Proceedings of the

1991 Data

Management Workshop

LTER Publication No.13

Editors:

W. Michener and J. Brunt

Contributors:

K. Baker, B. Benson, C. BLedsoe$_1$ E. Boose, C. Bowser, J. Briggs,

J. Brunt, G. Calabria, S. Chapal, A. El Haddi, J. Gorentz, J. Greenlee,

D. Henshaw, R. Ingersoll, T. Kirchener, M. Klingensmith, M. Klopsch,

W. Michener, B. Moller, M. MunIlo, R. Nottrolt, J. Porter, J. Vande Castle,

C. Veen$_1$ R. Waide

Published by the Long-Term Eological Reseerch Network Office

University of Washington$_1$ College of Forest Resources, AR-i 0

Seattle, Washington 98195, 206-543~4853, FAX: 206-685-0790/3091

TABLE OF CONTENTS

I.Executive Summary 2

II. Activity Reports 5

A. LTER Core Data Set Catalog 5

B. Data Management Symposium Proposal 5

C. Climate Database Proposal 5

D. Kellogg Workshop Report 5

E. Interaction With the Chinese Ecological

Research Network 5

F. Data Access Policies 7

G. GPS Proposal 7

H. Databits 7

I. LTERNET 7

J. LTER History/Reference File 8

K. LTERnet Bulletin Board 8

Ill. Working Group Reports 10

A. Data Access Facilitation 10

B. Minimum Site Capabilities and Establishment,

Evaluation, and Review of LTER Data Management 14

C. Future of Science and Technology Supplements and

Funding for Data Management Activities 18

D. Future Directions of Data Management In LTER 24

E. Quality Assurance/Quality Control 31

IV. Appendices

A. Site Flashes 35

B. Software Demos 42

C. Data Management Workshop Participants 44

D. Historical Aspects of Data Management at LTER Sites 46

I. EXECUTIVE SUMMARY

1991 LTER Data Management Meeting

San Antonio, TX; 1-3 August 1991

Prepared by William Michener (NIN) and James Brunt (SEV) 25 scientists (at least one from each LTER site) attended the annual meeting.

Lawson Spivey (Global Change Director for the Soil Conservation Service) gave the keynote address which was followed by extensive discussion on potential collaboration between the S.C.S. and LTER.

The meeting was devoted to assessing LTER Data Management (DM) activities since the 1990 Workshop and forming five working groups to address specific topics:

LTER DM Activities Since the 1990 Workshop:

- LTER Core Data Set Catalog was published. Electronic versions (ASCII & WordPerfect) are available via FTP.

- A DM symposium proposal (entitled "Environmental Information Management and Analysis: Ecosystem to Biosphere Scales") is being reviewed at NSF. A proposal for supplemental funds to develop an inter-site LTER climate database was supported by NSF.

- The April 1990 DM workshop report for Field Stations and Marine Labs is in press.

- Collaboration between LTER Data Managers and the Chinese Ecological Research Network has been initiated.

- Guidelines for developing LTER data access policies were completed and sent to all sites.

- The Network-wide GPS proposal was funded. Two high and two lower precision GPS units were ordered. Training for the high precision units was held Sept.30 - Oct. 4, 1991 at Boulder, CO.

- Databits (the LTER OM newsletter) continues to serve a critical need within the LTER Network. John Porter was commended for excellence in publishing.

- A SPRINTNET connection is being added to LTERnet.

.2

- An "LTER Data Management History and Reference File" has been compiled at the Network Office.

- The C-News electronic bulletin board and news software is now running on **LTERnet.**

1991 Workshop Accomplishments:

(1) Data Access Facilitation. Design characteristics for Interactive Data Access (IDA) systems were outlined. Six steps to support further development of LTER IDA systems were proposed. Other

mechanisms for facilitating access to LTER data (including online catalogs, data publication, interactive workshops, and education) were discussed.

(2) Minimum Site Capabilities (MSC) and LTER Data Management Review Criteria. An MSC working document was developed for review by the LTER sites. The MSC document can aid new sites in establishing DM systems, assist site reviewers in assessing DM, and potentially influence the availability of funding for DM activities. Criteria (including adherence to MSC) to facilitate review of an LTER site's Data Management System were proposed. These documents will be submitted at the next LTER Coordinating Committee for discussion.

(3) Science and Technology Supplements. The technical supplements have been very important in creating the current LTER technological infrastructure. They have enhanced collaborative research and allowed sites to expand the spatial and temporal scale of their research. The minimum standard installation is dynamic: it needs regular updating and continuing support.

Mechanisms have been proposed that would directly or indirectly encourage supplemental funding for the following activities: (1) enhance research at sites and across the network through the initiation and completion of data management projects that are personnel and/or computationally intensive; (2) allow support for refurbishment projects or new facilities for LTER data management activities to be included as cost university sharing; (3) allow sites to achieve a Minimum Site Capability (MSC) for management of LTER data; (4) support acquistion of new technologies which will facilitate intersite research, data management and analysis.

(4) Future Directions for LTER Data Management. Specific actions (4 of 8 which have already been initiated) which could greatly facilitate research activities within LTER and would increase the visibility of LTER Data Management efforts in the larger ecological community were proposed. These include:

1. Completion of a pilot research/data management project which would benefit the entire LTER Network.

2. Development of an LTER Data Management Slide Presentation.

3. Developing workshops on new analytical/statistical techniques.

4. Expansion of LTER Bulletin Board/Email activities.

5. LTER Data Publication.

6. Expanding the scope of the 1992 Data Management Workshop to include non-LTER Scientists (specifically, OBFS, SAML, and LMER).

7. Developing Collaborative Research with the Soil Conservation Service.

8. Outreach Activities with the Chinese Ecological Research Network.

(5) Quality Assurance/Quality Control (QAIQC). The QA/QC working group developed (and is in the process of implementing) the following action plan:

1. Prepare a summary report on QNQC techniques for ecological data management.

2. Evaluate the quality of electronically gathered data within the LTER network and subsequently develop recommendations regarding the management of such data. This will be done by designing an electronic mail survey, analysing the results, and consulting specific organizations or businesses for input (e.g. NWS, NOM, NCAR, **etc.).**

II. ACTIVITY REPORTS

A. LTER Core Data Set Catalog (William Michener, NIN)

The Core Dataset Catalog was completed and sent to all data managers. An electronic version is available via anonymous FTP in ascii text format or as WordPerfect 5.0 files. The catalog was enthusiastically received at NSF and is being distributed widely to potential LTER data users.

B. Data Management Symposium Proposal (William Michener, NIN)

The Data Management Symposium proposal is still undergoing review at NSF. The proposal was geared towards expanding the scope of the first Data Management symposium that was held at the North Inlet LTER site. The proposal emphasizes a regional and global perspective with a strong GIS component. A list of international participants will be developed. Tentatively, the symposium is scheduled for January 1993, contingent upon funding.

C. Climate Database Proposal (John Gorentz, KBS; Tom Kirchner, CPER)

The proposal for supplemental funds for an inter-site climate database has been recommended for approval by the NSF program. This project will include both approaches discussed at the 1990 Data Managers meeting in Snowbird. The lead site for development of the email access system will be KBS, and the IPC/RPC tools and example applications will be developed at CPR. BNZ, KNZ, NET, and NTL are also taking part. Tom Kirchner presented displays showing how the IPC/RPC tools will work, based on some example applications he has already developed. Most of the sites contributed examples of their climate data to assist John Gorentz in developing a software library of specialized aggregation operators for climate data types. The data managers discussed strategies for dealing with the problem of double indirect costs vs. accountability in multi-site proposals.

D. Kellogg Workshop Report (John Gorentz, KBS)

The April 1990 DM workshop report for Field Stations and Marine Labs is at the printers -- a few touchups still to be done. It should be out in Late August.

E. Interaction With CERN (James Brunt, SEV)

Pursuing Data Management activities with other ecological groups, agencies, and countries was an action item identified in the 1990 Data Management Report. Connections to the newly forming Chinese Ecological Research Network were seen as an ideal starting point for international outreach.

.5

At the LTER All-Scientists Meeting at Estes Park CO, Susan Stafford **(AND),** Barbara Benson (NTL), Jim Gosz (SEV), Rudolf Nottrott (NET), and James Brunt (SEV) met with Beryl Leach (NAS/CSCPRC/NRC) and Zhao Shidong (Deputy Director of the Chinese Institute of Applied EcQIogy). What resulted from their discussions was a proposal from Jim and Beryl to bring a delegation of Chinese scientists to the US to establish contact with LTER sites that have similar interests. The proposal was funded and in May of this year a delegation of 7 Chinese scientists was hosted at AND, NET, CPR, KBS, KNZ, and SEV LTER sites over a two and a half week period. The culmination of their visit was a 1-day workshop on data management at the University of New Mexico (SEV) in May of 1991.

The Chinese delegation was impressed with the relevance and utility of long-term studies. They were also convinced that networking is essential to LTER's success. The delegation expressed a desire to use LTER as a model, finding enough applied aspects (vs "basic science") to be convinced of the applicability to Chinese environmental problems. LTER Data Management stood out as "state of the art." The Chinese expressed a real need for technological assistance in Data Management and outlined a plan wherein:

1. A LTER delegation would visit China including Scientists and Data Management Personnel to assess the needs of CERN and CERN data management.

2. Six Chinese Data Management Personnel would then come back to the US to attend an intensive 3-month training course in Data Management.

3. 12 LTER Data Management Personnel would then return to China, and together with the six newly trained, would host a workshop for 40+ individuals from CERN research stations.

Carl Bowser (NTL), William Michener (NIN) and James Brunt (SEV) will be the Data Management representatives in the initial delegation. This delegation is set to depart from the US August 30,1991 and will return September 21,1991. While in China the group will spend several days in Beijing together and then split into three groups with one DM in each group to visit and assess field sites. The field site visits are divided into 3 major areas: Transition, Agriculture and Hydrobiology, and Forestry. The groups will then rejoin in Beijing to visit the major institutes that CERN will interact with. A report will be produced from this visit and assessment that will direct the DM training and workshops to follow.

The data management contingent of the delegation will use information on assessment of data management activities and minimum site capabilities in their work of assessing the CERN capabilities.

This activity will provide opportunities for network-to-network, site-to-site, and scientist-to-scientist interaction with the Chinese. Opportunities for participation in proposal writing, training, and workshop activities will also be available for interested data management personnel. Susan Stafford and Barbara Benson have prepared a draft of curriculum to be built upon in developing a training course in data management for the Chinese and others.

F. Data Access Policies (John Porter, VCR; William Michener, NIN)

Discussion was initiated at the 1990 Data Management meetings and continued during the All Scientists' Meeting where a group worked on a list of concerns. Guidelines for site data management policies were developed and sent out to the sites. The LTER Coordinating Committee appointed Myers, Hobbie, Magnuson, Michener, Stafford, and Porter to develop a framework for individual site policies. These site policies are now a requirement for LTER renewals. The future strategy is that each site would formulate it's own policy and the NET office can compare them to establish a common network wide policy.

G. GPS Proposal (William Michener, NIN)

The GPS proposal was recently funded allowing for 2 high-precision units to be purchased. Two additional Pathfinder units will also be acquired. A week long training workshop and field exercise will be held Sept.30 - October 4 in Boulder, CO, with the emphasis on high-precision units. Substantial theory on data processing will be presented. Training will also be given on standard surveying with theodolite technology. $500 per site in travel will be covered, and 1 person per site should attend, preferably a computer literate scientist or technician who will be using the equipment. Sites need to develop a list of objectives for the units since availability will be limited. Tentatively they will be available for use October - December 1991 and June-August 1992.

H. Databits (John Porter, VCR)

Much credit is due to **th~** contributing authors, and DATABITS has been generally met with positive response. John Porter requested feedback on the size and scope of DATABITS and it was generally felt that all is ok for now. Electronic access is available on LTERNET.

I. LTERNET Status (Rudolf Nottrott, NET)

- There are 2 Sparc stations networked together in the network office, one is the net server and the other is the GIS server. The hardware breakdown at the network office was described. The SRINTNET connection should be up in a month or so with the initial costs of 600 + 740 monthly charge for 9600 baud plus connection charges. LTERNET will pick up the bill for one year. The Electronic BBS is operational with C-News message software. A local version of USENET

is available and is compatible with the 'rn' program. The BBS currently has only DMAN news group posted. Rudolf will be moving away from mailing groups towards news format. INGRES will be **implemented.** Accounts are available to LTER scientists that want them.

J. LTER History/Reference File (Rudolf Nottro U, NET)

Following the decision at last year's data managers meeting, an "LTER Data

Management History and Reference File" has been compiled at the Network

Office. Copies of this file are available from the Network office on request. Two

copies of the file were given to the two new data management representatives,

Karen Baker for the Antarctic Palmer Station and Janet Greenlee from the

Jornada.

K. LTERnet Bulletin Board (Rudolf Nottrott, NET)

The C-News electronic bulletin board (BB) and news software is now running on LTERnet. The first newsgroup (equivalent to a bulletin board section) is 'dman', with aliases 'datamgt', 'datamnt' and 'datamng'. The 'dman' newsgroup has been used to distribute messages relating to the 1991 data managers meeting. To do this, the newsgroup 'dman' has been used transparently in place of the former mail forwarding group of the same name, with copies of every posting being distributed to all members of the former 'dman' group.

'rn' (for read news) is presently the main program used to access the C-News BB on LTERnet. Other access programs, such as 'nn', the **Xli** -based Xrn and the Emacs-based news reader are planned to be installed later. If you are not able to directly read the LTERnet BB (mostly those of you without an Internet connection), you may receive copies of all postings to particular section by email. In fact, BB sections usually will have mail groups attached to them, with all postings being forwarded to those groups.

Using **C-News,** an extremely widely used software package freely available at many locations on the Internet, has the following advantages:

C-News adheres to Internet standards (e.g. RFC 822 and the "Standard for the Interchange of USENET Messages" RFC 1036)

The "readers" (programs to read BB/news, such as 'rn'), are as wide-spread and freely available as C-news. These readers can be configured such that they read BB/news on remote machines (enabling you to read the LTERnet BB without ever leaving your local computer).

C-News interfaces easily and "naturally" with e-mail.

C-News and the programs to read it are relatively small (4 MBytes versus approx. 40 MBytes for ANDREW) and they have been compiled on almost any hardware/software. This will be beneficial for those of you with an Internet connection who want to read LTERnet BB remotely.

C-News is relatively stable. To cite the C-News creators: "C News has been tested pretty thoroughly. We're also thoroughly sick of it and make no promises that there will ever be another release. We may, repeat ***~~y*,** provide updates via some appropriate newsgroup..." (In contrast, the ANDREW system is still under constant development.)

Those who have already had access to the global USENET will find themselves on familiar territory.

There are more good reasons that make C News the BB of choice for LTERnet at the present time. Maybe the bulletin board itself will serve as a medium to discuss more details of this subject.

To get information of how to use the 'rn' news reader, type "man rn <RETURN>" on LTERnet. A summary of 'rn' operations is also scheduled to be published in the FaIl 1991 issue of DataBits.

The plan of using the ANDREW Message System from Carnegie Mellon University as the LTERnet BB system has been postponed for now due to the reasons listed above. ANDREW is installed on LTERnet and can be run as /usr/andrew/bi n/messages (or Iusr/andrew/bin/vui for vtl 00 emulators), but it's use is unsupported.

III. WORKING GROUPS REPORTS

III. A. DATA ACCESS FACILITATION

Committee:

Barbara Benson (NTL) session leader

Don Henshaw (AND)

Tom Kirchner (CPR)

Mark Klingsmith (BNZ)

John Porter (VCR) session leader

Cindy Veen (HBR)

The objective of the group was to find ways to facilitate access to LTER data by researchers. Several LTER sites have had successes using interactive programs to distribute data. HBR LTER has an online bulletin board for data distribution. It concentrates on pure data archiving with few query or browse capabilities. AND LTER has a system that sends out a form automatically to a data requester (see

Appendix B). It can accept e-mail or paper responses and does automated processing of requests (getting permission to release the data from the responsible **P1,** outputting documentation and keeping a log of requests).

Interactive data access (IDA) systems are one way to speed up the delivery of data to scientists and are discussed extensively below. Other access mechanisms that were briefly considered are online catalogs (already implemented for LTER core datasets on LTERNET), data publication, workshops based around specified datasets with a strong interactive component, and educational activities such as presentations, short articles, and training workshops on data access.

Use of interactive data access systems has a number of advantages over manual methods. First, it reduces the time required to fill a request. In principle, an investigator could have access to data only minutes after defining a data need. For the information/data manager it means that less time is spent filling requests. A properly constructed IDA system would make it easy to locate needed datasets. This capability would be especially useful for workshops and working groups and would facilitate data exploration and browsing by individual researchers. In essence, it would help to put data at their finger tips.

Some additional advantages of a good IDA system are that its use need not be limited to a specific site (although with built-in security, access can limited). IDA systems can provide data in multiple output formats (e.g., ASCII, spreadsheet, database and statistical package formats) and support and consolidate fundamentally different types of data (GIS vs ASCII). IDA systems can also improve records of what data is being used and can even automatically notify PIs of who is requesting data they collected. Finally, IDA systems can include

excellent security control. Users can be grouped into access classes to eliminate the possibility of unauthorized access to data.

Some disadvantages of interactive data access systems are that they are susceptible to network problems and can be difficult to implement and debug. If appropriate security measures are not implemented (as in open systems that are accessible to everyone -- even **non-scientists)**$_1$ there is a potential for misuse of data.

Table 1: Characteristics of an ideal interactive data access system.

ACCESS

Easy access to meta-data should be provided so that scientists can rapidly locate the types of data they need. This may include a browse mode for scanning meta-data, and logical functions for selecting datasets based on **keywords**$_1$ titles, etc.

Strong security measures must be implemented to assure that data are only available to authorized persohs.

Interfaces must be easy to learn, and user friendly. This requires context sensitive help and may include graphical interfaces, such as **pop~own** menus.

Export functions should support data transformation into a variety of formats such as **ASCII** or generic file structures for specific applications (Lotus **1-2-3,** Excel and DBase). Further it should permit subsetting of the data and have buift-in aggregation and linking capabilities or facilitate linkages to external software that can provide those capabilities.

Logs of system and dataset use should be automatically maintained, both **for'** monitoring data set use and to assure that system security is being maintained.

IMPLEMENTATION

To facilitate **inter-site** translers, an interactive data access system should be accessible from remote sites. Ideally the interface would be uniform across sites to minimize retraining for investigators.

Flexibility is an important feature. Muftiple computer hardware platforms should be usable. The interface should be able to sit on top of different database management systems with a minimum of reconfiguration. Thus, the system should be flexible in the way it permits data to be stored.

The system should be able to manage different types of data **(e.g.,** binary vs ASCII).

Software used to implement the interface system should be available "off the shelf", but only if it fulfills other design criteria. Additionally, software that is distributable (no license required) is desirable.

Security of **daia** is an issue of special significance. Many LTER site data access policies permit PIs to limit access to specific datasets until permission is

obtained. This means that not all data can be made immediately available online. With proper procedures, interactive data access systems can still be used, albeit with a slower turnaround time. For example, systems might automatically generate request for access and forward it via e-mail to the appropriate P1 and await a reply before permitting a dataset to be downloaded.

Another aspect of interactive data access systems is the possibility of creating automated data entry systems that allow P1's to directly enter both data and documentation (meta-data) directly into a database system. Advantages of such a system are that it saves time entering meta-data (it need only be typed in once, not typed in by a P1 and then subsequently retyped by the data management staff), and gives a P1

control over quality assurance of his data and documentation. Entry systems can lead to more reliable data and documentation if built-in checks, for ranges and internal consistency are incorporated.

A potential disadvantage of interactive entry systems is that they may fail to fully incorporate the expertise of information/data managers into the entry process. Some individual PIs may need help in evaluating the adequacy of meta-data. Although, interactive entry poses some advantages for quality assurance, via range checks, it does not facilitate traditional quality assurance procedures such as double entry systems. **Additionally,** if the system incorporates editing, as well as entry features, it would allow changes to archival datasets. Because it may be necessary to backtrack to an eariier version of a dataset, audit trails need to be included.

Implementation of interactive data access Systems at sites within the LTER network is a desirable goal. To further development of such systems a number of proactive steps could be taken. These include:

1) The LTER network should distribute **guideJines** for interactive data access systems. Suitable media include the Information/Data Management Workshop report and DATABITS, the LTER data-management newsletter.

2) Encourage NSF to fund developmental projects, both through technological supplements and independently funded grants.

3) Conduct a survey of existing systems and future plans for interactive data access systems within the LTER network. The results of the survey should be published in Databits and next years Information/Data Management Workshop.

4) Establish section for interactive interface development on LTERNET BBS to facilitate an exchange of ideas among developers.

5) Have individual information/data managers explore use of interactive data access systems outside the LTER that could serve as models for development of LTER-based systems.

6) Arrange for demonstrations of interactive data access systems at next year's information/data management workshop.

Proceedings of the

III. B. MINIMUM SITE CAPABILITIES AND ESTABLISHMENT, EVALUATION, AND REVIEW OF LTER DATA MANAGEMENT 1. MINIMUM SITE CAPABILITIES

Committee:

Karen Baker (PAL)

Carl Bowser (NTL) session leader

Gil Calabria (CWT)

John Gorentz (KBS)

Janet Greenlee (JOR)

Barbara Benson (NTL)

Research data management is an activity common to all NSF-LTER **sites,** and is considered essential for continued activities at the sites. Recently the National Science Foundation has underscored the importance by selecting at least one person with a data management background for NSF-Site Review teams. The demands for data management have grown with the increasing amounts of data obtained from each site, increasing complexity of data management systems, and the need to work both at the intrasite and intersite levels.

Greater demands on the data management personnel at each site can be attributed to: (1) increasing complexity of database platforms; (2) larger and more sophisticated data manipulation and data analysis programs; (3) increasing number of investigators at sites, each with their own commitment to collection of long-term data for research purposes; (4) the shift to automated data collection and data logging devices; and (5) the need to share data across sites.

As data management activities grow in scope and complexity it is also evident that all sites share a common need to be able to collect, archive, and transmit data in forms that optimize the process of aggregation of data into larger data sets, and to facilitate the exchange of data both within and among sites. Recognizing the individual character of each LTER site, differences imposed by the location of the research site relative to the base LTER institution, the character of individual investigator interaction, and differences in hardware/software preferences by researchers and data managers, demands flexibility in the approaches to data management. Similarly the **need'** to aggregate larger data sets, to communicate data, information, and programs across sites, and network wide acceptance of minimum standards for cataloging and documentation of data sets calls for some minimum capabilities for data management activities at each site.

To this end the LTER program endorses the concept of "Minimum Standard

Capability" with regard to data management activities at each of the LTER sites.

The intent is to ensure minimum capability at each site, and not to specify the

manner in which the capability is implemented. The framework of capabilities

promotes standard capabilities at all LTER sites, while at the same time is conscious of the individuality of approaches at each site, and encourages innovation, originality, and experimentation in data management at each of the sites.

Acceptance of common data management standards will encourage adherence to high standards and will reinforce the commitment that NSF has made to data management in this program.

A "Minimum Site **CapabilitiesH** draft document was prepared at the workshop and is being reviewed by all LTER sites for discussion at the next LTER Coordinating Committee meeting. This document includes discussion of continuity of data management at sites, documentation, and other site-specific capabilities.

2. ESTABLISHMENT. EVALUATION AND REVIEW OF LTER DATA MANAGEMENT

Committee:

Karen Baker (PAL)

Carl Bowser (NTL) session leader

Gil Calabria (CWT)

John Gorentz (KBS)

Janet Greenlee (JOR)

Barbara Benson (NTL)

Items Discussed:

1- Site Evaluation and Review Process by NSF

2- Minimum Standard Capability of Data Management

3- Information for Establishment of a Data Management System

Site Evaluation and Review Process

Data management has been an integral part of the LTER program since its beginning as is appropriate for a program which will depend on reliable long term records which are readily accessible both within and across the LTER sites.

The LTER data managers enthusiastically endorse the concept of the addition of a person with strong data management background on the NSF site review teams. In **addition**$_1$ data management personnel should be included in reviewing renewal proposals. Including a thorough review of data management will encourage adherence to high standards and will reinforce the commitment that NSF has made in this program to data management.

To further enhance the review process, adequate time should be allocated for review of both research and data management processes and status. We propose that the review process be extended by one half day to allow for one-on-one meetings of review committee members with site groups to undertake detailed reviews of several elements of the site research. Included among the detailed review would be an examination of the data management activities in terms of the MSC (Minimum Standard Capability) concept discussed below. The specific one-on-one evaluations would be decided by the Coordinating Committee, but might include one or more of the following areas:

(1) Data Management system review

(2) Review of approach and progress on the five core research areas of LTER program

(3) Review of the structure of LTER site administration and allocation of resources between shared and individual funding.

(4) Review of progress on intersite level research activities

To assist in the review, a Minimum Standard Capabilities (MSC) document (see next section) has been compiled listing the areas which should be addressed in any data management system. The MSC document should provide all LTER sites with an internal checklist against which they can evaluate themselves as well as prepare for site reviews. This process should lead to a standardization across sites in terms of quality and performance.

Minimum Standard Capability (MSC) of Data Management at LTER Sites

The committee proposes that all sites accept the concept of Minimum Standard Capability (MSC) with respect to Research Data Management activities. The MSC standards should be drafted by this committee for acceptance by all data management groups at their respective sites, and that the revised MSC document be presented for approval at the next Coordinating Committee meeting. An outline of such a document was produced during the meeting. The intent of the MSC concept is to focus on the capabilities of each site and not the specific implementation of these capabilities. The Data Management Committee recognizes that implementation of data management is dependent on many factors that are site specific, but that coordination of data management activities at the intersite level will require some cooperation in the area of standardization of reporting units, catalogs, documentation, etc.

Information for Establishment of a Data Management System

The LTER Data Management Committee endorses the need to provide information to interested parties on how to set up a modern, research oriented, data management operation. The working group recommends that this item be addressed more fully in the near future, and presented at the next Data

Management meeting scheduled for August 1992. Information can be made available through:

a) We recommend that public domain information files be **mai~tained** on the Iternet server that would provide information on the establishment of data management systems in ecological **research.** There are documents useful to any site getting started in data management or expanding into new areas. The LTER history file recently compiled by the network office addresses this need. The index to this file could be extended to include references to other documents that don't necessarily relate specifically to LTER, but which have been found useful at one or more sites. The mechanism for updating this list can include submissions by various **sites,** preferably accompanied by a few sentences explaining the utility or relevance of the **document.**

b) We recommend continued improvements in the information available to help each site learn what resources are available to use in evaluating and improving its own data management system, and to interact with data managers outside LTER. The Data Management group recognizes the need for funding to support justified travel to other sites, and seeks to obtain support from either the Network Office Budget or other appropriate funding source.

c) We recommend that the expertise in the various areas of data management be made known through an expansion of the current LTER directory. The directory currently lists the organisms and research interests of each investigator. We recommend that it be extended to list the particular areas of expertise of those involved in data management. It could identify those persons with experience in cataloging systems, DBMS software systems, interactive retrieval systems, and in other categories suggested by the list of Minimum Standard Capabilities. This would include information that would provide interested people in the types of systems each works with (e.g. UNIX, Macintosh, VAX, SUN, PC), their data management expertise (ingres, Foxbase, etc.), communications capabilities (internet, modem, telnet, etc.), and common software used for research, modeling, and data management (ingres, Excel, Lotus, SAS, SPSS, S, etc.).

III. C. FUTURE OF SCIENCE AND TECHNOLOGY SUPPLEMENTS AND FUNDING FOR LTER DATA MANAGEMENT ACTIVITIES

Committee:

Mark Klopsch (AND) session leader

A. El Haddi (CDR)

John Vande Castle (NET)

Background

Tech nological Supplements Requests (TSRs) were initiated by NSF in 1988 to allow LTER sites obtain state of the art technology for research. In particular, these supplements were for specific new and innovative tools to enable or expand research capabilities on a collaborative basis. Two types of supplemental requests were solicited: those for site level enhancement and intersite projects like the Global Positioning Systems.

The site level supplements were first used to implement a Minimum Standard Installation (MSI) for LTER sites which had been defined through a series of **meetings,** reports and workshops (Shugart et al, Gorentz et al, Foster and Boose). The key focus of the MSI has been to insure that each site has Geographic Information System (G IS) capabilities, Internet access, electronic communication with the field site, and appropriate archival storage.

Although designed to allow the sites to move up to a common framework without imposing changes on the site specific computational environment, the Supplements have largely defined the technical capability of the collective LTER Network. The communication, analysis, and GIS capabilities provided by the Technical Supplements have created a powerful framework for collaborative research which is already serving as a model for similar research networks.

Although not all the sites have yet reached the current Minimum Standard Installation and its incorporation into research programs is still underway, the benefits resulting from the infusion of technology are already obvious. While in most cases the equipment is only just starting to produce real output, the most important long term benefit, a conceptual change on the part of researchers, is well underway. Rather than being limited to site specific studies, researchers are realizing that the new

technology provides tools to expand research horizons to broader spatial and temporal scales. This change in scale has necessitated interaction with other agencies and researchers, laying the foundation for intersite comparisons.

Recommendations for Future Technical Supplements

Since the supplements are responsible for the technical capability of the LTER Network, the continuation of the program is essential for further evolution of the technical capabilities of the LTER Network. With ever tightening budgets it is important to continue to improve the quality of proposals.

We encourage the continuation of both site and intersite technical supplements and have addressed our recommendations for each separately. Our recommendations should be considered as more a framework for discussion than an exclusive set of proposals.

Site Level Recommendations:

The most important goal of the site TSRs has been to permit each site to achieve the Minimum Standard Installation. While the MSI is implemented at the site level, it serves as the foundation for intersite collaboration. The definition of MSI is dynamic and the fast pace of technology requires that it be continually updated. The MSI must continue to look toward the future since often new technology must be in place before research can even be conceptualized. Because the LTER Information Managers need to play an important role in the regular revision of the MSI, we recommend that its review be a standard agenda item at LTER Information Management meetings. Any recommendations resulting from that meeting would be passed on to the Coordinating Committee or be presented at a special TSR workshop.

Although obvious overlap exists, changes to the MSI can be roughly split into two types: additions and improvements. Additions are necessary to correct oversights in the original MSI and to incorporate new developing technology. Although other possibilities exist, we would recommend one very important addition to the MSI this year ... the acquisition of the necessary hardware and software at each site to support the new interoperable SQL server interface which should hit the market during 1992. Since most sites already have Internet connections and use database packages from manufacturers participating in the development of this interface, for most sites this addition should cost little more than the price of a software update. While the cost would be comparatively low, the benefit would be to significantly enhance the network capability for real time intersite data sharing and create the real possibility for the

exchange of database **applicati9ns** between sites ... even those with differing database management systems.

Improvements to the MSI are necessary because old solutions cease to be acceptable as needs and technology change. Today we find the high technology of a decade ago laughable and we can barely conceive of the research possibilities for the technology of the year 2000. Even during the short life of the TSR Program new computers have been introduced, superseded, declared

obsolete, and finally dropped from software support. As technology changes and utilization increases, the definition of what constitutes an acceptable Internet connection, archive device, or GIS facility will also need to change.

The storage space at existing LTER GIS installations, considered adequate at their inception, is becoming a major limitation as many of the sites gear up to work with GIS information and satellite imagery covering larger areas. At least several sites at the IM meeting indicated a need to deal with over 30gb of data. Most standard Read/Write laser disks hold only 250-300Mb on a side ... too small to hold even a complete Landsat scene. Although the capacity of the 5Gb Exebyte tape units is handy for backup purposes, the units do not provide the random access necessary for use as a regular storage media. Therefore, one example of an MSI improvement candidate is a large on-line data storage device (eg. 50gb optical jukebox). While the price of good quality laser jukeboxes with caching remains expensive, the cost has been dropping rapidly. Hopefully, with some assistance, they will become affordable before the lack of data shortage seriously restricts research. Other improvement possibilities include network upgrades to enhance communication and the acquisition of fast GIS hardcopy devices.

As the sites approach the goal of MSI it is important to remember that many other aspects of LTER research can benefit from an influx of new technology. Although many of these items would be of interest to many or all LTER sites, some such as a network of sensors to monitor water level changes over large areas may be very site specific. As MSI goals are met, the TSR program provides the opportunity for some very creative site specific uses of technology.

Many sites need new field and lab equipment: data loggers, hardware and software to speed data acquisition, and local LAN equipment. Another common need is equipment for enhancing communication of results and scientific presentation, specifically items such as color laser printers, film recorders, equipment to prepare multimedia presentations, etc. While some of these items may not sound very flashy and will require careful justification, they can make significant contributions to the quality of data collection and presentation.

Intersite Recommendations:

Many of the suggested ideas for TSRs were for items to be shared by all sites or groups of sites. These included tools for scientific visualization, new methods for processing information, and GIS modeling (SGI workstations etc). Other possibilities included new types of sensors and a whole range of devices

employing video and image recognition technology to do anything from estimate canopy cover to count plankton.

We would like to highlight two possible intersite proposals. The first would use the recommended SQL interoperability addition to build on the work of Kirchner and Gorentz to develop a truly network wide query system. The proposal would probably include a planning workshop, an implementation component for each site, and an LTER information center to act as the primary location for extrasite queries. In addition to several levels of "public" core data, the system could be used as a means of exchanging white page data on individuals, available equipment and software, and any information that we would like to share across sites. The benefits to both intersite and extrasite collaboration could be enormous.

A second intersite candidate we discussed, was the collection of high resolution (m^2) imagery for each site generated from low elevation overflights using both video and satellite type sensors. The information at these resolutions is invaluable in identifying and quantifying small scale patterns and extremely useful in scaling information from a site to landscape level.

General Recommendations

The group had a number of general thoughts about future Technical Supplements Requests. It is clear that the TSR program is the key to the technical competence of the LTER program. As a result it is very important that sites get a good oppQrtunity to suggest what they think is important. We would recommend that in addition to being a part of the IM meeting, suggestions for technical supplements should be either the subject of a workshop or be an agenda item at the next Coordinating Committee meeting. This **~ould** give sites time to formulate somewhat broader based recommendations than those likely to come simply from the Information Managers. It would also provide a forum for discussion of exceptions to individual site funding limits for specific projects or new sites requiring a major technological infusion.

The LTER Information Managers understand the need to avoid the use of TSR money for continuing personnel support. However, we feel that the innovative policy of providing initial personnel support is a good one and in many cases essential when introducing new technology. We encourage NSF to consider using this policy again under similar circumstances and ask that they continue to consider the expense of installation and training as an integral part of the cost of new technology.

Frequently the software associated with new instrumentation is either not fully developed or inadequate for research purposes. Therefore, before the

- equipment can be fully utilized, a significant amount of time and expense are required for the development of adequate interface programs. We encourage NSF to treat funding requests for software development in the same fashion they would treat the acquisition of commercial software packages.

So far, most of the exchanges between sites and agencies have consisted of email and **small,** fairly static data sets. As these exchanges progress, it will become increasingly important to share rather than simply exchange such information in order to keep it properly revised and to avoid problems resulting from multiple working copies. The types of remote access needed share data between the field sites, research institutions, and cooperating agencies make heavy demands on network resources. Even without proposing such things as teleconferencing, voice over data, intersite editing sessions, or centralized satellite imagery libraries; it is clear that our intrasite and extrasite communication needs are going to increase dramatically. Yet, even now the bandwidth, both local and Internet is basic obstacle to data communication and sharing. Although the necessary improvements are largely beyond the control of the LTER program, it is important that we encourage them whenever possible.

Another major concern that has surfaced is the cost of supporting the emerging information management and analysis system. Even before the Technical Supplements, the cost of information management was growing faster than budgets. While the Technical Supplements have been a significant benefit, they have also aggravated this situation. For most sites, the MSI has resulted in the addition of at least one GIS person and a substantial amount of equipment and software with annual maintenance costs running between 10 to 20% of the initial price. Therefore, the annual cost to each site for basic MSI support can easily reach $100,000.

Ideally these costs would be incorporated into the base funding of each site at renewal time. We would encourage NSF to look favorably at such requests or consider some alternative mechanism with better accountability and which doesn't penalize sites with several years before renewal. In the meantime, without NSF support to pay for these continuing costs many sites will be faced with tough choices. Due to the diversity of sites there is no single solution. Some sites are in positions where they can share the equipment with other agencies and researchers, thereby splitting the cost and enhancing cooperation at the same time. Other sites are more isolated and might even be faced with the inability to adequately support the equipment.

Conclusion

The Technical Supplement Request Program has been an enormous boost to the LTER program and has greatly enhanced intersite and interagency collaboration. The LTER program now needs to evolve beyond the concept of the MSI, toward a concept of Minimum Standard Capabilities which emphasizes not just a basic set of equipment but a program which includes the necessary support to use it adequately.

With a significant portion of the initial MSI goal met, the Technical Supplement program has entered a critical phase. If the LTER program is to maintain it's leadership and act as a catalyst for collaborative research, it is important that the TSR program continue. Although we have proposed a few specific items to scale the GIS facilities up to production levels and significantly enhance intersite communication, a crucial factor in the continuation of the program will be the quality and innovation of the proposals during the post MSI period.

III. D. FUTURE DIRECTIONS OF DATA MANAGEMENT IN LTER

Committee:

Caroline Bledsoe (NSF)

Emery Boose (HFR)

John Briggs (KNZ)

James Brunt (SEV) session leader

William Michener (NIN) session leader

Rudolf Nottrott (NET)

Plus Significant Contributions From: Karen Baker **(PAL),** Barbara Benson (NTL), Carl Bowser (NTL), Gil Calabria (CWT), John Gorentz (KBS), Janet Greenlee (JOR), John Porter (VCR)

Although there was considerable overlap with other working groups, our primary objectives were to examine and reccommend future directions for LTER Data Management. Discussion focussed on outreach activities. Following a brief description of general outreach activities, we propose several specific actions which could greatly facilitate research activities within LTER and would increase the visibility of LTER Data Management efforts in the larger ecological community. These include:

1. Completion of a pilot research/data management project which would benefit the entire LTER Network.

2. Development of an LTER Data Management Slide Presentation.

3. Developing workshops on new analytical/statistical techniques.

4. LTER Bulletin Board/Email.

5. LTER Data Publication.

6. Expanding the scope of the 1992 Data Management Workshop to include non-LTER Scientists.

7. Developing Collaborative Research with the Soil Conservation Service.

8. Outreach Activities with the Chinese Ecological Research Network.

Outreach Activities for LTER Data Management

For LTER to play a leading role in the ecological community, we recognize the need to increase our outreach initiative. This outreach should be exercised in all three levels of the LTER network: intrasite level, intersite level, and overall community level.

Data management training in hardware, software, and data management system design should be provided to data management users within the LTER Program as well as to interested outside users. To some extent this activity overlaps with the issues raised in the section above on "Information for Establishment of a Data Management System". Outreach activities could be provided at several levels.

(1) At the intrasite level (relationship of DM to research at sites)

(2) At the intersite level (data managers at other sites within LTER)

(3) Outside the LTER project.

a- For other research groups or individuals who are interested in setting up a data management system.

b- For agencies or research groups who have established data management systems to interface with LTER data management activities.

At the intrasite **level,** the interaction between the Data Manager (DM) and the Principal Investigators (PIs) becomes the basis for the outreach initiative. An essential DM role in LTER has been the documentation and archival of datasets. Even though this is a very important aspect of data **management,** we have recognized the need to go one step further and expand our role from dataset **"librarians"** to a more interactive "research facilitator" role. To increase this research role, the DM should provide tools to make archived datasets more accessible to the PIs. Most important of all, the DM and the PIs should try to establish links among these datasets to spur new, previously unexplored ecological questions.

To increase the interaction among the LTER sites, and to assist with the implementation of the MSC concept, the idea of an intersite cooperative training program was suggested. This program would support the exchange of expertise among all the LTER sites. For example, if a particular site needs some help in setting up their QA/QC protocol, the DM would visit another site which has this aspect of data management already established for training. These training sections would greatly improve the overall LTER network status and serve as an incentive to intersite research.

Outside the LTER program two types of outreach activities were considered, for non LTER sites wishing to establish comparable data management systems, and agencies with existing data management activities that wish to find ways to link or merge data sets for use at a larger scale of synthesis. This latter type of interaction will be especially important if LTER sites are to be integrated with world networks of ecological data (networks of networks).

Outreach could be provided in three ways: 1) information available in publicdomain files on Iternet; 2) written documents prepared by willing LTER data managers on the subject; and 3) offering workshops or short courses in conjunction with national meetings, such as ESA. This latter mode was enthusiastically endorsed by the working group, but will require further discussion at future Data Management meetings to consider such matters as the specific types of short courses to be offered, participation, and

coordination with the meeting organizers.

Pilot Research/Data Management **Project**

LTER data managers could offer to assist a current Network-wide project (e.g. the decomposition or DIRT experiments). By incorporating data management planning from the **outset,** and by drawing on the combined experience of the entire data management **group,** the project could illustrate the benefits of good data management and serve as a prototype for future intersite studies.

Development of an LTER Data Management Slide Presentation for Outreach

William Michener and Rudolf Nottrott will prepare a proposal which will seek to acquire the funds necessary for developing an LTER Data Management Slide Presentation. The background for this proposal, objectives, and logistics are discussed below.

Many researchers within the LTER Network routinely use information management facilities and expertise provided by data managers at their sites as well as other sites. The data management capabilities made available to LTER researchers include data entry, short and long-term data storage, electronic communications (data and text), data processing (statistical, modelling, etc.), data base management, and many other functions. Frequently, the full range of capabilities, including equipment and services, is not known to LTER researchers outside the particular site or to members of the larger ecological community. At their annual data management meeting in August 1991, the LTER data managers therefore resolved to compile an LTER Information Management Slide Presentation Set which would highlight these capabilities. In addition, data management research activities which could potentially benfit both LTER and non-LTER scientists will be highlighted.

The LTER DM slide presentation set would consist of 2 to 3 slides from each of the 18 sites highlighting the sites' information management facilities and expertise, and how they facilitate the long-term ecological research at the site and across the network. One Set will be sent to each LTER main administrative site for use by the site's researchers and data management staff. In addition, a copy of the Set will be maintained at the Network Office for archival purposes and possible loan to researchers within and outside the LTER network.

The presentation set will come with a short abstract describing each slide. Thus, a presenter of the set can be sure to accurately convey what the sites consider their salient information management capabilities, while still having the freedom of elaborating and filling in with other related information.

Contents of the Presentation Set will include: (1) one or two slides from each site depicting specific

research activities (e.g. decomposition experiments, meteorological monitoring stations, etc.) or site characteristics (e.g. aerial or

satellite **views**[1] dominant vegetation communities, etc.); (2) one slide which depicts each of the 5 core areas (total of 5 slides); (3) a brief history of the LTER Network since it's inception; (4) background text slides which define key or illustrate elements of the data management process (e.g. data entry, QA/QC, archival, data flow charts, etc.); (5) up to one or two slides from each site which highlight major data management accomplishments or illustrate research being performed in the field of data management/analysis.

Developing Workshops On New Analytical/Statistical Techniques

Hardware is no longer the principal limit on data analysis across the LTER Network. Sophisticated techniques for data base design and query, spatial statistics, remote sensing analysis, and other applications of our computer technology are being explored and developed at many sites, and could be shared across the Network at workshops devoted to the most promising of these techniques.)

LTER Bulletin Board/E-mail

The new bulletin board system will largely replace the current mail forwarding groups, reducing the day-to-day E-mail that many individuals now receive. The E-mail forwarding system will likely be extended in the near future to include the long-term monitoring section of the ESA. Eventually LTERNET and similar systems will probably service much of the national ecological community.

Data Publication

An informal working group was established to discuss issues related to the direct publication of data, for example, in the form of a CD-ROM. Data publications can take many forms, from individual data sets released to a specific recipient to copyrighted and peer-reviewed compendiums of multiple data sets. The group identified several potential advantages of publishing data over conventional methods of data management. These included: 1) security-- data that are widely distributed in a long-lived generic form are less likely to be lost than data that are maintained at a single location or within a single data management system, 2) more efficient access -- once data have been published, it reduces the time required to fill data requests, 3) outreach -- data publication can provide tangible products that are available to the entire scientific community, 4) academic credit

-- a data publication (especially a peer-reviewed one) is a citable product creditable to the investigators responsible for producing the data and

availability-- CD-ROM technology makes even the largest ecological data sets tractable for distribution. The group also identified several disadvantages. The first is that academic credit resulting from data publications may be quite small

and that suggesting such a radical approach to data management might undermine the credibility of information/data managers in the scientific community. Secondly, once data are published, errors and inconsistencies detected by subsequent quality control and quality assurance procedures cannot be incorporated into existing copies of the data sets. Finally, published data would be available to anyone, which might lead to misuse of the data.

Rapid decreases in the cost of CD-ROM (Compact Disk - Read Only Memory) technology has made it increasingly attractive, both for archival storage (media lifetime is estimated at 100 years) and data publication (capacity is approximately 500 MB per disk). Unlike WORM disk drives (which have similar archival properties and capacities but are not well standardized across systems), CD-ROM readers are widely available, inexpensive and incorporate standards that make them usable on virtually all computer platforms, from PC's to mainframes. Traditionally CD-ROMs have been targeted for mass-marketing. Costs of producing a master disk and 100 copies range from $1,000-$i ,500 with subsequent copies costing much less. Unfortunately, producing small numbers of disks with this technology costs almost as much as producing hundreds of disks. However, a new type of mastering unit that produces individual master copies for less than $25 each has been introduced. The cost of a mastering unit capable of producing individual CD-ROMs has dropped to $25-30,000.

There were brief discussions on the merits of copyrighting data, criteria that might be evaluated in a peer-review of a dataset (scientific value, adequacy of documentation), the level of aggregation appropriate to published data and problems that individuals have encountered when they attempted to publish raw data tables in a journal.

There were varied opinions on the merits and practicality of data publication within the working group. Nonetheless, several action items were proposed:

1. Support purchase of a CD-ROM mastering unit to be located at the LTER Network Office. Such a unit would be useful in itself for creation of archival copies of datasets for sites (CD-ROMs have a predicted lifetime of over 100 years, compared with a lifetime of 10 years for most magnetic media). It could also be used to support data publication efforts by individual sites.

2. Informally query LTER PIs about interest in producing data publications.

3. Explore producing a network-wide data publication targeting a specific theme or network-wide experiment (such as the decomposition experiment).

4. Encourage professional societies to produce data publications in support of their journals. Some societies already do this in the form of paper or microfiche copies, but electronically accessible data would be more useful.

Expanding the 1992 Data Management Workshop

A proposal to expand the scope of the 1992 Data Management Workshop to include representatives from the Organization of Biological Field Stations, Southeastern Association of Marine Laboratories, and Land Margin Ecosystem Research sites is being developed for submission to the National Science Foundation.

Collaborative Research With the Soil Conservation Service

Dr. Lawson Spivey from the Soil Conservation Service met with the LTER Data Managers to discuss the potential for collaboration between LTER and SCS. The SCS has recently taken a significant interest in global change research and believes that collaboration with LTER can result in mutually beneficial research. SCS was very interested in LTER Data Management and GIS activities. The Data Managers also proposed several ideas for collaborative research with the SCS. Potential cooperative projects include: (1) examination of the relationship among hurricane frequency, soil characteristics and vegetation regeneration; (2) developing or refining soil carbon models; (3) mapping groundwater recharge characteristics in relation to soil type; (4) refining and expanding the classification of wetland soils; (5) examine soil genesis in wetlands, specifically the conversion of agricultural fields to marsh; (6) evaluation of spatial and temporal distribution of soil moisture in the alpine tundra; (7) compilation of soil profile descriptions and integration of site level with landscape level soil characterizations; (8) evaluate desertification processes in relation to soil and groundwater characteristics; (9) incorporate soils data into hydrologic modeling efforts; and (10) examine impact of agricultural practices on soil characteristics.

It was pointed out during the discussion that many LTER sites have already collaborated extensively with the Soil Conservation Service. Many LTER sites have extensive groundwater well networks, meteorological stations, high resolution vegetation maps, and other research capabilities which could support SCS research efforts. The LTER sites could benefit significantly from access to fine resolution soils maps of their sites, detailed analyses of soil cores, and interaction with SCS scientists.

Interaction with CERN (Chinese Ecological Research Network)

The Data Management Committee discussed the "Terms of Reference" document (Attachment C of the 18 June 1991 Funding Proposal by **People's** Republic of China to the Norwegian Trust Fund). It was suggested that the "Minimum Site Capability" (MSC) terms would also be applicable to the CERN

group, and that they should be given a draft copy of the MSC criteria for consideration.

The group found it difficult to outline an assessment procedure for CERN without some detailed knowledge of the existing and/or planned operations. In general, however, the group would need to look at as much detailed information as possible about:

1. the types and quantity of data being collected and the intended uses

2. existing or planned methods/protocols for collection, entry, verification and archival for the data

3. data flow, tracing data from point of origin through processing, archival, and availability

4. data administration, quality assurance guidelines, documentation standards and data communication capabilities

5. hardware, software and networking capabilities.

The working group felt that other issues related to specific interaction with the LTER program or specific LTER sites, possible future workshops, and training courses, and issues related to participation of LTER scientists and data management personnel would be better discussed following the return of the U.S. delegation from China in September. The data management group is pleased to see the interests of the CAS-CERN groups in the LTER program activities, and is very interested in future participation with the CERN scientists and data management groups.

III. E. QUALITY ASSURANCE/QUALITY CONTROL

QA/QC are the procedures employed to improve the accuracy of a dataset through the reduction or elimination of incorrect information to an **acceptable,** documented level. These procedures require error identification which, in many cases, is a non-trivial task.

Errors may occur during manual entry of data, e.g. transposed numbers, missing **decimal.** points, and misplaced entries. If data are entered manually into a spreadsheet, one should make use of whatever quality control features (e.g. range checks, field **definitions**$_1$ and check some techniques) that are available with the spreadsheet software used. Moreover, it is recommended that raw data be documented and archived prior to further manipulation, e.g. variable transformation and calculations. A full-screen entry program is a good alternative to spreadsheet entry. It provides an acceptable level of quality control at the data entry stage and is comparatively inexpensive.

Errors may also be generated during the collection and storage stages of electronically **.** gathered data,e.g. via data loggers. These are frequently associated with long-term, continuously collected data (e.g. climatological data) which are made available quickly and widely distributed. Calibration and maintenance of the data collection equipment should be performed regularly and necessary standards observed. Sources of error occurring during automated data collection are:

1. External source errors

These errors result from conditions external to the sensors themselves **(e.g.,** accumulation of bird excrement or snow on sensors, frozen or corroded components, etc.). Such conditions are often easily

recognized and rectified. It is usually necessary to qualify the affected data with appropriate metadata, although determination of the exact time period of the affected data may be problematic.

2. Internal source errors

These errors originate within the data collection equipment itself (e.g., variable or inadequate sensitivity of sensors, hardware and software failures and incompatibilities). Such errors may be subtle and difficult to recognize or correct. Duplication of data collection equipment may be an appropriate solution, particularly in cases where the equipment is unattended for extended periods of time.

During the last decade there has been an increasing reliance on electronic equipment for the collection and storage of climatological data. Such data

collected at the LTER sites will undergo closer scrutiny in coming years for the purposes of the detection of signals of global climate change. Presumably any signals initially will be comparatively small. It behooves us to ensure that error ranges in LTER climate data are as tight as they possibly can be, in order to reduce the possibility that such signals are masked by error-generated **"noise".**

QNQC summary recommendations:

- Use graphics programs to visually inspect data for outliers, anomalies.

- Perform appropriate routine statistics when applicable. For manual entry verification:

- Use hardcopy output to verify against raw data sheets when possible. For data collection equipment:

- Calibration and maintenance of automated equipment must be performed regularly and necessary standards observed.

Action Items:

The QNQC group (Scott Chapal (NIN), Rick Ingersoll (NWT), Bernie Moller

(ARC), Michelle Murillo (SEV)) developed the following action plan:

1. A summary report on the topic of QNQC techniques for ecological data

management will be prepared.

2. Enlist 4 volunteers to form 1/2 of a working group designed to evaluate the quality of electronically gathered data within the LTER network and to subsequently **develop:** recommendations regarding the management of such data. This has been done with the members being Chapal, Gorentz, Ingersoll, and Moller.

3. Enlist 4 volunteers from the LTER Climate Committee to form the remaining half of the working group. This will be done in the next few weeks.

4. The working group will then develop a thorough survey that will provide detail on the current state of electronic data collection and subsequent management within the LTER network. Particular emphasis will be paid to electronically gathered **cIimatoIogi~aI** data. This survey will be distributed via e-mail to appropriate people at each of the LTER sites.

5. The working group will evaluate the results of the survey and then approach organizations (e.g., NWS, NOM, NCAR, etc.) who have potentially more expertise to develop recommendations for problem resolution.

6. If some problems still seem to have no solution within the scientific community, it may be prudent for LTER and some of the organizations above to approach the manufacturing industry.

7. The final objective will be to produce a report that summarizes the results of all of the working **group's** activities. It is intended that such a report should be possible before next year's data managers' meeting. Any information gleaned from the e-mail survey that might be of use to one or more sites will be made available as soon as is possible.

Proceedings of the

APPENDIX A.

SITE FLASHES

AND

We have recently submitted a Database Activities preproposal to NSF (Bob Robbins) to design an intuitive **user-friendly** interface to our existing Research Information **System,** select and evaluate appropriate software. This interface would allow users to manage their own data and allow independent access of data. On the outreach **front,** Susan Stafford and Barbara Benson are working on a preproposal to conduct a training workshop for CERN scientists (a **2-week** session at HJA site). Also, Yonghua Zhai from the Institute of Applied Ecology, Chinese Academy of Science, Peoples Republic of China will be visiting us for another 7 months assisting us with network support and Foxpro programming. We have recently been joined by Lisa Carlson who is working on a cooperative program with the USFS to set up a Research Natural Area Database. In hardware/software news, Foxpro 2.0 (which includes some SQL support and a compiler) has arrived. We have developed a Foxpro program to automate the process of handling data requests. The program logs data requests and generates form letters asking the requestor's intentions, requesting a release from the **P1,** and documenting the data. Over the last year we have installed a new Sparcstation II for GIS support, an SLC for use with modeling, expanded GIS storage with a laser disk, and added a DAT backup system. We have added a network at our headquarters site and are in the process of installing bridges to the campus LAN. We recently received news that tech supplement has been funded which will ease the pressure on our busy GIS machines and give our databank server more speed, space, and NFS capabilities.

ARC

1. We have been awarded two grants from NSF for three years. These grants will support the Experimental portion of our Aquatic and Terrestrial research at Toolik.

2. The LTER review team visited Toolik Lake for our site review on July 12 and 13,1991.

3. The Arctic Research Comission also visited the site on August 7, 1991.

4. Our database is on-line and available at Toolik for immediate entry of new data and for referencing old data.

5. We have been able to have much of the field data, nutrient and basic chemistry data analyzed and entered into the database by the day following collection.

6. The ever improving communication system at Toolik is working very well. This system

includes a satellite phone link with connections to the University of Alaska Computer Network and

access to our home accounts through Telnet. There is also a Fax machine at Toolik (907)659-

2417.

BNZ

Bonanza Creek has a full time data manager as of the first of this year. This position is supported 50/50 through LTER/UAF funding. Thanks to Phyllis Adams for fulfilling data management functions at BCEF

prior to this, while continuing with all her other responsibilities. We had our site review in July and were thanked for a well run site inspection. Received lots of

constructive criticism. I was chastised for spending too much time managing the SUN network. When I arrived on site my workstation was "still in the box". I now have it set up and fully operational. I have also spent a considerable amount of time with our SUN LAN and the results are promising. The SPARCserver 390 now provides all four of our SUN workstation with **FRAMEMAKER,** SAS and ARC/INFO. The network databases now in place will streamline system management and software upgrades. One of these workstations is in a separate building across campus and on a different subnet, which presents a challenging situation to administer. INGRES (2 node license) is installed on spruce and taiga. Spruce is being used for the climate database (inter/intra-site) and taiga for the data catalog and GIS data (as soon as we have the ESRI RDBI-INGRES module). I have been promoting site-wide data management at our biweekly meetings. I have instituted a set of forms and instructions for gathering dataset documentation.

CDR

We have acquired our Alphatronix INSPIRE (600 MB) optical drive and it is used as our permanent archival media. It is slow but much better than our old 5.25 inch disks and 150 MB magnetic tapes. Our data is still being converted from the old computer system to the new one. It will be all online as soon as the 1991 summer field season in over. There is great demand placed on our SparcStation (and on me) in terms of CPU usage and Disk space. It seems like the database size =exp(time*time*...). This is frustrating since our backup time has gone from about 1.5 hours to 5-6 hours/week. We hope to solve the problem by acquiring a 4 Gbyte tape drive. My preaching of powerful unix workstations is paying off. Project researchers are getting addicted to the power. Our PC-NFS and NFS-share setup are also protecting the unconverted (a few) and allowing them to use what they know without having to learn too many new things and thus concentrate on the science. In the **"real"** science arena, Cedar Creek LTER studies are paying off. A new paper looking at the effects of the 1988 drought on biodiversity and species loss was possible only because we had long term data for all individual species on the same locations, sampled in exactly the same way. Another paper dealing with chaos in the dynamic of grass populations (also only possible because of our long term data record) has just been tentatively accepted by Nature. We have also found some exciting patterns in an intersite comparison of soil properties and climate, on which Zak et al. have just completed a manuscript.

CPR

Our data management efforts for the past 18 months have focused on two issues; development of tools for easy data access and analysis, and organization and expansion of our data storage capabilities. Limited storage capacity over the past years was creating data management and organization problems. Recovery **bf** archival data required the intervention of our programmers to mount tapes and restore files to disk. Our recent acquisition of an additional 3 Gigabytes of disc storage as well as optical drives will solve our short and perhaps long-term storage problems. We plan to keep active data on the hard drives and use the optical drives for files that are used only infrequently. The active data include historical data sets which can be accessed without concem about ownership of the data. The ability to store the majority of our data on disks has stimulated the development of software tools to facilitate access to the data without programmer intervention. We now have available a collection of data manipulation tools that were derived from software developed for simulation modeling. These allow individuals to view the data in a graphical form and to create files containing subsets of the data to use for further analysis. The data files are accessed through servers running on the host UNIX computer. A typical client program is called LTERmenu. LTERmenu has been made available to investigators at other sites through the mechanism of anonymous ftp. LTERmenu uses a series of menus that allow investigators to select data sets to be manipulated. LTERmenu is a prototype for testing menus and the interprocess communication library we are developing. LTERmenu runs under the Sun

Microsystems Sunview shell. We plan to make available a version that runs under OpenLook (the X windowing environment), and will attempt to provide a version that will run under DOS on PCs.

Last year we distributed to the LTER data managers and other interested people a PC program and database that allowed them to retrieve entries from our local LTER bibliographic database. Entries could be located by author, keyword, or simply by browsing through the database. The entries can be saved in either ASCII or WordPerfect formats. We are currently extending this capability to our Sun computers, and plan to make this software freely available within the next year.

CWT

We have transferred all the Coweeta datasets from 9-in magnetic tapes to a Sun SPARCstation. During this process we have updated the databank, and compiled the Coweeta Hydrologic Laboratory Databank Catalog, which has been distributed to all the data managers at all the LTER sites. Now we are compiling a bibliographic database of all the publications on watershed management and ecological studies done at Coweeta. In addition, now that data catalog has been done, we have proposed to acquire a RDMS package (Ingres) to automate the data management process, and to facilitate data access to the P1's. We have also proposed to purchase a QC/QA tool (SAS Modules), and a optical disk for data storage. Our GIS system has been up and running for a while now, and it has been used extensively. Currently, it is being used to create a Potential Vegetation Model of Watershed 7. Our new Image Analysis System is also in operation, but it's running in "test mode". We have recently installed a ethernet bridge at the Institute of Ecology building which gave us direct connection to Internet. Since then, our LAN has been operating normally.

HBR

Site flash: 1) The Hubbard Brook sample archive is now a functioning system, with approximately 11,000 samples logged in. Each sample has a unique barcode which links it to a record in a database. Using a retrieval program, a person need simply scan the barcode to retrieve all of the information about the sample. 2) A directory of permanent plot studies now resides on the Hubbard Brook Bulletin Board. The directory is a compilation of responses to a questionnaire sent out by Charles Canham, Institute of Ecosystem Studies, in a quarterly newsletter called "Permanent Plotter". The goal of the directory is to keep an active and current listing of long-term vegetation research throughout the world.

HFR

Current research projects using the GIS include a study of the interaction of landuse history and vegetation 'in central New England, and studies of landsca~Ievel effects of hurricane disturbance in New England and in Puerto Rico (the latter as a collaborative project with the Luquillo LTER). Funding has boen obtained through Technology Supplement grants to establish a full Internet connection at the Forest, using a leased line and high speed modem connection to

· the Harvard University campus network. Progress has been made in documenting GIS files, entering long-term data on computer, compiling a data catalog, and establishing a data sharing policy.

JRN

The Jornada LTER experienced hardware and personnel transitions this past year. Our SUN 4/100 was returned to San Diego for remote sensing and GIS work at SDSU. We reestablished our email link through a host SUN machine at the NMSU computer center. We also acquired a new PC: a 486/33 through Gateway 2000. We are investigating RDBMS, in particular FOX-PRO, and tape back-up *(1/4"* streaming tape). Finally, we made it through both internal and external site reviews.

KBS

Lolita Krievs, the LTER GIS specialist at KBS, is now part of the data management staff. Thus far she has been involved mainly with data cataloging and documentation, and in exploring ways to integrate GIS data with other data. KBS, along with CPR, KNZ, NTL, and BNZ submitted a request for a technical supplement for the purpose of developing an inter-site climate database. As a follow-on to a prototype developed over the past **year,** KBS is joining with with Jim Beach, now of Harvard University, in developing a proposal for an inter-site networked herbarium database. Jim has accepted a **new** position with responsibility for computerization of specimen collections in three museums. Afthough this is not an LTER project per se, Jim is a good resource person for information on activities in the systematics, information science,and museum communities which are of relevance to LTER data management.

KNZ

1) New DM person (Haiping Su) hired 1 5the of May. His job will involve computer networking, data management and Remote Sensing/GIS. His background include a MS + PhD in agronomy from KSU.

2) A wildfire in April on the site has **"sparked"** some new research this summer. A SGER grant was funded to look at trace gas fluxes across watersheds.

3) A **tech.** supplement was funded this year. With these funds we hope to purchase optical disk for long-term storage to upgrade our GIS/Remote Sensing lab.

4) Our main **P1** (Tim **Seaste~t)** is leaving KSU for CU to head up Niwot LTER site. Alan Knapp will become the main **P1,** with John Briggs and David Hartnett. We plan to advertise for an ecosystems ecologist at the associate level.

NET (LTERnet)

The network office has installed the new Network Support System. This included the switch of LTERnet from the old VaxStation 2000, a relatively slow and small ULTRIX host, to a SPARCstation 2. Another SPARCstation 2, called Space, functions as a server for GISIRS related tasks. LTERnet and Space are integrated for file and device sharing using Internet protocols and SUN Network File System (NFS, incl. remote mounts and Yellow Pages). PCs connected to the system use PCNFS to have access to those resources. Connected devices include 4 Gigabytes of hard disk

storage, an Exabyte850o 8 mm cartridge tape drive (5 Gigabytes per cartridge) a 6250 BPI Cipher tape drive, a 150 Mb Sun Cartridge tape, drive, an erasable optical disk drive (600 MBytes per disk) and an optical jukebox for 10 optical disks. Other peripherals on the net are CDROM readers, a Calcomp digitizer, a HP plotter and color as well as BIW POSTSCRIPT printers.

At the time of the meeting, the planned connection to the commercial Sprintnet network (formerly Telenet) was being purchased from US Sprint. When completed in October, this connection will

enable LTER researchers without network access (e.g. while traveling or otherwise without network connection) to reach the LTER Network Support System by means of a phone call to the nearest Sprintnet access point (in most cases this will be a local or short-distance call).

Landsat and SPOT satellite data are now "pouring" into the Network office. Each scene is handled intact (Landsat-TM data are 291 MBytes/scene) for review and archive, with the original data tape sent to each site. Each of the scenes are archived on optical disk at the Network Office. An article describing the data acquisition, entitled "Remote Sensing and Modeling Activities for Long-Term Ecological Research" is currently in press for the October meeting of the American Society for Photogrammetry and Remote Sensing (ASPRS) in Atlanta.

NIN

Despite the enormous amount of effort expended to recover from Hurricane Hugo, significant

progress has been made in data management and support of meteorological data collection. The 9

LTER Core Data Set Catalog was completed as was the chapter "Data Administration" which was presented at the Kellogg workshop. A data availability site policy for North Inlet was completed and served as a model for implementation at other LTER sites. The SUN SPARCstation 2 has become the central component of the Data Management Network at the Baruch Marine Laboratory and the implementation of a file sharing LAN is complete. Acquisition of the Alphatronix Optical Drive has modernized the archival of all LTER da~a sets. The compatibility of the optical technology with identical equipment of the main campus insures redundancy in archival media and hardware. A portable uninterruptible power source protects the SUN and all associated peripherals from power outages, fluctuations and lightning strikes which are a common hazard. The new laboratory building, now in the design phase, will be equipped with dedicated UPS outlets in the computer complex and in all laboratories and offices. A digital fish length measuring board has been acquired from LimnoTerra, Inc. and implemented in order to automate the acquisition of fish length data. After Hurricane Hugo hit in September, 1989 several changes were necessitated in the collection of meteorological data. A temporary weather station was loaned by NCAR from October 14, 1989 through December 1990. Completion of the new Long Term meteorological station was constrained by the pace of reconstruction of Oyster Landing Dock. In December 1990 the new meteorological station was completed and retrieval, processing and quality assurance programs provided by Climatronics have been implemented for the archival of meteorological data. However, manual downloading of the data continues to be necessary, as the phone line will not be reestablished until the new laboratory has been built. The NWS weather station was replaced October 10, 1990 after the weather station enclosures were rebuilt. For the nutrient chemistry, two Technicon

Autoanalyzers were acquired to replace the equipment which was lost in the hurricane. Currently, all data are logged digitally to IBM PS/2 model 30 computers. The most serious constraint at this time is the lack of a full service network connection from the Marine Laboratory to the Columbia campus. There is no direct link from the Marine Lab LAN to the campus

ethernet to provide TCP/IP service on the Internet. EWSnet, a network of 25+ SUN workstations, has been established in the EWS building on the Columbia Campus, effectively integrating the Geology Department and the Marine Science Program in a UNIX (SunOS)

networked environment. It is our intent to replace the mainframe controller with a modem/router unit to provide serial TCP/IP network services via PPP (Point to Point Protocol) to campus. With such a configuration in place, the marine lab would be integrated with EWSnet and would be able to take full advantage of the Internet including remote UNIX protocols.

NTL

During the past year, we have converted our databases on the central campus VAX into INGRES. We plan to link front-end applications on end-users' microcomputers with the INGRES databases to provide researchers greatly enhanced access to databases using the Data Access Language in Macintosh System 7. We made considerable progress in organizing and maintaining the LTER wet samples including writing a protocol for maintaining the samples. Maintaining these types of collections represents a major achievement in an area of long-term research which is too often overlooked and undervalued. Our local file server was installed this year and is running Novell Netware. The file server is a 33 Mhz 386 machine with two 620 MB hard drives which use duplexing to protect against hard drive failure. Our data management system will be enhanced with several new additions from the 1991 technical supplement. Lab data reduction will be substantially improved with the installation of National Instruments Labview Software and DIA data acquisition board. The software will allow us to design **11virtual** instruments" to collect, analyze and report data from the **autoanaylzer,** ion-chromatograph, DIC/DOC system, and atomic absorption. We are also acquiring cartridge tape backup for the LANs at two of our research buildings and two erasable optical drives.

NWT

We have acquired a Sun SPARCstation 2 which is now connected to thin-wire ethernet. All of the LTER PCs in our primary research laboratory are ethernetted and are running CUTCP software that provides full ftp and telnet capabilities. Our 1991 technological supplement will be used to provide ethernet connections for our secondary research laboratory and for an upgrade of our climate data management program. A joint agreement among the University of Colorado

Graduate School, Institute of Arctic and Alpine Research, Environmental Population and Organismic Biology, and the Department of Geography has created 4 new tenure-track or tenured positions. All of the individuals hired to fill these positions will be directly involved with

LTER research. Tim Seastedt (formerly **P1** at **K'NZ)** was hired to fill the position of ecosystem science.

PAL

1)The new LTER Antarctic Palmer (PAL) funded jan91 to dec96 has several components including both

field work (water column, seabird and seabird prey measurements) and modelling. Our first year's highlights:

MAR: All Antarctic LTER Scientists' Workshop and MiniSeries (UCSB)

JUN: Steering Committee Meeting (UCSB) Antarctic Regional Logistics Meeting (Denver,CO

with Antarctic Support Services (ASA)

Palmer Communications and Computer Networks Meeting (UCSB) with ASA

OCT: First field season begins

NOV: First annual time series cruise

2) First field season begins 1 **1Octobe~1** and ends in March. During most of the season there will be 10-12 people on site. There will be an additional 6 people for cruise work.

3) A new staff research associate, Tim Neuberger, will join the Palmer LTER Antarctic team. He will provide continuity during the field season for the water column and prey components.

4) We're very happy to announce that a Polar Automatic Weather Station (Steams, Univ. of Wisconsin) has been funded. If all goes according to plan, the station will be installed this year at

Bonaparte Point approximately 1/4 mile from Palmer Station. Wind **speed,** wind direction, atm pressure and air temperature as well as several optical measurements will be made.

5) A request has been initiated with DPP for two GPS units for Palmer for finding stations during weekly sampling at sea from **zodiacs.**

SEV

Sevilleta LTER just completed its first review process hosting the sites internal review committee in May and then an NSF review in July. Both reviews went very well. The results and recommendations from the reviews will be taken into consideration this tall in an all **P1** retreat. Site hosted a workshop on Data Management for the Chinese delegation in May. Supplement proposal to enhance storage capacities for Data Management and GIS was funded. The proposal included DAT tape drives, magnetic disk drives, and optical disk drives, and analytical software (SAS) for the Sevilleta Information Management System. This field season represents the **"closing** of the gap" on a number of our data sets - turnaround time between data entry and summary analyses. We have been experimenting with computer conferencing software from the

UNM group that developed Khoros, hope to have a proposal out on this soon. Also, UNM biologists have received funding to build a real field station that can house LTER summer crews and visiting researchers.

VCR

There have been several advances in data management at the VCRILTER. The first and most important has been the addition of a half-time graduate student to aid in processing of incoming data sets and filling data requests. This position is currently being filled by a graduate student in Computer Sciences who will be working on a thesis dealing with data management issues, particularly quality assurance. Our Sun computer systems, although primarily used for GIS and remote sensing, have worked out well for data management by providing network and backup support. Addition of a Pinnacle read-write optical disk to the SUN systems has helped improve our perpetual disk space crunch. We will be getting INGRES **SQL~atabase** software for the SUN computers. This will be used for both conventional data management and for interfacing with the ARC/INFO GIS software using an RDBI.

APPENDIX B.

SOFTWARE DEMOS

ANDREWS DATA REQUEST HANDLING PROGRAM DEMO

We are establishing a more formal mechanism for handling outside requests for data. We have developed an automatic data handling and tracking system to make fulfilling data requests more efficient. A Data Request Form has been created so that requesters can provide information to us for tracking purposes (e.g., name address, institution, dataset description, purpose and publication intentions), and this information is maintained in a Requests Database. A cover letter accompanying the Request Form includes a statement indicating potential costs, and a standard acknowledgement statement for use in forthcoming publications or reports.

The Data Request Handling Program is outlined by showing available menu options as follows:

1. INITIAL REQUEST: Requester's name, **address,** e-mail, and requested dataset codes are entered and are automatically placed in the Requests database.

2. SEND FORM: Requester's name is selected from popup menu, and a

cover letter and Request Form are generated as an ASCII file for paper or email. (A special NSF acknowledgement is included with appropriate grant

numbers for core LTER datasets.)

3. FORM RECEIVED: Entry screen provided to enter form information (automatic parsing for e-mail files).

4. P1 PERMISSION: A memo is generated alerting the P1 of the data request and asking for **~ermission** to distribute the data. If permission is granted, date of permission is stored in database. Otherwise, a memo is generated to the requester to inform them that access is denied.

5. DATA DELIVERY: A README file is generated to the requester to accompany data delivery and to note the data documentation **filename,** DATA.DOC. The acknowledgement statement is reiterated. If there are data charges, a bill is automatically attached. The DATA.DOC file is created through user-friendly menus so that database documentation can be selectively included for each dataset. DATA.DOC includes a list of delivered files and their titles, a data abstract (optional), selected data format information and variable descriptions, and coded field definitions.

6. BROWSE: A Requests Database editing feature for viewing or editing all requests.

A status indicator is automatically updated each time a menu option is **completed,** and tracks the

current stage of processing for each request.

In summary, our program generates all necessary correspondence, as well as **documentation,** in the handling of requests for data, and provides a system for tracking the use of our datasets.

APPENDIX C. Gil Calabria

Coweeta Hydrologic Lab

University of Georgia

LTER DATA MANAGERS Institute of Ecology

MEETING: PARTICIPANTS Athens, GA 30602

San Antonio, Texas, August 1-3, Scott Chapal

1991 North Inlet Marsh

Baruch Marine Field Laboratory

P.O. Box 1630

Karen Baker Georgetown, SC 29442

Palmer Station, Antarctica **A.** El Haddi

Computer Systems Lab Cedar Creek

University of California

Santa Barbara, CA 93106 University of Minnesota

Dept. of Ecology and Behavioral Biology

Barbara Benson **3l8ChurchSt.SE**

North Temperate Lakes Minneapolis, MN 55455

University of Wisconsin-Madison John Gorentz

Center for Limnology kellogg Biological Station

680 N. Park Street Michigan State University

Madison, WI 53706 W.K. Kellogg Biological Station

Caroline Bledsoe Hickory **Corners,** Ml 49060

National Science Foundation

Biotic Systems and Resources Janet Greenlee

1800 "G" Street **NW,** Room 215 Jomada Experimental Range

Washington, DC 20550 New Mexico State University

Las Cruces, NM 88003

Emery Boose Don Henshaw

Harvard Forest

Harvard University H.J. Andrews Forest

Petersham, MA 01366 Oregon State University

Forestry Sciences Lab

Carl Bowser 3200 Jefferson Way

North Temperate Lakes Corvallis, OR 97331

University of Wisconsin-Madison

Dept. of Geology and Geophysics Rick Ingersoll

329 Weeks Hall Niwot Ridge

Madison, WI 53706 University of Colorado

INSTMR, Campus Box 450

John Briggs Boulder, CO 80309

Konza Prairie

Kansas State University Tom Kirchener

Division of Biology Central Plains Range

Ackert Hall Colorado State University

Manhattan, KS 66506 Natural Resources Ecology Laboratory

and Dept. of Range Science

James Brunt Fort Collins, CO 80523

Sevilleta National Wildlife Refuge

University of New Mexico

Dept. of Biology

Albuquerque, NM 87131

Mark Klingensmith Cindy Veen

Bonanza Creek Hubbard Brook Forest

Forest Soils Lab USDA Forest Service Service

University of Alaska Forestry Sciences Lab

Fairbanks, AK 99708 P.O. Box 640

Durham, NH 03824

Mark Klopsch

H.J. Andrews Forest Robert Waide

Oregon State University Luquillo Experimental Forest

Forestry Science Department Terrestrial Ecology Division

3200 Jefferson Way Center for Energy and Environmental

Corvallis, OR 97331 Research

GPOBox3682

William Michener San Juan, PR 00936

North Inlet Marsh

Baruch Institute

University of South Carolina

Columbia, SC 29208

Bernie Moller

Marine Biological Lab

Ecosystems Center

Woods Hole, MA 02543

Michelle Murillo

Sevilleta National Wildlife Refuge

University of New Mexico

Dept.of Biology

Castetter Hall

Albuquerque, NM 87131

Rudolf Nottrott

LTER Network Office

University of Washington

College of Forest Resources AR-i 0

Seattle, WA 98195

John Porter

Virginia Coast Reserve

University of Virginia

Dept. of Environmental Science

Clark Hall

Charlottesville, VA 22903

John Vande Castle

LTER Network Office

University of Washington

College of Forest Resources, AR-i 0

Seattle, WA 98195

APPENDIX D.

HISTORICAL ASPECTS OF DATA MANAGEMENT AT LTER SITES

by Carl Bowser

Department of Geology and Geophysics, University of Wisconsin

North Temperate Lakes LTER

The following talk was presented at the annual meeting of the LTER data managers in San Antonio, Texas (August 1991). I have chosen to submit it in a more expanded form and may have added a couple of comments that were not part of the original **talk.** It is a draft copy of what I hope will serve as an article on Research Data Management and, as such, it may not be copied or distributed without my knowledge and proper creditation.

Proceedings of the

46

HISTORICAL ASPECTS OF DATA MANAGEMENT AT LThR SITES

by

Carl Bowser

Department of Geology and Geophysics, University of Wisconsin

North Temperate Lakes - LThR

I- INTRODUCTORY COMMENTS:

This talk is a version of the talk presented to the representatives of the CERN (Chinese Ecological Research Network) on the occasion of their visit to the U.S. last May. The talk was presented at their last stop, at the Sevilleta site in Albuquerque, and reflects an attempt to indicate to the Chinese delegation the basis for our growth and evolution as a data management group.

As one of the Principal Investigators involved in the First group of funded LThR sites I've been involved in the LThR project from its outset. Having been involved early on in the LTER program I think I can offer some perspectives on where the program has come, and reflect on some of the "growing pains" of data management ifi the program. This talk is about some of the approaches to data management, problems we encountered, and solutions we've attempted over our years of growth, especially as they may affect some of the newer sites in the LThR network. And Fmally I've attempted to use these "historical developments" to suggest what I believe are some of the lessons we've learned and some thoughts about the issues that face us in the future.

II- OUTLINE OF ISSUES DISCUSSED

A- NATURE AND GROWTh OF LThR PROGRAM CONTROLLED APPROACH TO DATA MANAGEMENT

B- THE BALANCE BETWEEN MONITORING AND RESEARCH

C- INDWIDUAL SITE VS GROUP APPROACHES

D- THE OLD VS THE NEW (Established vs Novice sites)

E- TOP DOWN VS BOTTOM UP DESIGN

F- RELATION OF DATA MANAGEMENT TO RESEARCH

G- INDIVIDUAL RESEARCHERS VS RESEARCH GROUPS

H- AU"I'HORITY AND STABILITY OF DATA MANAGEMENT GROUP

I- LESSONS LEARNED, EVOLVING TRENDS, AND NEW PROBLEMS

ifi- EXPANDED COMMENTS:

A- NATURE AND GROWTH OF **LThR** PROGRAM CONTROLLED APPROACH TO DATA MANAGEMENT

The nature of the LThR program at its inception controlled our approach to data management. The availability of various hardware/software platforms was **limited,** and the microcomputer revolution was just beginning.

Site history is important as it reflected problems of growth. Growth has been in steps, with growing "pains" at each step as sites joined the LThR program and learned to share **ideas,** agree on common goals, and to incorporate new ideas into those of the established sites. Six LTER sites were funded in 1981. Even at this early level we were a collection of sites with differing levels of committment to data management and varying degrees of expertise)

AND = Andrews Experimental Forest

CWT = Coweeta Hydrologic Laboratory

KNZ = Konza **Pr~irie**

NIN = North Inlet MarshiEsturary

NTL = North Temperate Lakes

NWT = Niwot Ridge

A Second cohort of five sites included two years later, and the first meeting of these

11 sites was made difficult because of the differing ideas of some of the new

groups as to how we should approach intersite activities, and how we should use

the existing resources to sponsor scientific interchange at the intersite level.

CDR = Cedar Creek Natural History Area

CPR = Central Plains Experimental Range

JRN = Jornada

ILLINOIS RIVER

OKEFENOKEE

In 1987-88 the original group nearly doubled in size. Two sites were not renewed and eight sites were added, bringing to a total of 17.

ARC = Arctic Tundra

BNZ = Bonanza Creek Experimental Forest

HBR = Hubbard Brook Experimental Forest

HFR = Harvard Forest

KBS = Kellogg Biological Station

LUQ = Luquillo Experimental Forest

SEV = **Sevilleta**

VCR = Virginia Barrier Island-Estuarine

ILLINOIS RIVER (dropped)

OKEFENOKEE (dropped)

Most recently, through funding from the office of Polar Programs, one more site was added in 1990, bringing out total to the present 18.

PAL = Palmer Station (Antarctica)

Data management was identified in all early LTER proposals, but in varying degrees. Early indications were that IM'RASITE data management was of prime importance, but as sites grew in number and sophistication we began to realize a need to "converge" in our thinking about data management. NSF at that time was reminding us that we were peer reviewed **ar~d** funded to focus on the research on site specific goals.

B- THE BALANCE BETWEEN MONITORING AND RESEARCH WAS **~TALLY** A PROBLEM

What are the goals and objectives of the program? To monitor change alone? To detect change and study its controlling processes? What is the

proper balance between monitoring and process oriented research? These were questions we faced from the outset in LTER.

MO~~RtNG

Monitoring implies dedication to long-term measurement of chosen variables with careful attention to replication of measurements. Monitoring also had (and perhaps still does) a perjorative status, in that it implied more attention to collection of data than to the use of the data for research on long-term phenomena.

Simple Monitoring of ecological data is a relatively more straight forward question? "WHAT IS CHANGING AND BY HOW MUCH?"

This was the easy part. We "simply" needed to collect **data,** reduce the **data,** evaluate the data quality, move the data to a **database,** and to provide access to those needing the data (for whatever purpose; planning agencies, limited scope research **questions,** etc.).

RESEARCH ON LONG-TERM PHENOMENON

This was the harder part. What should we best monitor to understand change and long-term processes? What are the underlying causes to long-term change? Long-Term Research implies collection of data to gain NEW insights to ecological systems, and research on long-term processes. Collection of data of diverse types must be LINKED (or Linkable) to allow new relationships to be discovered and interpreted

We are not a group of sites funded for long periods of time, but instead we are a group learning about long-term processes. With the help of 5, 10, perhaps 30 years of data it is our task to help predict the **beha~vior** of large scale ecosytems on time scales of centuries to perhaps millenia. Clearly we cannot afford to simply collect data for these periods of time, we must learn to move across temporal (and spatial)

scales with the help of our shorter term observations and models build from our understanding of shorter-term processes.

Data collection strategies required awareness of other data structures and attempts to make diverse data types linkable. Therefore, data management needs to **fmd** convenient ways to optimize linkage of diverse data structures and to provide them to researchers in usable form, and in a timely fashion

Sites were driven by research questions, but still had to commit to some balance between new research and maintaining a stable monitoring program.

C- INDIVIDUAL SITE VS NETWORK APPROACHES TO DATA MANAGEMENT

The unique aspects of divergent sites required site specific solutions. Aquatic, grassland, forest systems needed different approaches to data management. Highfrequency data is different from annual organism population estimates, or plant inventories, etc. The background and experience of researchers at sites were (and still are) different allowing for several different approaches to data management.

The need to gam new insights about larger systems forced data management and data communication to an intersite level. At First data management activities at intersite level were designed to discover those areas of common approach and **Qust** as important) areas of divergent approaches/solutions. From the outset not all sites agreed that a larger network was needed, and that too much remained to be done at the local LTER site.

Our focus on site research and data management still led to some useful thoughts about common data management activities. Figure 1 shows the our site's concept (NTL) of how we were progressing in our data management tasks in 1982. The "wiring" was meant more as an example of our progress toward solving data management tasks, and also serves as a useful list of some of the elements of data management, or at least our view of it early in the LTER program.

Early efforts at site communication led to recognition for need for intersite communication. Data management representatives started meeting on regular basis to, compare approaches, coordinate activities, discover and define common needs/goals, and communicate approaches and solutions to data management. Data management activities have grown from the outset, and now have the reputation as one of the most visible and productive groups involved in intersite activities[1]. The following list illustrates the **"evolution"** in time of data management activities.

1. (1981; August) The First meeting leading to the proposal for the data management workshop held at the Kellogg Biological Station was at the ESA meeting in **Bloomington,** IN (Attenting Ron Bonell, Carl Bowser, John Gorentz, George Lauff, John Magnuson, Susan Stafford?)

2. (May 1982)KBS workshop on Data Management (NSF sponsored)

3. (Fall 1982) First meeting of LThR data manager's meeting (Univ. of Illinois)

4. (Nov. 1983)Research Data Management conference held at the Kimberly Center (Univ. of South Carolina)

5. (Nov.1984) Publication of Michener (ed.) book on "Research Data Management in the Ecological Sciences" (The First product of the data management group)

6. (Spring 1985)Data Management Workshop held at NMSU on Climate data. This was an early test of intersite data management activities using common data to all sites and addressing questions on:

a) Climate variablility at sites

b) Climatic differences among sites

c) Whether the LThR funded years were "average" compared to prior climate observation periods

Out of this group the Climate Committee originated, but the results of the

workshop were never published.

7 Later the Coordinating Committee agreed to fund an annual data management meeting.

8. Data management representation at Coordinating Committee meetings is now policy.

1The activities of the data management group are considered by NSF as elements of intersite research, and not as a management activity. Methods of data **linkage,** data communication, distributed data systems, etc. are all research elements that not only improve intersite communication, but also lead to improved data management/research activities at the sites.

9. NSF support for GIS and network level activities were initiated through the Technical Supplement grants.

10. Network level support for data management activities was initiated through establishment of the network office and the hiring of Rudolf Notrott and John Vandecastle.

Data management has come a long way since the beginning of the LTER project, and if the past is any measure it has much to look forward to!

D- THE OLD VS THE NEW (Established vs Novice sites) -

Sites had (and still do) varying degrees of sophistication in data management. Some of the existing sites had data management activities in place and integrated with larger research activities (Andrews site is a good example). Some data management groups were far ahead of other sites in their **thin~~~g** about data management. Some existing groups were functioning with older, obsolescent technologies at a time of rapidly changing technology [The question was (and still is) basically how to deal with data management in a time of growing/changing technology.]

The most "expert" sites in terms of data management experience were those who **had~been** in the business for some time. Their systems were largely mainframe based with heavy use of site and research project specific FORTRAN coded programs. Some were still using Hollerith cards. Some groups were pushing use of the CPM operating system as the "system of the future".

Those sites starting from scratch had the opportunity to take maximum advantage of the growing power of microcomputers, and many did. Some sites chose to learn solely from others, but at the risk of these sites being less influential in the future directions of data management in LThR. To some extent these sites "mined" the expertise of other sites, and did not provide full participation in our data managementactivities. Currently even those "novice" sites are caught up in the problems of dealing with changing technologies. We've had to learn how to keep data management systems up to date and functional ASSUMING that hardware and software systems will change.

For **example,** in the past 10 years the North Temperate Lakes site moved from Univac 1108 to VAX mainframes and from Apple **II** to a mixed Macintosh/IBMPS2 desktop systems (including Compaqs for RS/GIS activities). GIS activities allowed expanded view of data management, but initially GIS activities were (and somewhat still are) carried on as separate from the prior data management activities.

At present it's become clear that we must assume that things will constantly change (usually for the better), but at the cost of a certain percentage of data managers time and LThR funds toward keeping up in critical areas.

E- TOP DOWN VS BOTTOM UP DESIGN

A basic and fundamental problem of approach faced by LTER project from the outset and one still relevant to our discussions at these meetings is that of top down versus bottom up implementation of data management. We recognized diversity among sites, and that sites were initially funded to do research based on more site specific concerns and not at the intersite level.

Sites are different. They differ in their management structure, relation of the research site to the **P1** home base(s), the nature of the facilities at the research site, and year round site access limitations. The following lists a few of the differences in site structure that could lead to different solutions to the data management activities at each site.

1) Sites and University are "near" one another (< 100 miles) [i.e. one could drive to the site, do research, and return home in the same day.J

2) Site **"remote"** (> 100 miles) from the University requiring nearly a day just for travel

3) Site has minimal research facilities at the site, and therefore most sample processing and data entry had to be done back at the University.

4) Site has full time personnel resident at the site to provide sampling assistance.

5) Site is available, and used year round.

6) Site is only available in the Summer months

7) P.I.'s are mostly all at the same institution

8) P.I.'s are spread over several **institutions,** all at some distance from the "home" institiution

Because of the differences among sites, each site needed some freedom to tailor the data management system to their own needs. Sites recognized some need for standardization and convergence, but differed in ways to solve the problem. The problem of convergence was split into two opposing views:

1- All sites purchase compatable hardware and software or develop identical formats for documentation and catalogs

2- Recognize "diversity" in hardware/software solutions based on the needs justified by each site.

The control of data management by centrallzed decision versus convergence resulting from the experience of diverse groups can be debated. It has both advantages and disadvantages. Top down enforcement of data management structure is potentially more "efficient" in that it leads to uniformity of **systems,** and group familiarity with data management hardware and software **o~er** the larger network. This was a difficult problem for the LThR program because of the fact that sites had chosen their systems for data management independent of other sites.

Bottom up evolution of data management is less "efficient" and arguably slower. The fact is that the LTER sites chose this mode, partly in recognition of the diversity of sites and their data management **approaches,** but also based on a negative history of a prior research program. (IBP). It took advantage of a wider array of data management solutions avallable with different computer systems. It also encouraged originality in approach sometimes leading to new, and very useful, data management approaches

We needed to **fmd** ways that encourage experimentation with new or alternate approaches to data management, and to take the best of these ideas and attempt to make them work across a larger number of sites. We found that it was possible to coordinate our efforts at an intersite level while still allowing for differences among sites.

Connectivity is improving, and we should have less problems in the future dealing with the changing system environment, but it IS something about which we have to be aware. Questions about the ease of connectivity revolve around the common format for file interchange and convergence toward common operating systems. It is true that ASCII files are the common media (currency) of exchange

of data and text. More recently file compression programs, and conversion of binary files to pseudoascil (binhex) format for transmission have been written to **-~allow** exchange of programs across the network in ASCII format, but now ftp and related protocols (Telnet,...) readily allow transfer of files directly in binary format.

UNIX supported software now allows remote logon and use of programs, a vision we only dreamed of just a few years ago. The lternet office Bulletin Board is hardware independent in that users need only have terminal access to the internet. Working **in~a** heterogeneous environment is becoming a relatively straightforward process, and improving daily. We should not discourage that trend.

F- RELATION OF DATA MANAGEMENT TO RESEARCH

Researchers and data managers can have quite different views of their role in the LThR project, and it has been very important to recognize that those views can differ **markedly.**

The Database as a "library"

One type of librarian wants to catalog and preserve books, and is not particularly happy when too many books are checked out. Other librarians (appropriately) view their job as a resource manager, and are happiest when their facility is used a great deal.

In the framework of the LThR project the data manager could be considered the librarian and the researcher the user of the library. The library user (researcher) **--wants** access to the data, and in a form that's most useful to the questions they are trying to answer. Past problems with data management at specific sites reflect these two views of the process.

Successfully cataloging and archiving the data is not enough for data management, but was thought by some to be sufficient. I Recall a data management meeting several years ago in Corvallis where one of the data management "statements" being recorded in the minutes is that "...data management is not a problem at the LTER sites." Had it been no problem we wouldn't have been meeting, and several site RESEARCHERS were complaining bitterly that they could not get easy access to the data in forms useful to them. We must constantly be reminded that Data Management is not a process to archive data and provide data set catalogs and documentation. It is a PROCESS that serves an end user, i.e. the

RESEARCHER.

"Research data management" serves the research function. It's a term that was used as a title for the Baruch Symposium volume to emphasize that management of research data is perhaps different than traditional data management. Effective data management involves coordinated work with data managers and researchers from the outset. To the extent that data managers were removed from the research processes they tended to be less appreciated, and treated as "librarians" rather than active participants in the research process. With such a structure we risk loss of support of data managment activities with long term consequences to all of the **LThR** program.

As the data managers are able to demonstrate to their own **P1's** and the LTER

- - program that they can provide NEW insights to research by providing new "mappings" of data on one another, or by generating new graphical relations for looking at data, then they will develop stronger support from the site project leadership.

We've struggled to have a mix of researchers, data analysis, and data management people involved in our data management activities to help inform one another of the problems and potential solutions about use of data in a centrally managed database system. **It's** worth preserving.

Solutions? Better integration of data management into research (liason with science), active participation of researchers in the data management, representation of a data management representative on the coordinating council, attendance of researchers at annual data managers meeting.

G- INDIVIDUAL RESEARCHERS VS RESEARCH GROUPS

By its nature the LThR program is a collaborative enterprise. However most research personnel came into the program having a background of working more independently of one another, or with collaboration among a relatively small group of self selected researchers. We have a tradition in the natural sciences to work more individually. Why? One can only speculate.

1- The peer review process differentially rewards "first author" papers

2- The stereotypic "ecologist" is one who prefers to work alone in natural settings, and is not driven by "big science" projects.

3- Faculty hiring tends to promote diversity in specialties providing little overlap and promoting individual research projects.

4- The research problems used to be more tractable for single investigators.

5- The competetive urge and investigator egos can stifle collaboration.

6- Single funded researchers by tradition did not have collaborative programs "forced" on them, but instead forged individual collaboration on a person by person basis.

From the point of view of the LThR enterprise lack of stimulation for collaborative research can be viewed as a problem. Meant less as a criticism as much as it is a statement of fact, it's one that we've had to face. We've made progress at dealing with the situation by:

1) Meetings to get people familiar with one another (All Scientists meeting, data managers meetings)

2) Watching some groups achieve new levels of **understanding'** based on a model of cooperative research.

3) Seeing a growing literature of multi-authored papers.

4) Reallzation that some problems are ONLY possible with an interdisciplinary or multidisciplinary approach.

INTERDICSIPLINARY VERSUS MULTIDISCIPLINARY APPROACHES TO RESEARCH

Interdisciplinary research and multidisciplinary research are fundamentally different approaches, and deserve some discussion. Recognition of such is the key to understanding interaction among researchers and data managers in the LThR setting. Multidisciplinary and interdisiplinary approaches to research differ mainly in the way intestigators allocate their research funds, and interact in the research setting. Multidisciplinary research essentially involves more individual investigator independence, both in funding support and in activity. Interdisciplinary research

focusses more on shared funding of common resources (lab facilities, technical support, data management support, etc.). Other differences are:

Multidisciplinarv research

1. Groups of researchers working on the same problem

2. Researchers work more or less separately from one another

3. Generally have their own budget "piece of the pie"

4. Operate to provide specialty expertise to the group, but tend to share data at less frequent periods

5. Tend to keep their own data in forms most useful to them.

6. Researchers are generally happiest working with only their own data and in their own specialty area.

7- Researchers tend to be in a competitive situation with one another, usually trying to convince one another of their "value" to the project by highlighting their own research contribution.

Interdisciplinary research

1. Groups of researchers working on the same problem

2. Work more closely with one another on a day to day basis.

3. Generally do not have their own budget, but share specialists among themselves

4. Operate to provide specialty expertise to the group, and share data on an open basis with one another.

4. Provide data to commonly managed system, and in forms determined by group needs.

5. Operate to provide specialty expertise to the group, but in a more interactive mode.

6. Researchers work with pooled data as well as with data from their own specialty research.

7- Researchers tend to be in a less competitive situation with one another, focussing more on problem solving than establishing "group worth".

In table 1 I've tried to indicate the differences between multi- and interdisciplinary approaches, highlighting the differences in the function of data management between the two approaches.

Depending on the specific character of researchers and their ability to get together and meet on a regular basis, one could argue the merits to both approaches. Most researchers appreciate credit for their contribution to the project, and it is a challenge to be more communal in the research enterprise, and still fmd reward for individual contributions. Nonetheless the nature of research involving a larger number of research personnel will eventually demand some acceptance of the interdisciplinary mode of interaction.

Much of the research in the LTER program involves synthesis at higher levels than represented by individual specialties, and therefore, linkages of specialties is important. This implies a need to link data sets or to map data into common relations that will allow synthesis, and that implies some committment to a data management structure that promotes data linkage, and "on line" sharing of data. Researchers who work more independently may not respond well to these group needs, thus, making linked data sets more difficult to achieve.

Linkage is the key word. To link diverse data requires cooperation of resarcher and data manager. The data manager assumes more central control of data, and their help is often needed to retrieve the data for individual researchers. This was

> (and still is) a major barrier to some researchers; they feel that they've lost direct control over data that they've collected. Similarly this problem lies at the root of data security, data sharing in networked systems, and open availability of data required by federal agency policies.
>
> Our site (NTL) has struggled to find ways to share and link data of diverse character. Data differ in type and frequency of measurement. Our data manager plays a central role in helping to link such data sets, and serve to provide data retrievals that cannot be done by any but the most "computer literate" researcher.
>
> Data types
>
> a) Continuous (temp, stream flow, Nitrate conc., etc.)
>
> b) Discontinuous (presence/absence data, numbers of organisms, etc.)
>
> c) Ranked Variables
>
> d) Attributes **Datafrequency**
>
> a) Hourly weather data
>
> b) Monthly data
>
> c) Annual Population estimates
>
> d) Infrequent observations
>
> e) Short term graduate studies not considered as "Core" data for the LTER project.

Dispersed data vs centralized **data**

> Dispersed data under the multidisciplinary model (above) requires data management focus on data catalogs and data documentation (knowing where to request the data). Centrally coordinated data also requires data documentation and catalogs, but more importantly focuses on ways to link the data or to provide data in forms that help groups of investigators working together (modeling, graphic displays, statistical analysis, etc.).
>
> What is true at the SITE level (individual researcher versus site interests in data) is also true at the INTERSITE level (site interests different from network interests in the data) - [i.e. groups of groups]

H- AUTHORITY AND STABILITY OF DATA MANAGEMENT GROUP

To be useful to the **LThR** research enterprise data management activities must be well funded, adequately staffed, and given sufficient authority to carry out its function. The level of funding of data management is a critical issue, and was one of the major differences among sites, a difference that still **exists,** but diminishingly so.

Data management as a cost to the site

Among some research personnel data management activities were considered as taking funds away from research projects. Perhaps short-sighted, but when research is involved there is always competition for funds, and data management activities did not always compete well for these funds. There was a risk that the data manager could become isolated from the research process, and as that happened there was a tendency for data management activities to suffer diminishing support. Clearly the closer the data management group is to the research group the

better it is for both. They are NOT independent activities, but had become so at some sites. Where this attitude existed it has been a challenge to the data management group to demonstrate the role that RDM plays in active research.

Parenthetically the same holds true for data management activities at the intersite level. All data managers have a stake in the success of the LTER project, and as such need to be concerned that their activities, no matter how seemingly important, must not diverge from the interests of the research **P1's** in the program.

Appointment level

The level of data management appointment and data manager authority has been diverse among sites. We've literally seen the complete range of appointment level for data managers from principal investigator, data management specialist, graduate student, secretary, data analyst, hourly help, computer data administrator, ecologist with an interest in computers, and computer enthusiast with an interest in the "environment". The lower the level of appointment or the more removed the data manager was in familiarity with ecological research the weaker the data management position became.

To have a research investigator assume the role of data manager took away from research time and was probably not be the best use of their time. Using hourly help to "data manage" is a misunderstanding of the concept of data management. That person merely served to transfer data to computers with little or no help in organizing an effective data management system. One need not go through a detailed description of the differences each site chose for the level of data management position to realize that strong data management groups have developed at sites where there was strong principal investigator interest in the data management, and where the level of appointment was appropriate.

We do not need principal investigators as data managers, but we DO need principal investigator level appreciation of the data management tasks, and some backup in the ehforcement of policies among all site researchers.

Stability of data management position

Short term hires have been problems for some sites, and may reflect their level of committment to the coordinated data management activities characterized by the LThR project. Some sites had a new data manager show up at each intersite data management meeting **prQviding** little "memory" for their site, and consequently they provided less help in developing our present approach to data management. Generally the most successful data management groups have come from sites where there has been relatively little turnover in staff.

Authority of data manager

Temporary or short-term employees tended to be people with little or no authority to forge data management agreements among researchers at the sites. These people became the "librarians" of the data management world. Committment to full-time or fully-funded data managers was needed to develop long-term data management systems that served needs of the local site as well as the network of sites. They needed higher levels of authority to put some **"teeth"** into data management requests. They also needed backup from site **P1's** to ensure that

agreed upon protocols for reporting and documentation of data were accepted, and enforced.

At the Northern Lakes site it has worked well to have one P1 responsible for working with the data managers, and to have the data manager directly accountable to the lead P1.

I- LESSONS LEARNED, EVOLVING TRENDS, AND NEW PROBLEMS

Are there lessons from ten years of growth of LTER data management? I certainly hope so. Where do we stand, and where do I think data management is heading? The following are a few thoughts.

Top Down vs Bottom Up design of data management are both viable issues in **LThR**

Bottom Up structure reflects individual site needs

The top down versus bottom up issue is likely to remain with us for some time. **Let's** face it, we would all have an easier time if all of us were to be using the same computer and the same **software.** The problem is to agree on what those best "systems" will be. Each has their own ideas based on their prior experience with computing environments. All too often one **person's** idea of the best computer is based on a lot of experience with their own computing environment, and precious little knowledge of other possible systems. Competitive opinions about what is the best system will always be around, but they should at least be based on more than familiarity with just one system.

My personal definition of the best computer is that it's the one sitting on MY desk. We benefit directly with what we have learned to use effectively. Switching to other computer systems or learning new software is not cheap. People must invest a considerable amount of time to move to another computing platform, and we should all think of the costs to the other person when we propose some convergence in computing systems and software.

The UNIX vs VAX vs Macintosh vs IBM vs **"whatevef'** discussions are not very productive **in.my** opinion. Certainly we should be aware of new products and approaches, but that doesn't justify some of those interminable discussions as to why our own system is better than someone else's.

I would argue that the fuel for creativity and originality is to have diverse approach to data management (in **concept,** and in implementation, i.e. hardware and software). I believe we can look back on the successes of data management in the LThR program and appreciate that fact. In the 1960's when IBM was able to capture most of the computer market the only "creativity" we the user saw was price escalation. To some extent that era of computing also fostered the "you come to me and learn how to use MY system" attitude that still pervades some university computing facilities. We are better off for having diversity.

The need to conform to common formats for shared data should not be confused with percieved needs for common hardware/software implentations. We have to remember that agreements for common hardware/software will impose a

real cost on sites that they may not be able to afford at the time. It is appropriate for our activities to capitalize on diversity and do our level best to implement a data management system that assumes heterogeneity from site to site.

Had we all conformed to the same system at the initiation of the **LThR** project we would still not be using the same platforms we would have chosen at the outset. In the early 80's mainframes were the platform of choice (each with their own unique operating system). Spreadsheets, were unknown on mainframes. Recall that Visicalc, the first spreadsheet, sold as many Apple **II's** as the other way around. The computer revolution was just beginning to have an

effect on the user community in the early 1980's. Hollerith card input, keypunch machines, a much older style of Fortran, and long waits for batch processed jobs were essentially the only way to operate prior to the 1980's.

Top Down structure reflects needs of LThR at the intersite level

There is a need for some convergence in approach (top down), but that convergence is needed for common approaches to data cataloging and documentation, standards for data formats that facilitate data sharing, and procedures that help research at the intersite level. The history of the LThR data management may have been one of convergence stimulated by bottom up structure, but there ~:s also some need for a level of common agreement, implying top down structure.

We have grown by listening to one another's approach. We've converged on what we feel is best, and rejected that which does not work at our site. For that reason some of the relatively new sites have the potential to take advantage of the experience of research data managment within the LTER program, and some of us "old site types" will have to be prepared to grow with some of the newer trends in database systems.

Among those areas of common concern that will lead us to sorne convergence are:

1) Data Catalogs and Documentation

Uniformity of data descriptions from sites and the growing need to pool some data will demand some convergence. Remote query of sites to determine what data sets are available are not far away, requiring more in software solutions than in hardware to implement such.

2) Exchanging and pooling data from other sites

The ability to exchange and integrate with other LThR sites and perhaps other ecological networks is becoming necessary. Some of the first efforts were initiated in the data management workshop held at NMSU in 1985 on the subject of climate comparison across sites. The newly funded data management effort to put climate data "on line" for all sites is another step ir' this direction that will be a prototype for other shared data sets.

3) Future distributed databases systems

Possible future of "distributed data systems" where all data structures will have to be in some form that will allow convenient linkages. While of some advantage to all sites and outside interested groups it will also bring to the fore the problems of data sharing mentioned earlier.

Learning to work in heterogeneous environments is essential to future growth

Software and hardware capabilities have changed considerably since the initiation of the LTER project and there is no reason to think it will not continue to change in the future. Sites may differ in their specific hardware/software implementations, but have comparable cababilities. We've learned to exchange data across systems, and with substantially different software. Much commercial software has provided solution to some of our data management problems, and more appears every day.

I have a personal preference for commercial software solutions, where available. Customized programmmg may be needed to solve specific problems, but suffer in that the programmer does not always provide for upgrading and evolution of programs. We've many examples of customized programs that have become obsolete, the programmers have left, and nobody has the time or patience to work their way through these programs.

We now have the ability to read tapes, diskettes, and CD-ROM's with different operating systems, to the ability to move data on the network using ASCII format we have relatively little problem in moving data files across systems in existing heterogeneous environments. Networks of computers are common, and several software implementations have arisen in recent years to handle moving of data, mail, compiled programs, satellite and other images, and sound. Compression routines allow data transmission rates fully acceptable to most users, and "on line" use of remote

computers is commonplace. All of these capabilities exist with little need to have the same computer system, only software capable of exchanging data with other computer systems.

Macintosh, SUN, and IBM appear to have the market for desktop computers, and UNIX and VAX operating systems seem to be dominating the mainframemimcomputer market. Desktop computers bring to most of us more computing power than existed on mainframes just a few short years ago. NEXT computers, parallel systems, and distributed computing systems are possible with existing technology; what can we look forward to in the next ten years?

Data managers in the LTER program should assume that hardware and software will change in the future, and should plan their data management approach accordingly. It has been estimated that computing power increases about ten times every five years, and we need resources to take advantage of these more powerful systems. It's a FACT that we're simply going to have to learn to deal with.

It is a problem to maintaln our existing systems, and also to provide for changes that seem to occur regularly. Accordingly we need to set aside some portion of our data management activities and financial resources to the continual process of upgrading of hardware and software at each of our sites. The LTER "Technical Supplement" has helped our sites tremendously, but how much can we rely on it in the future? We can't afford to find ourselves falling behind in this field simply because the technology has passed us by, and we have to find ways to prevent the need for "massive infusions" of money to bring ourselves up to date.

We need the ear of all **P1's** at the sites to convince them that the funding of data management mcreases their ability to do research and that it doesn't take resources away from short goal oriented research priorities. Loss of support for data management activities at our sites may provide short-term funding for research, but it will prove disadvantageous to the sites in the long-term. We need to be able to

convince all researchers of that fact. (as well as those in appropriate positions within the funding agencies). Improved support from the site **P1's** also requires that the data managers need to have full involvement in the facilitation of research not just management of data.

Sites need to agree to some sort of minimum standards of data management

The need for some convergence in our data management programs, and the leadership role that the LThR has exhibited will require that all sites meet some minimum standards of data mangement capability. To stress the work "capability" means that it is the concept of data management that should meet minimum standards of ability, and not the implementation of that capability. in other words we need to stress the goals and concepts involved in data management, and not anyone's idea of how those goals are implemented at each site.

The ability to move data across sites, to have adequate documentation and data set catalogs, the ability to maintain.appropriate records of data management activities, and to for all sites to have mimimum commitunent to research data management is called for.

Integration of our GIS activities with our field monitoring activities.is not near complete

This is a goal of the data management group, but I fear not yet a reality. The "technical supplement" was largely used to bring sites to some level of sophistication with remote sensing and GIS activities, and to some extent it overshadowed some of the ongoing data management activities. The demands for training and expertise in these added activities is tremendous, and in most cases required the hiring of additional specialists with remote sensing and GIS experience. To some extent that has led to distinct groups at the LTER sites working somewhat independently of one another. The power of graphical analysis and image manipulation of large remote sensed images is evident, but that capability will only be reallzed when the RS/GIS efforts are truly integrated into the base monitoring data, and with reserach uses in mind. That potential has yet to be realized at most sites.

"Outreach" to **othe~** organizations will force added demands and careful distinction between our capabilities and the implementation of those **capabilities.**

Collaboration with **o~er** ecological organizations is beginning. Caroline Bledsoe has had **oc~casion** to talk about networks of networks in the ecological community, and has invited a representative of the Soil Conservation Service to our meeting this year. The activities with our CERN colleagues, are another example. This latter relationship already has significant involvement of data managers at several sites, and should put demands on the data management group that will help underscore the growing outreach problems we face.

A goal of the data management group is to facilitate **collaboration,** data sharing, and research at scales ranging from local site to intersite to international. It is to our credit that we're gaining a reputation for leadership in Research Data Management in the ecological sciences, and it's a position that we **can't** take too lightly. It isn't going to be easy to maintaln that leadership position, but will require a willingness to stay involved in data management activities OUTSIDE our

own sites, and an interest in working with a "diversity" of data management solutions appropriate to different organizations.

IV- CLOSING COMMENTS

When I look back on where data management has come since the early days of the LThR project I can always remember problems and difficulties, but most of those percieved problems have been swept away by our frequent meetings, collaborative orientation, cooperative spirit, and (I think) our need to obtain a satisfactory blend of "bottom up" cooperation and "top down" driven needs for data sharing and more global level interests in Long-Term Ecological Research. We've a lot to be proud of. We can look forward to an even greater period of accomplishment Those accomplishments will come easier if we constantly remind ourselves of how we got where we are today and what important guiding principles emerge from the history of our development.

Clearly we research data management is fundamental part of the process of research in **ecology,** and not a separate activity driven by a need to satisfy funding agency requirements. To the extent that our sites appreciate that, and know from seeing the successes of data management in forging research we will continue to gain respect for our contributing role. To the extent that we do not, we risk becoming the "data librarians" with little or no appreciation of the excietment of research and the knowledge that **you've** all had an important hand in the process of research.

7

PRIORITIES & PRO6RESS

- Data Collection

· Data Logging

· Data Verification

· Transfer to Main~me

· D.B. Schema Definiton

· D.B. Loading

o DataArchiving

o Data CaijAbstracting

● ● Report"' Generation

· Data Downloading

o Data Examination

o Graphi~I'Tabular

o Statistical MalysIs

o Data Modelling

o Interpretation

o FeedbacktoData _____ Collection

●

o **Twi~d Tog~her** Open Cifcuit

F-,

MULTI INVESTIGATOR RESEARCH

TYPE OF RESEARCH COOPERATION

MULTI- INTER-

_____ DISCIPLINARY DISCIPLINARY

1 Research Group Working on the same Working on the same

_____ problem problem

2 Individual More Independent More Collaborative

Researcher _____ _____

3 Researcher Specialty Expertise Specialty Expertise

Provides: _____ _____

4 Distribution of Portion of budget is More of budget given

Research Funds assigned to each to common staff

researcher (technical support, etc.)

5 Researcher's Generally kept Pooled into centralized

Data independently by each data management

_____ researcher system

6 Access to data _____ _____

a For Researcher Easiest for researcher Depends on level of

computer skills;

generally more

_____ difficult

b For Site Group Difficult; requires Easier if managed

researcher response properly; requires data

manager response

c For Intersite Very difficult to Potentially easy, but as

Groups impossible yet not implemented

7 Data Manager Data Set Catalogs, Data Linkable "relations

Provides and Dictionaries from diverse data sets,

Needs Database schemas

Data set catalogs

_____ Data dictionaries

8 Data Manager "Librarian" Research Facilitator

Function _____ _____

Table 1.

LTER NETWORK OFFICE PUBLICATIONS

No.1: Long-Term Ecological Research. *Bio&ience,* 1984

No.2: The Climates of the Long-Term Ecological Research sites. University of Colorado, Institute of Arctic and Alpine Research (INSTAAR), Occasional Paper 44, 1987

No.3: Standardized Meterological Measurements for Long-Term Ecological Research Sites. Bulletin of the Ecological Society of America, 1987

No.4: 1990s Global Change Action Plan. Network Office, 1990

No.5: Long-Term Ecological Research Network Core Data Set Catalog. LThR Network Office and Belle W. Baruch Institute, 1990

No.6: Climate Variability and Ecosystem Respesse. Network Office and UDSA Forest Service SE Experiment Station, 1990

No.7: Internet Connectivity in the Long-Term Ecological Research Network. LThR Network Office, 1990

No.8: Contributions of the Long-Term Ecological Research Network. *Bio&ience,* July/Augs'st 1990 (single article).

No. 9: Long-Term Ecological Research and the Invisible Present, *BioScience,* July/August 1990; Long-Term Ecological Research and the Invisible Place. *Bioscience*$_1$ July/August 1990 (three articles, including LTER Publication No.8)

No.10: Prcceedings of the 1990 LThR Data Management Workshop, Snowbird, Utah; Network Office, 1990

No.11: Long-Term Ecological Research in the United States: A Network of Research Sites 1991 (6th edition, revised), LThR Network Office, 1991

No.12: Technology Development in the Long-Term Ecological Research Network: Status of Geographic

Information Systems, Remote Sensing, Internet Connectivity, Archival Storage & Global Positioning Systems. LThR Network Office, 1991

No.13: Proceedings of the 1991 LThR Data Management Workshop, San Antonio, Texas; Network Office, 1991

Planned Publications:

LThR Atmospheric Chemistry Workshop Prcceedings

LTER Stream Research Catalog

Other Pubilcations:

***LTER Network News* (biannual newsletter; back issues available)**

In most cases, the above publications are available in ljiiiited cpwntiaes at no cost. For inform~on on their current availability and/or status, please contact:

Stephanie Martin

Publications, LThR Network Office

University of Washington

College of Forest Resources, AR-b

Seattle, Washington 98195

PH: 206~5434764; FAX: 2O643~O790/3O91

Internet: ***sMartin@tternet.washingto".ed~~/Bitnet: sMartin@fternet***