974) sites are also recognized for participating in the development of this prototype. LTER sites are funded all or in part by the National Science Foundation. We also wish to acknowledge the efforts of Caroline Bledsoe for her strong support and continued interest in this project.

LITERATURE CITED

Bledsoe, C., J. Hastings, and R. Nottrott. 1996. Xclimate workshop. Davis, CA. *http://lternet.edu/im/xroots/aclim.htm*

Brunt, J. W. 1996. Developing an LTER network information system for the 21st century. *http://lternet.edu/is/is18Jan96.htm*

Greenland, D., T. Kittel, B.P. Hayden, and D.S. Schimel. 1997. A climatic analysis of Long-Term Ecological Research sites. *http://lternet.edu/im/climate/climdes/*

Greenland, D. 1986. Standardized meteorological measurements for Long-Term Ecological Research sites. Bulletin of the Ecological Society of America 67:275-277. *http://lternet.edu/im/climate/standard86.html*

Porter, J., D.L. Henshaw, and S.G. Stafford. 1997. Research metadata in Long-Term Ecological Research (LTER). Proceedings of the Second IEEE Metadata Conference. Silver Spring, MD. *http://www.computer.org/conferen/proceed/meta97/list_papers.html*

Stafford, S.G., J.W. Brunt, and W.K. Michener. 1994. Integration of scientific information management and environmental research. Pages 3-19 in W. K. Michener, J. W. Brunt and S. G. Stafford editors. Environmental information management and analysis: ecosystem to global scales. Taylor & Francis, Bristol, PA.

DATA AND INFORMATION MANAGEMENT IN THE ECOLOGICAL SCIENCES:

SYNOPSIS FROM A FIELD STATION PERSPECTIVE

Hilary M. Swain

Archbold Biological Station, Lake Placid, FL 33862

William K. Michener

Joseph W. Jones Ecological Research Center, Route 2, Box 2324, Newton, GA 31770

Abstract. This paper represents a synopsis of the Data and Information Management in the Ecological Sciences (DIMES) workshop from the viewpoint of the Organization for Biological Field Stations (OBFS). In addition to presenting highlights of the workshop, we examine the effectiveness of the workshop for the member field stations associated with OBFS. This paper is based on closing remarks at the workshop (Swain), interspersed with post-workshop observations by one of the organizers (Michener).

INTRODUCTION

Objectives of the DIMES workshop were to: initiate technology transfer; facilitate interpersonal networking; communicate training opportunities; produce hard copy and digital versions of the DIMES proceedings to serve as a resource guide; and identify future data management needs at field stations and research sites. Implementing onsite data management and integrating data management among sites were described in the opening session as "two of the primary challenges facing field stations over the next decade." The workshop organizers challenged all participants to two proximate workshop objectives; meet ten new people and learn ten new things. Post-workshop evaluations ranged from comments such as "overwhelming," to "great -- right on target," to "not technical enough," and reflected the broad spectrum of backgrounds and interests of the attendees. However, the general consensus of participants was that the workshop largely exceeded expectations. The diverse speakers were extremely effective at conveying information to attendees about data management, and most participants left with the sense that attention to data management is "increasingly overdue" at many field stations, and that many tools and techniques are available to facilitate data management at field stations and other institutions.

PARTICIPANTS

The DIMES workshop attracted approximately 100 participants. A survey of 65 attendees showed that the workshop reached its target audience. Geographic and institutional (Table 1) representation was diverse. Almost half (48%) of the attendees described themselves as data managers.

Table 1. Affiliation of workshop attendees.

Organization	Percentage
Organization for Biological Field Stations	38%
Field Stations and Research Centers (non-OBFS affiliated)	45%
LTER sites or LTER Network Office	28%
University (Faculty and Student)	26%
National Laboratories (e.g., Oak Ridge National Laboratory,	5%

National Center for Ecological Analysis and Synthesis)	
State or Local Government Agency	5%
Federal Agencies	15%
Other (e.g., consultants)	5%

Table 2. Strengths and difficulties of field stations and field research sites from a data management perspective.

	Strengths	Difficulties
1.	The size, diversity, and longevity of the legacy data sets held at field stations, and the institutional site-based knowledge, are an invaluable and irreplaceable ecological resource.	The rate of entropy (loss of information content over time) of the legacy data sets presents a seemingly paralyzing data management backlog for many field stations.
2.	Many of the legacy data sets are fairly site-specific with fewer thematic components, and therefore potentially easier to integrate.	Most field stations and sites are now also tackling regional analyses and cross-site comparisons. The expansion of spatial, temporal, and thematic scales of ecological study requires scaling up to much more extensive data management.
3.	Field stations typically embody a depth of natural history knowledge that complements the quantitative ecological data sets. In many cases, this tradition has included retaining original data forms and field notes on site.	The natural history aspects of many ecological data, and the scattered documentation of such knowledge, means that data management at field stations must deal with extensive metadata requirements.
4.	Several field stations and sites, particularly the LTER sites, have ongoing data management protocols and institutional policies which can act as models for other field stations.	Ongoing data management problems at field stations are 80% cultural. Large numbers of skeptics among research scientists are yet to be convinced of the value of integrated data management. The proprietary aspects of data have not been resolved at many field stations.
5.	Field stations are entering the computer equipment market at a time when prices have come down considerably, and there is increased capacity to network existing computing facilities.	Chronic budget shortfalls and lack of institutional support for data management are common at many field stations. The costs of data management are high and include: personnel (which may exceed data collection efforts), long-term curation and maintenance, archival facilities and metadata consultation.

A FIELD STATION PERSPECTIVE

Strengths and difficulties of field stations for data management

Previous studies have summarized the inherent strengths and weaknesses of field stations and field research sites from a data management viewpoint (e.g., Gorentz 1992, Gross et al. 1995, Lohr et al. 1995). Speakers at this workshop did not dwell on these issues, but clearly understood that successful data management at field stations is based upon acknowledging existing strengths and accommodating intrinsic difficulties. Discussion of field station strengths (Table 2) was accompanied, in most cases, with an understanding of the offsetting difficulties. The extent to which most speakers recognized the varied field station contexts into which their recommendations have to be implemented was reassuring to field station personnel.

Opportunities and challenges facing data management at field stations

The DIMES Workshop provided an overarching summary of the opportunities to use current tools for data management. The consensus was that "the tools are there" for each step of the data management process. Authors that specifically addressed data management tools in their contributions to this volume are listed below:

- Infrastructure design including hardware (Chapal), communications (Nottrott),

and software (Baker)

- Data entry (Briggs)
- QA/QC (Edwards)
- Database management system processing (Porter)
- Metadata (Michener)
- Archival (Olson)
- Scientific visualization (WWW (Benson) and San Diego Supercomputer Center (Helly))
- Data and information resources (e.g., World Wide Web (Benson and others))

Although the tools for data management are generally available, implementation at field stations and field research sites presents a series of challenges. Presenters were encouraged to include "tricks" of the trade that they use to overcome cultural barriers to effective data management. Successful data management is only achieved in social environments that are receptive because there are long-term benefits as well as incentives to participate. Components that presenters viewed as critical for implementation were: institutional incentives and recognition; effective software support; and initial marketing to participants. Successful completion of a *site needs assessment* is critical to facilitate data management design and implementation. Site needs assessments include: identification of data and site objectives; developing policies for data sharing and data ownership; and assessing the infrastructure, personnel, and budget. Workshop participants were interested in seeing real-world examples of cross-site comparisons or interdisciplinary studies where the results clearly demonstrate the scientific value of participating in shared data management, to help market the advantages. Specific challenges include demonstrating how data management has effectively: delayed "data entropy" (*sensu* Michener et al. 1997); supported the use/re-use of data by the data originator and data re-use by others; and facilitated expansion of spatial, temporal, and thematic scales of ecological study.

Field Stations recognize there is a full spectrum of tools available for data management, but have low budgets and limited trained personnel. Workshop presenters provided advice on "low-end" and well as "high-end" solutions (Table 3). Further guidance is needed, however, as to "where to get on the ramp," depending on current circumstances and future needs. Specific topics of interest to field stations and research sites include: *technical interoperability* such as field station infrastructure (e.g., hardware, software, communications) and ecological data archives; *semantic interoperability* including standards (metadata, methods, syntax) and metadata tools (entry, search); *social interoperability* including data and information sharing and technology transfer (training, meetings); *funding* for the computational infrastructure and data recovery; and *reward systems* like recognition for data and metadata publications and other incentives.

Table 3. Synopsis of DIMES Workshop recommendations for low-, medium-, high-end technological solutions for various stages of the data management process.

Task	Low	Medium	High
Data entry	spreadsheet (e.g., EXCEL^{TM})	full-screen data entry program with programmable QA/QC (e.g., EasyEntry^{TM})	full-screen data entry program with QA/QC and database functions (e.g., SAS^{TM} and relational DBMS)
Quality assurance/ quality control (QA/QC)	Manual	Range checks, field validation, etc. (e.g., EasyEntry^{TM}, SAS^{TM})	Comprehensive graphical and statistical QA/QC (e.g., SAS^{TM})
Database management system (DBMS)	non-DBMS with data management functions (e.g., merge, subset, Boolean operators, etc. (SAS^{TM}))	User-friendly PC-based DBMS (e.g., ACCESS^{TM}, PARADOX^{TM})	Comprehensive PC- or UNIX-based DBMS (e.g., ORACLE^{TM})
Archival	redundancy (i.e., disks and paper copies stored in two locations)	Tape, optical disk	off-site data archival facility (e.g., Oak Ridge National Laboratory DAAC)
Metadata	Paper	Word processor	DBMS
Hardware	PCs and printers	Workstation & color output	mixed PC & UNIX, multi-media
Software	WORD^{TM} & EXCEL^{TM}	SAS, graphics	ARC/INFO^{TM}, ERDAS^{TM}
Network	Modem	Internal network (e.g., NOVELL)	Internet & WWW connectivity

FUTURE DIRECTIONS

Training and mentoring

The DIMES Workshop was a recognizable starting point for data management networking based on personal contacts developed at the workshop. Other ideas for training and mentoring included development and utilization of "hands-on" training centers (possibly in conjunction with the National Center for Ecosystem Analysis and Synthesis (NCEAS), Oak Ridge National Laboratory DAAC, or other established facilities). In addition, the concept of site visits by "Rapid Assessment Data Management Teams" was suggested. Such a team might include groups of 2-3 individuals drawn from a pool of experienced data managers who could "jump-start" the planning, design, and implementation processes.

Future meetings

Significant interest was generated at this Workshop for follow-up workshops and meetings. Possible venues include: NCEAS, other workshops sponsored by NSF-DBA, annual "Data Management" workshops/symposia at ESA or other Society-affiliated meetings, and a Journal/ Bulletin Board. It will be worth considering what other potential participants/groups were missing from the attendees at this workshop and how best to include them in future workshops and training efforts.

A closing note

The DIMES Workshop provided a superb compilation of the tools and techniques available to participants for implementing data management. Missing from the discussion, however, was the debate about a broad vision of collective success, in terms of ecological data management. How do various organizations integrate data management across multiple sites and regions? Clearly, LTER sites play a leadership role in this task, but what is the collective vision to tie together data management among the LTER Network, the Organization for Biological Field Stations, the Association of Ecosystem Research Centers, and members of societies such as the Ecological Society of America? Such a collective vision of success will lay the ecological and data management cornerstones upon which future generations can build.

ACKNOWLEDGMENTS

The many members of OBFS who participated in this workshop sincerely appreciated the encouragement and the organizing skills of all the conveners (Michener, Gosz, McKee, and Porter) and also acknowledge generous NSF support for field station participation at this meeting. Participants and Resource Guide users are encouraged to provide feedback to NSF and the workshop organizers (Michener, Gosz, McKee, and Porter).

LITERATURE CITED

Gorentz, J.B. 1982. Data management at biological field stations. Report of a workshop at Kellogg Biological Station, April 1990.

Gross, K.L., C.E. Pake, E. Allen, C. Bledsoe, R. Colwell, P. Dayton, M. Dethier, J. Helly, R. Holt, N. Morin, W. Michener, S. T. A. Pickett, and S. Stafford. 1995. Final report of the Ecological Society of America Committee on the Future of Long-term Ecological Data (FLED). Volume I.

Lohr, S.A. P.G., Connors, J.A. Stanford and J.S. Clegg. 1995. A new horizon for biological field stations. Rocky Mountain Biological Laboratory. Miscellaneous Publication No. 3.

Michener, W.K., J.W. Brunt, J. Helly, T.B. Kirchner, and S.G. Stafford. 1997. Non-geospatial metadata for ecology. Ecological Applications 7(1): 330-342.

ECOLOGICAL METADATA

William K. Michener

Joseph W. Jones Ecological Research Center, Route 2, Box 2324, Newton, GA 31770

Abstract. Metadata represent comprehensive documentation of the content, context, quality, structure and accessibility of a data set. In this chapter, relevant geospatial and non-geospatial metadata "standards" are reviewed, World Wide Web sources of information pertaining to metadata are identified, a metadata implementation "recipe for success" is proposed, and remaining challenges are discussed.

INTRODUCTION

Metadata are the information necessary to understand and effectively use data, and include documentation of the data set contents, context, quality, structure, and accessibility. From the perspective of the data originator, metadata are necessary to support further processing and analysis. When a scientist's goal is to re-use data collected by others, comprehensive metadata may be essential to support identification and acquisition of suitable data, as well as to facilitate additional processing and analysis.

Metadata are receiving increased attention by the scientific community. For example, ecologists, scientific societies, and state and federal agencies are recognizing the importance of high quality, well-documented, and securely archived data for addressing long-term and broad-scale questions (e.g., Gross et al. 1995). In addition, ecological data, such as those collected by individual and teams of scientists at field stations, marine laboratories, natural areas, parks, and preserves, represent a significant national resource that are essential for understanding and monitoring the health of the dynamically changing environment. Comprehensive metadata are required to counteract the natural tendency for data to degrade in information content ("data entropy" *sensu* Michener et al. 1997) through time.

PROGRESS IN METADATA STANDARDIZATION

Although all ecological data have a "spatial" element (i.e., are collected at one or more points in space), ecological data may be generally categorized as being either geospatial or non-geospatial. Geospatial data include those types of data that are explicitly associated with a geographical location. Examples include remotely sensed imagery, geographic information system (GIS) data layers, data derived from broad-scale sampling efforts (e.g., National Weather Service, National Atmospheric Deposition Program), as well as fine-scale sampling of spatially explicit patterns and processes. Non-geospatial data, on the other hand, and for the purposes of this paper, might include data from laboratory and micro- to mesocosm experiments, as well as other ecological data that are collected at a finite number of points. For these cases, precise geographic coordinates of the sampling sites are relatively unimportant and often unrecorded. Most metadata standardization efforts have, thus far, focused on geospatial data.

Geospatial metadata standardization

As part of the ongoing evolution of the National Biological Information Infrastructure (NBII) and standardization of geographical data in the Federal government, significant attention has focused on standardizing geospatial metadata. One of the most significant products to emerge has been a document entitled "Content Standards for Digital Geospatial Metadata" (Federal Geographic Data Committee 1994) which contains a comprehensive list of geospatial metadata descriptors. Seven categories of metadata descriptors are included in the document: (1) identification; (2) data quality; (3) spatial data organization; (4) spatial reference; (5) entity and attribute; (6) distribution; and (7) metadata. Additional efforts are underway to add extensions to the Content Standards that are relevant to vegetation classification data, as well as cultural, demographic, and other types of geographical data. Additional information on Federal Geographic Data Committee activities, metadata generation tools (e.g., NBII MetaMaker), and related material can be found at the World Wide Web sites listed in Table 1.

Table 1. Geospatial metadata World Wide Web sites.

"FGDC Metadata FAQ"
http:05/12/98/geochange.er.usgs.gov/pub/tools/metadata05/12/98/tools/doc/faq.html
"National Biological Information Infrastructure (NBII)"
http:05/12/98/www.its.nbs.gov/nbii/index.html
"NBII MetaMaker Version 2.10"
http:05/12/98/biology.usgs.gov/nbii/metamaker/metamaker.html
"Metadata Tools"
http:05/12/98/badger.state.wi.us/agencies/wlib/sco/metatool/mtools.html
"Metadata Tool Evaluation"
http:05/12/98/www.fgdc.gov:80/metadata/mitre/task2/index.html

Generic ecological metadata descriptors (non-geospatial)

Ecological studies often require the collection of an extremely diverse array of data including attributes that characterize and quantify the chemical and physical environment, organism physiology, population and ecosystem dynamics, community composition, landscape structure, as well as anthropogenic influences. It is unlikely that a single metadata standard, no matter how comprehensive, could encompass all types of ecological data because of this complexity. Consequently, a generic set of non-geospatial metadata descriptors were recently proposed for the ecological sciences (Michener et al. 1997). The list of metadata descriptors was suggested as a template that could serve as the basis for more refined subdiscipline- or project-specific metadata guidelines. Five categories of metadata descriptors were delineated: (1) data set descriptors; (2) research origin descriptors; (3) data set status and accessibility; (4) data structural descriptors; and (5) supplemental descriptors (Table 2).

IMPLEMENTATION STRATEGIES

Metadata may be recorded in a variety of forms ranging from free-flowing text to incorporation into a structured database management system (DBMS). Some of the most important metadata attributes are often recorded in the field using pencil and paper. "Natural history" observations are frequently critical for correct interpretation and analysis of field data. Field notes and other metadata can later be maintained in paper files or incorporated into word processing files, SAS programs, DBMS programs, or World Wide Web-accessible documents. The choice of metadata media is often dictated by availability of software, trained personnel, and time. Guidelines for metadata structure and supporting technology (WWW forms, etc.) are currently being discussed and developed at the San Diego Supercomputer Center, National Center for Ecological Analysis

Table 2. Generic non-geospatial metadata descriptors for ecological research (adapted from Michener et al. 1997).

I. Data Set Descriptors

A. Data set identity

B. Data set identification code

C. Data set description

1. originator(s)
2. abstract

D. Keywords

II. Research Origin Descriptors

A. "Overall" project description

1. identity

2. originator(s)

3. period of study

4. objectives

5. abstract

6. source(s) of funding

B. "Specific sub-project" description

1. site description

2. experimental or sampling design

3. research methods

4. project personnel

III. Data Set Status and Accessibility

A. Status

1. latest update

2. latest archive date

3. metadata status

4. entry verification

B. Accessibility

1. storage location and medium

2. contact person(s)

3. copyright restrictions

4. proprietary restrictions

5. costs

IV. Data Structural Descriptors

A. Data set file

1. identity

2. size

3. format and storage mode

4. header information

5. alphanumeric attributes

6. special characters/fields

7. authentication procedures

B. Variable information

1. variable identity

2. variable definition

3. units of measurement

4. data type

5. data format

C. Data anomalies

V. Supplemental Descriptors

A. Data entry

1. data forms used

2. location of completed data forms

3. verification procedures

B. QA/QC procedures

C. Related materials

D. Computer programs and data processing algorithms

E. Archival

F. Publications

G. History of data set usage

---

and Synthesis, the Long-Term Ecological Research Network, as well as numerous other organizations. Regardless of the availability of tools that can facilitate metadata entry, storage, and retrieval, there are several non-technological activities that can be performed at the level of the individual investigator, field station, project, or "group" to facilitate successful metadata implementation.

Metadata: a recipe for success

The first and probably most important component of metadata implementation is to perform a site or project needs assessment. Such an assessment entails identifying data objectives (e.g., projected or desired data longevity, potential for re-use, value), establishing guidelines and procedures for data sharing and data ownership, assessing infrastructure (e.g., availability of hardware, software, people, funds), and categorizing and prioritizing metadata activities. For example, at a field station, meteorological data may receive a high priority for metadata implementation because of their perceived value to a large number of ongoing studies, historical usage patterns, and potential for repeated use over time. In contrast, infrequent field surveys performed as part of an undergraduate research project may receive a lower priority for archival and metadata implementation. Once categories of data are prioritized, it is necessary to either adopt an existing metadata standard (e.g., geospatial metadata standard (FGDC 1994)) or identify a set of minimal and optimal metadata descriptors that meet perceived needs.

The second recommended step in metadata implementation is to perform a pilot project using one to three relatively "simple" data sets. Based upon successes and difficulties encountered in the pilot project, it is useful to re-evaluate site needs and objectives. For example, a formal or informal cost-benefit analysis may facilitate future prioritization and balance completeness of metadata versus funding and personnel availability. Following this evaluation process it is necessary to formalize metadata activities. It may be desirable, for example, to develop relevant policies and procedures, identify available metadata tools or initiate programming efforts to develop appropriate tools, and establish a reward structure for providing comprehensive metadata. Metadata and other data management activities should be re-evaluated on a periodic basis to insure that they are meeting specified objectives. Several simple "rules of thumb" may facilitate successful implementation:

- Keep it simple! Start small and build upon successes. For example, the time and effort expended on a pilot project are usually paid back several-fold in the long run.
- Build consensus among scientists and data managers from the start. Data management initiatives, regardless of their potential benefits, are often unsuccessful when the "user community" is excluded from the process. Data management must be fully integrated into the research planning process and involve the scientific community it serves (Stafford et al. 1994).
- Data longevity is roughly proportional to metadata comprehensiveness. However, establishing a goal of complete metadata that can meet all future needs may be exorbitantly expensive and, ultimately, unattainable.
- Data and metadata should ideally be platform-independent. Hardware and software change frequently. Today's "standard" may be gone tomorrow. Thus, it pays to avoid proprietary storage formats whenever possible.
- The degree to which high-quality ecological data and accompanying metadata are securely archived and accessible for future research is directly related to the extent to which an ethic of data stewardship is promoted and rewarded (Porter and Callahan 1994).

FUTURE CHALLENGES AND OPPORTUNITIES

Flexible metadata tools that support entry, search, and retrieval are essential for facilitating metadata implementation. There is a significant need for research and development in this area. Many of the scientific benefits that are associated with the availability of high-quality data and metadata have been discussed here and elsewhere (Gross et al. 1995, Michener et al. 1997). Although future research endeavors will inevitably pay more attention to metadata and other aspects of data management, the clock is running out on many extremely valuable long-term and unique ecological data sets. There is a significant need for established funding mechanisms and data archives to support metadata development and secure long-term storage of these irreplaceable data (Gross et al. 1995). Development of attendant reward systems (e.g., peer-reviewed data and metadata publications, equating database construction with publication efforts) will be essential for further promoting an ethic of data stewardship (Porter and Callahan 1994).

Much future discussion will likely focus on standardization issues. If ecological metadata are or should be standardized, then who decides on the standard? Should standardization occur at the level of the institution (e.g., field station, university), society (e.g., Ecological Society of America), discipline (e.g., litter decomposition), funding agency (e.g., NSF), or globe (e.g., International Long-Term Ecological Research Network)? What constitutes minimal and optimal criteria and standards? Like other standardization efforts, the true test of any emerging metadata standard will ultimately rest on whether the standard is simple to use and easily understood, and whether or not it makes our science better.

Finally, it should be reiterated that there are costs associated with metadata implementation, data archival, and other data management activities. Personnel costs associated with developing metadata can, in some cases, exceed data collection efforts. Issues related to long-term curation and maintenance of data and metadata cannot be dealt with effectively in most 1-, 2-, or 3-yr grant cycles. Devoting resources during a short-term project to data management (e.g., metadata) costs money and personnel effort and can result in fewer short-term publications. On the other hand, when high quality data and metadata are securely archived, they can be "mined" for many years or decades into the future. Proper balance of short-term costs versus long-term gain is an issue that warrants continued thought and discussion.

ACKNOWLEDGMENTS

This paper benefitTed from discussions with James Brunt, John Helly, Thomas Kirchner, and Susan G. Stafford.

LITERATURE CITED

Gross, K. L., C. E. Pake, E. Allen, C. Bledsoe, R. Colwell, P. Dayton, M. Dethier, J. Helly, R. Holt, N. Morin, W. Michener, S. T. A. Pickett, and S. Stafford. 1995. Final report of the Ecological Society of America Committee on the future of long-term ecological data (FLED). Volume I: Text of the report. (*http://www.sdsc.edu/~ESA/FLED/FLED.html*).

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.

Porter, J. H. and J. T. Callahan. 1994. Circumventing a dilemma: historical approaches to data sharing in ecological research. Pages 193-202 in W. K. Michener, J. W. Brunt and S. G. Stafford editors. Environmental information management and analysis: ecosystem to global scales. Taylor & Francis, Bristol, PA.

Stafford, S. G., J. W. Brunt and W. K. Michener. 1994. Integration of scientific information management and environmental research. Pages 3-20 in W. K. Michener, J. W. Brunt and S. G. Stafford editors. Environmental information management and analysis: ecosystem to global scales. Taylor & Francis, Bristol, PA.