DATA MANAGEMENT AT BIOLOGICAL FIELD STATIONS AND COASTAL MARINE LABORATORIES

January 1992

Report of an Invitational Workshop
April 22-26, 1990
W.K. Kellogg Biological Station
Michigan State University

sponsored by
Organization of Biological Field Stations
and
Southern Association of Marine Laboratories

prepared for
National Science Foundation
Division of Biotic Systems and Resources
Biological Research Resources Program

edited by
John B. Gorentz
W.K. Kellogg Biological Station

TABLE OF CONTENTS

PREFACE

The data on the natural populations and biological processes of a biological field station's habitats are a research resource, just as are the buildings, research equipment, and habitats themselves. Or more accurately, they are a potential resource, a potential that is realized only when the data are organized, documented, and cared for to make them usable and accessible.

Although this issue of data management has been given increasing attention the past several years, and much progress has been made, it may be that the task of developing these data resources lies largely ahead of us.

A workshop was held at the Kellogg Biological Station in 1982 to encourage and foster the development of data management at field stations. Since nearly a decade has passed, it seemed an appropriate time to assess the progress that has been made, to reexamine our goals, and to determine what can be done to encourage and lead the way to the further development of databases and their utilization.

We sought support from the National Science Foundation for a data management workshop at which representatives from field stations and coastal marine stations could examine the state of data management, share information, and propose goals and new projects to advance this important work. As terrestrial and coastal marine stations wrestle with ways to allocate their limited research resources to this need, they can and should learn from each other's successes and mistakes. Field stations as a group have some unique objectives and requirements, giving them a common interest in data management that is somewhat distinct from other data management activities and goals. We believe the essence of the workshop deliberations held at the W.K. Kellogg Biological Station during April 22-26,1990, has been effectively captured and documented in the report that follows.

The workshop was supported by a grant from the Biological Research Resources Program, National Science Foundation, was co-sponsored by the Organization of Biological Field Stations (OBFS) and the Southern Association of Marine Laboratories (SAML) and hosted by the Kellogg Biological Station.

Thirty-six participants were invited to the workshop, representing data managers, scientists, and administrators representing biological field stations and marine laboratories of the United States. They represented sites newly embarked on data management programs, as well as those with well-established data management facilities.

The workshop was organized into three working groups, each led by two rapporteurs. These rapporteurs compiled the findings of their respective groups, and authored the first three chapters of this report. It should be recognized, however, that each chapter contains material originally contributed by those in the other groups; the topics are interrelated and it was impossible for any one group to consider its agenda in isolation from the others. On the day prior to the workshop, to provide some background for the participants, a pre-session symposium and the results of the a pre-workshop survey were presented. These materials are summarized in the fourth chapter. Because so much of the discussion at the 1990 workshop was made in reference to the 1982 workshop, it was decided to reproduce the 1982 report in the final appendix to this report.

Co-Principal Investigators
James J. Alberts, SAML
John B. Gorentz, KBS
George H. Lauff, OBFS

RAPPORTEURS and AUTHORS

Data Administration:

William K. Michener

Ken Haddad

Data Standards:

Warren Brigham

James W. Brunt

Computer Systems for Data Management

John H. Porter

Jeff Kennedy

Summary of Survey and Pre-Session

John B. Gorentz

Michael P. Hamilton

Editorial Consultant

Edie Erwin

INTRODUCTION

Science is based on the free and open exchange of information, whereby scientists can build on the work and data of those who have gone before them. Since science builds on previous work, including that represented in previous databases, scientists have a responsibility to preserve data for those who will follow after them.

In this context, the data gathered at biological field stations and marine laboratories constitute a national resource which should be preserved and made accessible for the purpose of advancing science. Long-term records of populations and biological processes in natural habitats as well as the physical and chemical environment in which they occur, are a research resource necessary to the study of ecological processes of regional and global significance.

Data sharing through the traditional system of refereed publication is not always adequate; there are unpublished data, never-to-be-published data, and raw data behind publications that need to be made available as a resource for others. Although some disagreement exists over whether available resources should be spent testing hypotheses rather than on preserving data without a clear hypothesis to be tested, it is generally agreed that the main purpose of long-term data management is to provide descriptive background data which can serve as a context for experimental studies. Research should always drive data management, rather than vice versa.

For the purpose of this publication, data management means caring for certain data so that, whatever their original purpose, they are preserved and made available for more general use, now or in the future. A field station's data management is distinct from computer management or investigator-specific data management, although it encompasses both. Comprehensive data management goals, realistic long-term planning, and solid institutional commitment are necessary for the care of data at each station. However, field stations and marine laboratories cannot manage data in isolation from each other. They need not only to collaborate and cooperate in data exchange, but also learn from each other's experiences, successes, and mistakes in developing their data management systems.

This publication is the result of deliberations by 40 representatives from stations and laboratories of all sizes. Their object was to produce a usable decision-making tool for data management planning and implementation. It is hoped that their shared wisdom will benefit the researchers, administrators and data managers at all field stations and marine laboratories.

The first three chapters of this report represent the conclusions of each of the three working groups—Data Administration, Data Standards, and Computer Systems. The first chapter summarizes the results of a pre-workshop survey and a series of demonstrations presented by participants at a pre-workshop symposium.

The goals of the three working groups were:

DATA ADMINISTRATION

- Identify the benefits of an institutional data management program for those sites deciding whether or not to embark on one.
- Identify the types of data management that can be of use: the data management services that can be provided, the types of data to be managed, and the types of resources and staffing.
- Distinguish between that data management which is appropriately undertaken by a site and that which is best left in the hands of individual researchers.
- Identify administrative structures by which data management programs can be successful, identifying appropriate relationships between data management, research, and site administration.
- Identify realistic funding levels and methods of funding data management.
- Identify growth trajectories appropriate to field stations of both large and small size and levels of activity.
- Identify means of long-term care and storage of data.
- Consider the **role of Geographic Information** Systems in **relationship to more** traditional data management.

DATA STANDARDS

- Identify areas in which standards are needed to make data management for collaborative research more efficient, and areas in which they are best avoided because they may hinder research more than help.
- Identify the potential benefits of standards in data management.
- Identify existing protocols that might be adopted.

- Identify mechanisms by which researchers and data managers can communicate with each other to develop such standards as are needed,

COMPUTER SYSTEMS

- Provide guidelines for choosing system capabilities that can aid data management,

- Identify computer systems appropriate to both large-scale and small field stations and marine laboratories.

"Discuss the impact of new technologies on data management, including not only computers and software but also local-area and wide-area networks.

"Identify costs of networking, both initial and recurring, to assist preparing budgets.

EXECUTIVE SUMMARY

The following is a summary of the major findings and recommendations appearing in Chapters 1-3 of this report.

Database Administration

- A data management program can benefit inland and coastal field stations by increasing scientific productivity and increasing the effectiveness of site administration.
- Those sites possessing effective data management systems remain the exception rather than the norm.
- Each field station and marine laboratory should perform a needs assessment to determine where data management fits into its overall mission, and should establish policies and directives accordingly.
- General guidelines for developing data management systems are 1) start small, 2) learn from other related institutions, and 3) find the right persons. Data management plans should allow for incremental growth.
- Training, though expensive, is likely to provide long-term benefits in productivity.
- Close communication between investigators and data managers is essential, but a site's data manager(s) should report directly to the site administrator, rather than to an individual investigator. Investigators and other site users should be involved in continual evaluation and review of data management.
- Site policies should reconcile the conflict between investigators' proprietary rights and general accessibility to data. A data ethic should be encouraged, which maintains that it is unacceptable for data sets with general utility or long-term value to remain permanently inaccessible.
- Data management should be viewed as an appropriate and necessary expense in research budgets.
- Those persons who evaluate research proposals or perform site reviews should examine how data resources are being cared for. However, funding agencies should not enforce unreasonable standard data formats.
- A mechanism is needed by which small investments, perhaps in the $5,000-$15,000 range, are available to get data management programs started, especially at new or small sites. These programs should be focused on specific general-use databases.

Data Standards for Collaborative Research

- Long-term studies and research on regional or global phenomena require the development and use of standards for documentation and exchange, so that data gathered at different times and places can be brought together for comparative analysis.
- Standards should be developed only for specific needs, with full consideration and involvement of the people who are intended to benefit from them, and should not be arbitrary or overly restrictive.
- The test of adequate documentation is that it should contain sufficient information for a future investigator who did not participate in collecting the data to be able to use it for some specific purpose.
- The Intersite Archives File Structure (Appendix D) is a recommended protocol that can be used by field stations to store and exchange data and documentation.
- A series of workshops should be funded to provide training, help field stations exchange information on data handling, and produce shareable databases. In the process, standards will be developed or adopted as needed.
- Multi-site, network-accessible databases should be funded as pilot projects.

Computer Systems for Data Management

- The single most important component of a computer system for data management is dedicated staffing to implement and operate it.
- No single computer system will be appropriate for all stations and laboratories. Systems must be tailored to achieve specific levels of data management and fit within resource constraints.
- A "top down" approach should be used in selecting computer hardware and software. The selection process should focus on data management and research tasks and the software and hardware needed to address them.
- Connection of a field station or marine laboratory to one or more wide-area networks can greatly enhance opportunities for scientific collaboration and help reduce the isolation that researchers at field stations often experience.
- Rapid changes in technology make good communication (electronic or otherwise) between data managers at different field stations critical.
- The best protections against loss of archived data are continuity of management and a strong data archiving policy. Technological backwaters and deterioration of media can be avoided by data managers who remain alert to changes in their computing environment and are aware of media limitations.
- An expansive definition of a computer system for data management can include facilities for visiting researchers. In some cases computers and computer access by visiting scientists to resident data bases are critical to the success of scientific investigations.

CHAPTER 1-DATABASE ADMINISTRATION

William K. Michener
Baruch Institute
University of South Carolina

and

Ken Haddad
Florida Marine Research Institute

1.1.0 INTRODUCTION

In the scientific process, answers to questions about the real world are coaxed out of data sets containing observations of patterns and processes. Various methods may be employed, but all rely on the availability of high quality, well documented data. All scientists participate in data management activities to varying degrees. Data management may therefore be viewed as a critical component of the scientific process.

Science builds on past knowledge which serves as a basis for future advances. The research community associated with field stations collects environmental data that represent a national resource which should be conserved for posterity. These data can, in many cases, be used to examine the effects of global change, loss of biodiversity, and habitat degradation. Scientists working on site-related or ecosystem-related research have a responsibility to future scientific efforts. Through improved preservation, access, and management of data, scientific research can be enhanced.

Ideally, all field research sites, stations and laboratories would have a data management system to serve the needs of current and future research. A data management system consists of both physical and functional attributes. Physical attributes include the people, hardware and software that are necessary to manage a site's database. Basic data management functions that are typically implemented to varying degrees at inland and coastal field stations include:

a. Record keeping of ongoing research (who, what, where, and when)
b. Organization of historical information (history of research and land use activities at the site, facilities development, site personnel, institutional support, etc.)
c. Facilities support
d. Individualized project support (data entry, file maintenance, security, documentation)
e. Acquisition and maintenance of basic databases for use by multiple investigators (specimens, maps, species lists, meteorological and hydrological data, etc.)
f. Data archiving
g. Communication of data (maintain public database, network with multi-site projects)

Data exist in two primary forms at field stations and marine laboratories, site information and researcher specific data. These site-information data sets include:

1. Data on the user base

   Lists of researchers and projects
   Mailing lists
   User statistics

2. Bibliographic data

   Library catalogs
   Published papers about the site
   Theses and dissertations and reprints

3. Site characterization data

   Meteorological data
   Hydrographic records
   Notes on land use

4. Inventories

   Species lists
   Collections
   Maps and photos

Researcher specific data are generated by individual research projects and may or may not be of interest to subsequent researchers at the site.

In the following sections, we examine the benefits of data management; its current status at field stations throughout the country including obstacles to implementation; a blueprint for planning a system; suggestions for implementation; and a discussion of costs and evaluation. Since many administrators are exploring ways to store, retrieve, and analyze spatial data relevant to their sites, a separate section (Appendix B) is devoted to discussion of geographic information systems (GIS).

BENEFITS OF DATA MANAGEMENT

An effective data management program can directly benefit a site in two ways: (1) it can increase scientific productivity and (2) it can increase the

efficiency of site administrative activities. Additional indirect benefits such as expansion of the field station's financial resource base may accrue as funding agencies continue or expand support in relation to the increasing value of that site's data as a resource.

Increased Scientific Productivity

A data management system which reduces duplication of efforts, facilitates awareness and communication of a site's data resource, and leads to better coordination of research efforts can significantly increase scientific productivity.

When data are made more freely accessible, use of data is expanded, reinterpretation of previous studies is possible (perhaps with the help of new types of analyses), an historical record for research and site use is established, duplication of effort is reduced, data are incorporated into the literature more rapidly, loss of data is prevented, and misuse of data is more easily discovered. For data sets with general utility or long term value, permanent inaccessibility is unacceptable.

Every site can benefit from a "data ethic" based on a self-evaluation of its treatment of data resources. Such an evaluation can lead to greater awareness of the current and potential value of a site's database and a recognition that specific data management activities may preserve and even enhance the value of that resource.

Increased awareness of data availability at a site through the production of a catalog of data, site bibliography, and data archive can often reduce the need to perform pilot studies and may facilitate experimental design and implementation.

Many data sets (e.g. meteorology, water quality, habitat characteristics, species lists, etc.) are of general interest to a large number of scientists. However, each scientist cannot always justify the costs of individually collecting, storing, documenting and performing the data management activities necessary to maintain the complete variety of data sets which may have relevance to his or her specific research interests. Even when scientists do have the resources to compile such data sets, they usually do not have the resources to provide the long term care necessary to make them available to a wider audience.

Sites may choose to fund, collect and store some data sets as a site activity. Relevant examples are the locations of field sites (past and ongoing), meteorological data, and other data sets which are site-specific, but of general interest and long term value. Well documented and archived meteorological and habitat data can facilitate the planning of experiments and sampling regimes by providing details about seasonal weather patterns and historic sampling locations. These activities must be carefully selected on the basis of generalized needs of the site's users. Sites may also wish to act as custodians or archivists for individual researchers' databases.

Where long-term data exist, it may be possible to place short term experiments into a broader temporal context. New studies may be more efficiently designed and implemented when they can be coordinated with ongoing research projects.

When data are managed as a long term resource, new investigators are often attracted to a site, and the potential exists for participation of that site in larger scale (intersite, regional, and global) comparative studies. Research sites appreciate in value as their historical databases grow.

Service to Researchers

A data management system which has the appropriate support staff can increase the efficiency of individual scientists by taking over responsibility for routine data management activities. In addition, data management consulting services provided to on-site investigators and visiting scientists regarding design and implementation of data sets, analytical tools available for interpretation of data, and hardware/software training can greatly increase scientific productivity,

Investigators working without the assistance of an organized data management system may not realize how much of their time is spent on data management. Having a site-sponsored data management system in place will not eliminate the need for investigators to spend time on data management activities; however, their productivity should increase as less time is required for more routine data management tasks.

Development and implementation of quality assurance and quality control procedures can facilitate scientific research through detection of data corrupted by human and machine errors as well as by media degradation. Other tasks, including data documentation support, data archival, and translation of data from one format to another, can often be more efficiently performed by experienced data management personnel. However, the investigator should always be involved in the process whereby the data are merged into a site's long term database system if quality is to be assured and documentation maintained.

EFFICIENT SITE MANAGEMENT

Data management can provide the means to document the project and site output that forms the basis of financial support for a field station's resources and activities. It can also enhance communication with off-site investigators, funding agencies, and institutions. Maintenance of ongoing and historical data sets can facilitate monitoring of the biological integrity of the site and provide data necessary for site impact assessment studies. The

data necessary for balancing the selection of new research sites with the need to preserve the integrity of historic research sites can be cared for.

Many activities, such as visitorship, laboratory space management, and vehicle and equipment scheduling, are not usually perceived as data management, yet most field station managers perform this kind of data management on a day-to-day basis. This information is lost when there is no policy or mechanism for retention, and the data are discarded after use. The loss of these data results in lost opportunities for long range planning and improving the economies of site management.

CURRENT STATUS AND OBSTACLES TO IMPLEMENTATION

Despite the potential for increased scientific productivity, expansion of a site's financial resource base, and facilitation of site administrative activities, sites with effective data management systems remain the exception rather than the norm.

The reason for the slow and sporadic development of data management systems is sometimes attributed to the lack of an adequate staff and sustained funding. Understaffing is a problem on all operational fronts (see Chapter 4, Administration and Personnel), and data management is a time-consuming task whose needs are often underestimated.

However, effective data management systems may also be slow to develop for a number of other reasons related to: (1) a lack of recognition that, in addition to habitats, physical facilities and personnel, data are the most valuable resource that a site possesses; (2) an unrealistic or inadequate assessment of site-specific needs; (3) a lack of agreement on goals and priorities; (4) a lack of integration of data management into the overall site administrative scheme; and (5) a lack of communication among site administrators, researchers, and data managers.

BLUEPRINT FOR DATA ADMINISTRATION

The basic administrative tasks involved in establishing a data management system can be briefly stated as:

1. Identifying the user community, inventorying data and assessing their importance in light of the field station's mission.
2. Developing a data management policy appropriate to the mission and user/data profile.
3. Developing a list of data management priorities and assessing the methods and hardware/software options necessary to address those priorities.
4. Developing a justification for enhanced allocation of staff and budgeting resources devoted to data management needs, based on the preceding analyses.

Without the support of site administration, a viable data management system cannot be realized. Site administration, in conjunction with the research community, must be responsible for design, implementation, and continued support of data management. The design phase requires adequately addressing the data management needs of the present and future community of researchers likely to use the field station. Performance of a needs assessment will help determine where data management fits into the overall site mission.

Implementation of a data management system requires that considerable attention be paid to staffing, incorporation of data management into the administrative hierarchy, and funding. After initial implementation of a data management system, continuing support activities (including evaluation and management of incremental growth) must be performed. The design and implementation phases are discussed in further detail below.

INSTITUTIONAL COMMITMENT

Without an institutional commitment there can be no guarantee of continuity, and data management activities will likely be characterized by responses to short term, project-specific requests rather than the comprehensive support which is possible with a broad and well-integrated system.

NEEDS ASSESSMENT

Each station should do an assessment of its own needs and priorities. Stations differ in their needs and their ability to support data management. Some can support higher levels and intensities of data management than others. The following list presents some questions which should be examined as part of a needs assessment.

1. Mission, goals and objectives of the site:

   Is it a preserve?
   Is it a teaching facility?
   Is it a research facility?
   Does it support its own researchers or seek to attract visitors?

2. Type of scientific data being collected:

   Is it descriptive and of general interest to various researchers?
   Are there ongoing projects? Projects of historical interest?
   What databases exist?

What databases not currently available can potentially be recovered and made available?
What databases are anticipated in the future?
Do the databases relate geographically/biologically?
Are the databases of a short-term or long-term nature?
Are databases in analog or digital form?
Do important data sources consist of photos, imagery, video, etc.
What levels of scale and/or scope are represented by the various databases (subcellular to landscape)?

3. Volume of activity:

What is the **current and projected number of** researchers/students?
What is the size of historical and current databases?
How many potential and actual users exist?

4. Sophistication of data generating, processing and managing activities:

What computerized storage facilities exist?
Is there access to off-site resources?
What is the potential for storage and access?
What kind of processing services and equipment are available?
What levels **of expertise do the on-site personnel** possess?

5. Infrastructure:

What are current and potential sources of support?
Is there an on-site library and what are its capabilities?
Can the library be used as a service node for data access?
What kind of data acquisition equipment is available?
How many support personnel are on-site?
Is it a seasonal or year-round operation?

Planning

In setting priorities, a station should identify the potential level of data usage, determine common needs, and identify potentially valuable long term data sets, including historic, current, and future data sets. Data management priorities, like research priorities, can be viewed as a compromise between what can be done and what should be done. Addressing the following questions in the light of research priorities may help set priorities for data management: (1) What do I, as a scientist, wish I knew about the history of a site? (2) If I could go back 50, 100, or 1000 years, what would I record for the future? (3) What present conditions are important enough to record for posterity? and (4) If I were presented with an historic data set, what ancillary information would I need in order to effectively make use of the data? These may be difficult questions to answer but may suggest actions to be taken.

Investigators should be consulted before a site establishes guidelines. Speculation and contemplation of future needs and priorities should be encouraged. Agreement among the on-site researchers and the external scientific community should be sought. By addressing the needs of the research community through an assessment process, one can avoid forcing unnecessary or unreasonable standards on investigators for such things as data storage and data transfer formats.

Priorities will also be affected by changes in the goals of the station and the parent institution. Funding sources can affect priorities, but they should not drive the process.

Data Archives

Many stations, after assessing needs, will conclude that they need to archive data for subsequent retrieval. Research at a site will be greatly enhanced when other data sets from that site are available. Many data sets have broad or long-term significance and should not be lost. Funding and infrastructure will be needed to support them. Stations that take on this responsibility need to ensure that important data are appropriately deposited in a system that is secure, yet allows reliable retrieval. This can be done on-site or off-site. In either case, the issues of access, longevity and quality should be addressed:

1. Access

volume of data
volume of requests
level of interest
documentation
data formats
cataloging
ownership of data
remote/on-site

2. Longevity

primary storage media
changing formats
physical stability of media
redundant storage

3. Quality

multiple versions
documentation
expertise for monitoring quality
standards

Plans for data archiving should take into consideration the volume of data, projected number of requests for access to it, and the anticipated level of scientific interest. The potential use can vary from a small number of scientists interested in addressing a site-specific question to a much larger inter-disciplinary scientific community that would use the database together with those from other institutions to address regional or global issues.

Regional data storage is reasonable if the volume of data and the use levels are high. Data documentation and cataloging of the data sets are crucial for access whether on-site or off.

A site should consider initially storing data sets in a standardized generic format (ASCII). This would allow flexibility in moving data sets and in accessing them remotely. The question of data ownership should be addressed early in the design phase.

The issue of longevity requires consideration of changing formats and the physical stability of media. There may be a need to ensure access through redundant storage. Disasters can destroy data on-site and off-site. If data sets have been prioritized and the most used and most critical stored in more than one location, access is preserved. The management of redundancy must be integrated into the site's data management plan. The plan should address updating data sets to avoid multiple versions which lack adequate documentation. Data sets should be tracked to manage the numbers of copies and to allow for purging of outdated data sets.

Data quality is of special concern to scientists who use data they did not themselves gather. A station must assure the highest degree of quality control over its own data and provide full documentation of data obtained from elsewhere for its own researchers. Disclaimers should be stated where appropriate.

Quality of data is linked to the development of standards for data generation and documentation. Researchers should be encouraged to fully document data before submitting it for archival storage. Participation of the scientific community in designing and implementing data set documentation can be a very valuable step towards insuring that data sets are usable in the future.

Accessing and archiving data costs money. Ideally, cost should not be a barrier to access. To encourage shared databases, stations should strive to supply data sets free of cost to those scientists who participated in their development. Data should be accessible to others at minimal or no charge, but cost recovery may be appropriate and necessary. Legal and institutional obligations regarding data accessibility will need to be addressed at each station.

One solution for small stations may be participation in development and maintenance of regional or national data banks. Data archiving in such data banks could prove cost effective, but the concept needs additional study.

The best protections against loss of archived data are continuity of management and a strong data archiving policy. Technological backwaters and deterioration of media can be avoided by data managers who remain alert to changes in their computing environment and are aware of media limitations. A data archiving plan or policy can and should ensure that a data manager remains alert and sensitive to potential problems.

IMPLEMENTATION

Successful implementation of a data management system requires that site administration pay particular attention to: (1) development of a reasonable plan which supports incremental growth, (2] effective incorporation of data management into the administrative organization, (3) assessment of costs and procurement of necessary funds and staff, (4) proprietary rights, (5) continuity of management, and (6) continuing evaluation. The research community and site administration, along with data management and funding agencies where appropriate, should discuss and agree on goals and priorities before beginning the implementation.

A plan for data management should allow for incremental growth. The guidelines are: (1) start small, (2) network, and (3) find the right person.

Start with a pilot project small enough to be accomplished within a reasonable time frame but important enough to have a beneficial impact. An example might be development of a reprint list, taxonomy lists, collection list, or acquisition and cataloging of aerial photography.

"Networking" means seeking advice from sister institutions with similar size, resources, and mission. Many pitfalls can be avoided by learning techniques used at other sites.

Choose a data manager who has research experience and a scientific understanding of the site's research program but also has data management skills. The person should have an interest in and enthusiasm for the data and products to ensure success. A data management system cannot necessarily be sustained over the long term as a "labor of love," but high enthusiasm and intensity are needed to get it firmly established. Good communication skills and relationships with data users are essential.

Upon completion, the pilot project should be evaluated. Maintenance costs incurred should be considered, because even when the system is established, it will not maintain itself. Lessons

learned on costs and benefits should be used in planning the next project.

Subsequent projects should be chosen by consensus, considering overall site needs. These projects could include more individual specialized research projects but should be prioritized on the basis of cost vs. benefits. At some point it will become apparent that the next incremental step will require people and equipment necessary to accomplish the next tier of objectives. It is generally safe to assume that data management tasks will be more complex and time-consuming than anticipated. Time schedules proposed for projects will inevitably become more realistic as both data managers, scientists, and administrators become more experienced in administration and implementation of specific projects.

Eventually a site archival facility should be established. The data management system should be designed to survive the loss of key data management personnel and changes in research emphasis (driven by both investigators and funding) at the site. Specific arrangements for archiving data either on-site or off-site should be addressed by the site administration. Options include submission of data to the recognized site data management system, the parent institution, or a regional or national data bank. In the event that a station is closed, mechanisms should be established whereby responsibility for maintaining the database passes to the parent institution or "field station community" (a "sister institution" or possibly a regional/national data bank).

A key issue for data management administration is the resolution of any conflict between the investigator's proprietary rights and the need for general accessibility to data. Permanent inaccessibility to data is unacceptable, but the investigator should be able to control access to his/her data for a reasonable period of time. Several issues must be addressed when considering the issue of proprietary rights. These include the potential for allowing others to publish findings before the investigator who collected the data does, misinterpretation of the data by someone unfamiliar with the experimental design or habitat characteristics, and using the data out of context. Questions of legality, including "Who owns the data?" and "Who is liable for misuse or misinterpretation?" must also be answered.

Many universities and funding agencies have regulations regarding the fate of a data set. However, each site should have a policy regarding proprietary rights or should negotiate with individual researchers. In either case, issues of data ownership and access to particular data sets should be clarified with researchers before the project is begun.

If release will hinder research or use of the site, or violate local, state, or federal regulations (e.g., regarding endangered species), the site may wish to restrict access to the data. However, individual sites must look into the pertinent regulations and develop a policy which incorporates them.

Oversight

Ideally, the data manager(s) should report directly to the site administrator (Figure la). At many small sites, a single individual may function as both the site administrator and the data manager. Close communication between the investigators and the data manager is essential, although the site administrator has the ultimate responsibility of directing the data manager while also addressing the needs of the scientific community and responding to influences external to the site. These influences may include: (1) database requests from other scientists, agencies, and institutions; (2) funding agency requirements; and (3) institutional, state, or federal obligations.

Site administrators are cautioned against implementing an administrative hierarchy whereby one of their scientists assumes control of the site's data management personnel (Figure Ib), since this has a high probability of fostering conflict within that station's scientific community. For example, the perception that the personal agenda of the scientist administering data management receives precedence over the broader station objectives frequently arises. This perception, whether based on reality or not, may serve to isolate data management from the other scientists at the site.

Some sites, particularly large ones, may wish to implement an administrative hierarchy whereby the site administrator oversees a "data management steering committee" which periodically reviews the data management system and participates in establishing and prioritizing objectives (Figure lc). For example, although data management personnel may report directly to the site administrator on a routine basis, they may also participate in monthly or quarterly reviews by the steering committee. The steering committee would ideally be comprised of the site administrator as well as a manageable number of scientists who represent the various research programs at the site. This scheme provides an important feedback mechanism to the site administrator and facilitates communication among scientists and data management personnel.

Staffing

Successful data management systems are usually staffed by persons who have a strong interest in the scientific research conducted at the site. This not only promotes effective communication with the

scientific community, but often helps assure retention of data management staff at field stations which cannot offer competitive salaries.

Requisite skills and expertise needed by data management staff will be largely affected by the size of the organization and the length of time the system has been in place. Generally, the more complex the operation, the greater the need for more specialized personnel. At small sites or those with low research activity it is essential that the data manager have expertise in both science and data management and that this individual have access to appropriately qualified consultants. At some sites, primary training in biology may be appropriate, whereas at others a background in chemistry, geology, physical oceanography or other relevant disciplines may be appropriate. In any case, the initial staff member should be a scientist first.

For larger, more active sites a systems analyst/programmer should be added next. It is critical that the data management staff be able to communicate with the scientists and also have the expertise to accomplish what is needed. This means that at least some of the data management staff need to be very broadly trained. Increasing the size of the data management operation brings increasing specialization.

An alternative to making data management an adjunct to computer support is to staff data management as an adjunct to an existing library or museum. This can be appropriate at institutions where the library or museum already has strong ties to an information management and retrieval program. Links to computer support would still be needed but on a secondary basis.

The use of graduate students as a cost saving means is problematic as it may restrict continuity and could cost in additional training time. However, it does allow for student educational/financial support and with careful choice could provide talented personnel. Sites may wish to examine the possibility of developing one- to three-year undergraduate and/or graduate internships or independent study programs to accomplish specific tasks. Some tasks, such as data entry, routine quality assurance, and graphics production, may be more appropriate for temporary personnel.

The ability to communicate with a wide range of people is the most important qualification for a data manager, assuming an appropriate level of technical skills. Data managers must be able to effectively articulate the purpose and needs of the system to site administrators, researchers and the parent institution, as well as field questions and demands from on-site and off-site users. Consistent communication with other data management personnel promotes better and more creative systems.

Site specific technical skills might include organizational or curatorial expertise, experience in data collection, storage and retrieval skills, programming knowledge, networking experience, proficiency at hardware and software maintenance, and GIS experience if appropriate. Again, the technical skills are site specific and vary depending on the size and activity level of the site. Organizational skills are basic and can include library expertise.

Strong administrative and financial support is necessary to attract and retain data management staff. Not only should personnel be adequately compensated, but data management should be provided with the necessary staff, equipment and operating budget for continual database maintenance and update. Administrative support should be demonstrated by ensuring that the data manager reports directly to the site administration. Consensus on priorities is necessary so that data managers can focus their attention on well-defined projects. Close communication with (but not supervision by) the researchers on-site will enable the data manager to be an integral part of the site's research activities, All site-related publications should acknowledge data management personnel and identify where data have been deposited, much as one does with plant and animal specimens.

Costs Associated With Implementation of Data Management

The equipment, staff, space and budgetary resources committed to data management vary widely among sites, reflecting the wide-ranging missions and academic clientele of the nation's network of field stations and marine laboratories. One site with ten scientists may require only a part-time data manager, whereas another with the same number of researchers may require two or more staff members to meet a broader range of data management duties. It is not possible to state a simple formula for the cost. However, it may be helpful in planning the implementation or expansion of data management to consider some scenarios along a continuum in which staffing is the most important limiting factor.

In the first scenario, there is no dedicated data management staff, and no formal data management, although some informal records may be kept. Documentation is spotty, if it exists at all, and back-up copies of data are not maintained. Such a data management system does not require any computer and software resources. There are no long-term benefits to researchers. Without proper data administration it is not a question of whether data will be lost, but when they will be lost.

The second scenario features a part-time data manager typically capable of maintaining only one or two of the types of site data. Initial effort may be focused on identifying, acquiring, and documenting data. Except in isolated cases, data belonging to individual researchers are not managed under this arrangement. Part-time data

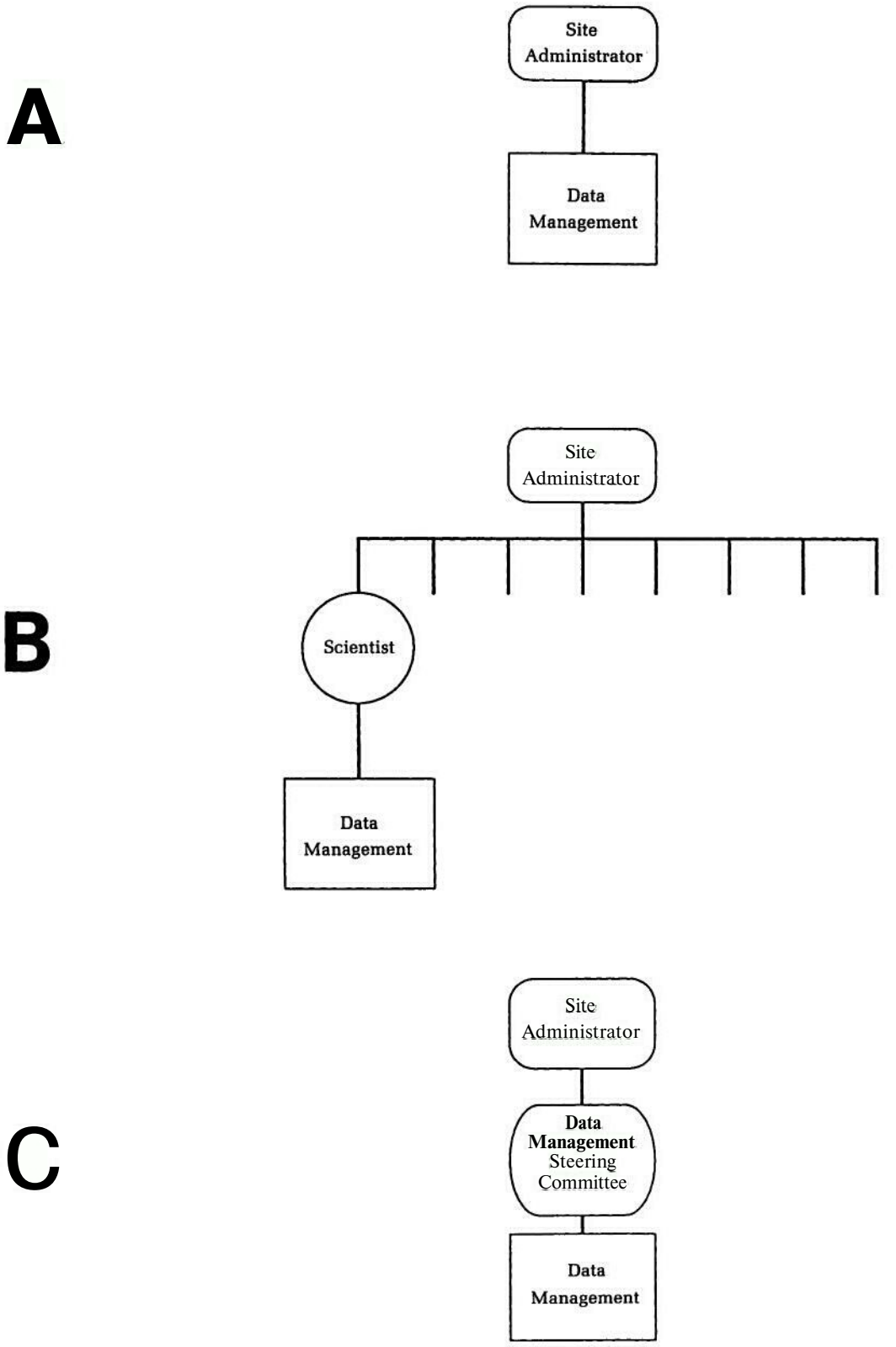


Figure 1. Administrative Structure for Data Management.

managers are typically unable to provide data entry services to researchers or support comprehensive or rigorous error checking.

A data dictionary is likely to be informal, non-integrated, and not automated. A file cabinet or single microcomputer may be the only hardware used for these activities. Software should be "off-the-shelf packages that are in wide public use and which support generic data structures, because there will be little time available for customization.

Expenditures for training and annual maintenance may be minimal, though not because of a lesser need. Part-time data managers are likely to be successful only if they have access to those who can provide training and advice, and if they take advantages of maintenance and other support services for software and hardware. Access to electronic mail networks can be used to get help and advice from other data managers, as well as to facilitate access to data by researchers.

The primary research benefit under this scenario is the creation of a persistent institutional memory, at least about a few selected types of data.

To handle the backlog of historical data sets, or the startup of a new computer system, additional resources typically will be needed.

The third scenario features a full-time data manager. This level of staffing might be appropriate to a site with ten scientists. In addition to managing site characterization and administrative data sets, a full-time staffer may also be able to manage data for a small number of individual research projects. The extent to which this is possible depends on the size, type and complexity of the data sets, and the number of data management "clients." Provision of data entry services for a few individual research projects becomes possible, but may require contracts for service bureau data entry.

Computational environments are more variable at this level, with the more powerful personal computers, workstations, and even mainframes being used to handle the larger volumes of data.

Benefits of such a system include easy access to site information, which can in turn facilitate integration of new research projects and researchers. Because a full-time data manager is available for consultation, individual researchers can be more efficient with the time they spend on data management tasks.

The last scenario is a data management staff comprised of several individuals, typically at a more active site with a large number of scientists. The individuals may be trained in systems analysis, database programming, computational ecology, statistics, or other technical fields. There may also be a full-time data entry staff.

The computational environment is typically complex, with several different types of computers, each used for specific tasks. Computer hardware may include a microcomputer network, minicomputer, or multiple workstations. The larger the site, the greater is the need for more storage capacity and processing power. Network connections, with electronic mail, remote terminal access, and file transfer capabilities, are desirable to facilitate off-site archiving, access to external databases and transfer of data to remote researchers.

Table 1. Four scenarios representing the range of costs associated with implementing data management systems at varying levels of intensity.

Site Activity (# of scientists and users)	Personnel Required (FTE)	Hardware/ Software Costs	Annual Maintenance Expenses	Training
1-5	0.20-0.75	$ 4,000	$ 300-500	Self-taught
10	1	$ 8,000	$ 600-1,000	$5,000
50	2.5 (plus data entry staff)	$ 100,000	$ 15,000 (10% of total budget)	$25,000
100	3.5 + (plus data entry staff)	$1,000,000	$ 200,000 (10% of total budget)	$25,000 +

Annual maintenance expenses for hardware can be expected to consume approximately ten percent of the annual data management budget. Another rule of thumb is that annual recurring costs for hardware and software (i.e. updates, repairs, maintenance, license renewals) are likely to be 8-12 percent of the initial cost. Creative ways can often be found to keep these costs down, but usually at the expense of personnel time.

Training is of major significance. Hardware and software are evolving at a rapid pace and it is difficult for data management personnel to individually track these advances, learn new "tricks of the trade," or readily become proficient with new techniques and hardware/software tools. It has been demonstrated that continued expenditures in training provide a long term benefit in productivity both by the data management staff and the researchers. Training should be given a high priority even though initial expenditures might seem high. It is preferable to maintain an annual training budget and schedule.

Potential Funding Strategies

A successful data management plan must have adequate and stable funding. Potential sources of funding may be the parent institution's facilities budget or re-routing of appropriate portions of overhead costs (e.g. program management charges or indirect costs) to data management. Either mechanism requires full support of the parent institution. Any revenues generated through overhead and indirect costs associated with grant support cannot be considered completely reliable. Though probably not adequate for full support, user fees may be useful to sites catering to visiting researchers and can often result in partial cost recovery. However, this approach may tend to reduce the efficiency of data management if a "pay as you use it" approach is adopted and no mechanism is established for long term support of the facility and data management personnel.

No site/institution can rely on short term (1-2 year) grants to support ongoing data management. Funding for base level data management activities should be done with hard money. However, short term grant money can be very useful for providing start-up hardware/software, training, and other special projects. Data entry and/or conversion of previously collected data to appropriate formats, support for incremental improvements to existing systems, and the underwriting of publication costs (electronic or traditional) are possible uses of grant money. Short term grants can be used to implement specific data sets (station bibliography and data catalog, species lists, etc.). However, mechanisms should be established to cover the recurring costs of maintaining data sets after this initial investment.

Role of Funding Agencies

Data management represents a real cost for research. Therefore, it should be viewed as an appropriate and necessary expense for grant budgets. The challenge to funding agencies is to encourage ties between data management, field stations, parent institutions and the research community. Funding agencies can foster the development and support of data management systems by providing start-up funds for hardware and software, by supporting research in data management (e.g. development of more efficient database structures and quality assurance procedures, etc.), by supporting training programs, and by developing mechanisms for supporting database development.

Data management can be included as a line item in proposal budgets and as a topic to be examined during the review process. Although funding agencies should not force unreasonable standards on scientists for items such as data storage or transfer formats, they can encourage retention of data at field stations and elsewhere by asking scientists to state in their proposals what, if anything, will be done with the data when the study is completed, or as data are acquired. Referees should be encouraged to consider these factors when evaluating proposals. However, scientists should be assured of proper acknowledgment for the use of their data in any subsequent publications.

Funding mechanisms are needed for getting field stations started in data management. An initial investment of $5,000-$15,000 may be all that is needed to get a data management program off the ground, especially at new or small sites. A system of "mini-grants" would assist small stations in implementing a basic data management program. The possibility of internal reviews or mini-reviews for proposals of this size should be examined. The "seed project" model presently used by some funding institutions may be appropriate for initiation of data management at field stations.

Funding agencies might explore the possibility of funding the development of regional/national data banks tor archiving of data from field stations. This might reduce the need for an extensive data management system at small sites.

Another important potential role for funding agencies is support for training data management personnel in the needs of field stations and providing them with the necessary skills and support group to meet these needs. This might be accomplished by sponsoring regional two or three day workshops or an exchange program whereby personnel visit sites with data management systems in operation.

EVALUATION

Appropriate site priorities depend on continual evaluation and review. Site users as well as administrators are essential participants in the evaluation. All sites need an evaluation process, but the larger the organization, the greater the need to establish a formal process. Site review committees, whether for funding agencies or inhouse obligations, should make data management a high priority. The overriding objective is to establish and maintain effective communication among and between the scientists, data management, and site administration. An additional objective is to develop and monitor communication with external users and data management personnel.

The intensity of data management will likely vary for different databases at a particular site, based on some prioritization of its value to users. The data management program should be reevaluated periodically to reflect changing user needs and priorities. Data management objectives should ideally always be closely linked to research and administrative objectives.

CHAPTER II—DATA STANDARDS FOR COLLABORATIVE RESEARCH

James W. Brunt
University of New Mexico

and

Warren Brigham
Illinois Natural History Survey

RESEARCH NEEDS

Increasingly, environmental scientists are being encouraged to focus attention on regional and global issues such as biodiversity and global change. The wide geographic distribution and diversity of ecosystems encompassed by inland and coastal field stations represent a major national resource. To address large scale environmental questions, scientists will require resources such as the data generated at these facilities.

Large scale scientific issues elevate the importance of data management beyond the needs of the individual investigator. When data are regarded as "belonging to science" and, therefore, to be shared with other researchers now or in the future, data standards to enhance communication become necessary.

By using standards, researchers can save time and prevent costly mistakes in interpretation of data. The activities that suffer most from lack of standards are the arranging and organizing of data, documenting of what has been done, and sharing and exchanging of data with other researchers.

Implementation of standards is particularly important where several researchers are working on a joint project. For example, if all investigators on a project adopt standard location descriptors, all localities referred to in data sets can be communicated to other data users reliably and accurately, and the data sets can be used for comparative analysis. In addition to project-wide data standardization, site-wide standardization can result in similar benefits to both the investigators and future users of the research site and its historical data. Current data sets can be compared with future data sets.

Electronic networks are making the sharing of scientific data for comparative analysis much more feasible. Field stations, herbaria, museums and other biological information providers benefit greatly from the data communication channels provided by research networks such as the Internet. Biological databases can be made immediately accessible to ecologists, systematists and conservationists around the world (Appendix C). But the success of network accessible databases in biology depends on the ability of the disciplines and subdisciplines to reach consensus on elementary data models and database structures.

Data standardization at the various levels, from the raw (primary) data to structures for user access and network exchange, should have as its primary goal the advancement of the science. The emphasis should be on those standardization strategies that maximize the conduct of science and the use of the data, at any level of the information management process.

Application of standards does involve costs, however. Perhaps the greatest cost is in instances where databases must be converted to comply with "newer" standards. This implies that carefully designed standards are best applied early in the development of data management at a particular site.

Creation and implementation of data standards should not be done in an arbitrary or overly restrictive manner such that the researcher's ability to collect and process data is restricted. Proposed data standards should be examined and applied only if they enhance data management. The need is not for standards that are in some sense sophisticated or elegant, but rather, standards that active researchers will in fact use to document and archive their data.

There can be benefits to having discipline-specific standards for representing space, time, and the relevant physicochemical data associated with biological information, but the appropriate persons to develop such standards are those researchers who need them.

TYPES OF DATA STANDARDS

Organization of Data

Organization of data refers to the logical structure of data — what all the variables are, how they should be organized into different types of records, and how the variables and records should be arranged with respect to each other. Standards for organization of data make it easier for scientists to analyze and re-analyze their own data as well as share it with other researchers. The 1982 Workshop Report, Data Management at Biological Field Stations (Appendix G, Chapter 2), describes data standards with respect to design of data sets which, if

applied, would enhance data management at any site with little adverse impact on a researcher's activities. The information presented in that document is still relevant to the management of biological data.

Data Documentation

Any researcher who has tried to produce syntheses integrated over space and time using previously collected data, including data from other researchers, has probably experienced frustration due to inadequate documentation of the data.

If the documentation describing a particular data set is lost, the data become useless. While this is particularity true for archived or historical data sets, lack of proper documentation can affect any data file. Thus, for exchange and archiving, data documentation should be incorporated with the actual data as soon as practical, possibly even in the design phase of the research.

The test of adequate documentation is that it should contain sufficient information for a future investigator who did not participate in collecting the data to be able to use it for some scientific purpose.

The 1982 Workshop report describes a standard for data documentation (Appendix G, Chapter 2). Some of the information categories may not be applicable to the data at every site and some additional categories may be needed for "non-traditional" ecological data sets (i.e. remote sensing and Geographic Information System files), but the essential elements are present.

Data Exchange

Field stations and marine labs represent a heterogeneous research and computing environment. The independence and isolation of field stations has led to a tremendous variety of data management approaches usually tailored to local needs, but which make data exchange and collaboration difficult. Although inter-university and international computer networks are becoming accessible to field station user, some standards must be followed to use them for data exchange.

There is a need for a non-restrictive but powerful common-denominator structure for data sets that will encourage good practices of documentation and communication. A complete data set should be an entity that contains all of the relevant documentation, as well as a history of the data. To be complete, documentation should include comments and annotations about the data set as a whole, and also about the individual records and observations where necessary.

One generic file structure that can be used is the Intersite Archives File Structure (Conley and Brunt, Appendix D). It can facilitate an orderly approach to the design and implementation of field station and marine lab data exchange capabilities. The structure is one that includes full documentation and comments with the data. It solves the problem of possible separation of the data from the documentation. The data itself can be of any basic type, such as statistical data, text data, graphics data (e.g. files that can be written to a graphics plotter), gene sequence data, or bit map image data.

IMPLEMENTATION

The Development and Adoption of Standards

There are several possible routes to the development and adoption of data standards. At one extreme, de jure standards can be put in place by fiat. At the other, de facto standards can be adopted by survival of the fittest.

Standards imposed from above, without full consideration and involvement of the people who are intended to benefit from and use them, are usually ignored. Equally suspect are standards promulgated for political purposes, by institutions eager to enhance their own standing, without regard for research value and technical merit.

On the other hand, standards developed through a completely *ad hoc* process tend to be developed inefficiently, with much reinventing of wheels. Standards developed in this manner tend to lack rigid definition, so that there is no way of knowing whether compliance is apparent or real. These standards, too, may have political value, in that one can easily (and truthfully) claim compliance; but very little efficiency is gained.

A relatively non-dictatorial process somewhere in between these two extremes will involve researchers and data managers in developing, testing, and using standards relevant to the full array of types and uses of ecological and environmental data.

We recommend that specific standards be developed (1) through a series of workshops at which technical resources will be examined to address specific standards and data topics, and (2) by increased use of communication networks (including both electronic and personal networks) of biological field stations and marine laboratories.

Workshops

These workshops should not be limited to the narrow issues of standards, but should include information on technology for scientific data handling in general, such as data acquisition systems, data analysis tools, data handling in general, and data exchange. Opportunities to exchange this sort of information are currently quite limited. Training programs and seminars, as well as joint efforts to create shareable databases, are needed.

Progress in data standards will be made through common consent and practice, utilizing the expertise of those with relevant experience. Training can be provided by persons knowledgeable in fundamental data management principles as they apply to scientific data.

Data managers and researchers who have dealt with issues of bringing together two or more data sets for comparative analysis have much to contribute. Librarians in the field of information science, systematists, and museum curators who are affiliated with field stations will also have relevant experience.

In order to move beyond generalities and down to practical issues, each workshop should deal with specific issues, such as electronic networking or data archiving, or with specific types of data, such as species lists, site bibliographies, data catalogs, climatological data, or spatial data (Table 2).

The result should be shareable databases that relate directly to the testing and evaluation of scientific hypotheses. In the process, standards will be proposed, developed, and tested.

Network-Accessible Databases

The development of multi-site, network-accessible databases, such as those being developed in the plant systematics community (Appendix C) should be encouraged. Especially valuable are projects which bring people from different sites together to apply their complementary areas of expertise. Bringing data together from two or more sources will require development or adoption of standards. Even more importantly, making those data accessible on the network will immediately test the usability and usefulness of those standards and of the entire concept of shared databases.

COMMUNICATION AND EVALUATION

Any workshops or network projects should place great emphasis on communicating information about their findings and products to the community of over 200 field stations and marine laboratories. It is expected that this would encourage further testing and evaluation of the utility of standards and databases.

Table 2. A series of workshops to provide training and information exchange and produce shareable scientific databases.

	Technology-Oriented Workshops
Electronic Networking	Collaboration via data exchange will benefit from communication technology and expertise. Participants will learn how to use electronic mail, network file transfer, and remote access capabilities. The product of this workshop should be a plan for network access via Internet/NSFnet.
Data Archiving	Methods for data storage and archiving are developed. Some attempt should be made to identify the types of data appropriate for intersite access.
	Product-Oriented Workshops
Species Lists	Example: Systematize lists of species at the field station and marine labs. Species inventories are basic to biodiversity studies. Strategies for data update and exchange should be studied. For certain groups, development of a central database may be appropriate.
Site Bibliography Development	Data catalogs and site bibliographies need to be developed for every site in an exchangeable manner.
Meteorological & Hydrologic Data	Develop standards and methods to share these types of nearly ubiquitous data.
Spatial Data	Develop spatial data standards for geographic information systems, global positioning systems, and remote sensing, etc.

CHAPTER III—COMPUTER SYSTEMS FOR DATA MANAGEMENT

John H. Porter
University of Virginia

and
Jeff Kennedy
University of California Natural Reserve System

INTRODUCTION

Research data management is increasingly linked to computers and associated technologies. Properly integrated hardware and software systems are crucial to managing large amounts of data. This chapter examines:

1) hardware and software selection
2) typical data management computer systems
3) uses and implementation of networks (both local and wide-area)
4) technological innovations that will influence data management methodologies, and
5) computer systems required for archival storage with special emphasis on the needs of marine laboratories and biological field stations.

The 1982 workshop report, Data Management at Biological Stations, (Appendix G) provides excellent guidelines for the management of scientific data at field stations. It includes comprehensive and thoughtful discussions of software and computer systems for data management, providing a blueprint for a complete data management system. However, at a majority of field stations, the 1982 recommendations for computer systems and software remain unimplemented. This is somewhat surprising, because advances in computer and network technologies solve many of the problems identified in that report. For example, the anticipated "proliferation of microcomputers with nonstandard floppy disk formats" failed to materialize, and a few standard formats have emerged. Moreover, the extensive use of wide-area and local-area networks, which was unanticipated in the 1982 report, has reduced the need to exchange data on physical media and thus has reduced physical barriers to data exchange.

The computational environment has also become increasingly homogeneous. The distinctions between the capabilities of mainframe, mini- and microcomputers drawn in the 1982 report are rapidly diminishing. Increasing numbers of software packages run on both mainframes and microcomputers. The dominance of certain software packages in microcomputer markets has led to emergent standards for exchanging data between different brands of software, computers, and operating systems. We anticipate that the rapid improvements in price and performance will continue at an accelerated pace.

Many of the shortcomings of relational database and statistical software described in the 1982 report have also been reduced. Although there are still improvements to be made regarding data documentation (i.e., data about data, or "metadata"), many of the problems associated with the entry of textual information have been reduced or eliminated. The increasing use of Structured Query Language (SQL) by relational database packages also provides opportunities for increased standardization.

Technical advances have led to new challenges for data management. For example, the increasing use of scanners and graphical and audio data (video and remote sensing imagery, maps, photographs, and sound recordings), with their large file sizes and specialized formats, creates problems regarding data storage requirements and file exchange compatibility.

Given that the technical barriers to achieving a functional data management system have decreased, the general lack of success in fully implementing the systems envisioned in the 1982 report seems paradoxical. The consensus among workshop participants and questionnaire respondents was that these recommendations remain unimplemented in significant part because the single most important component to a successful data management system is dedicated staffing to implement and operate it. Even the most "user friendly" software interfaces and most powerful computer systems are useless for data management without dedicated individuals (possessing the requisite computer expertise and interpersonal skills) to run them.

MANAGEMENT STRUCTURE

At many biological stations and marine laboratories, data managers are expected to provide computer system support (e.g., configuring hardware, managing and installing software, and administering networks) as well as perform data management functions. It is our recommendation that data management and computer system support duties be separated whenever feasible. The advent of complex multitasking operating systems on personal computers (UNIX, OS/2, Multifinder) and sophisticated networking software can cause a significant increase in the time taken by system support tasks and concomitant decrease in the time available for data management.

SELECTING COMPUTER HARDWARE AND SOFTWARE

It is a striking comment on the rapid innovations in computer technology over the past few years that choice of a computer system is increasingly arbitrary. The distinctions between the general types of tasks that can be performed by mainframe, mini- and microcomputers has all but vanished (although there still may be significant speed differences between different size computers in performing large tasks). The convergence of mainframe, mini- and microcomputer hardware and software means that it is increasingly necessary to take a "top down" approach, centered on data management and research tasks and the software needed to address them, rather than a "bottom up" approach that starts with the selection of computer hardware and software (Figure 2).

The first step is to identify the tasks to be performed by the system: What sorts of data will be managed? What sorts of manipulations or analyses will be needed? How large are the data sets likely to be? How many users need to be supported? Based on the answers to these and similar questions software can be identified that are capable of supporting these tasks.

Software

Typically, there will be several products with similar capabilities (e.g., all relational database programs share certain basic features). Deciding between them will depend on the particular characteristics of each roduct: functional strengths and weaknesses, cost, speed, ease of use, prior familiarity, compatibility with other software, availability of adequate user support (from the vendor, from a knowledgeable user or users group, or from the department or parent institution), and the financial stability of the vendor (i.e., will the company still be in business five or six years from now).

Most popular software packages have their share of proponents and detractors, whose differing opinions depend largely on their familiarity with the package in question. For this reason, it is good to query several sources regarding each package. Preferably, you should test the software yourself, using your own data.

Hardware

Once potential software packages have been identified, it is time to select the type of computer to purchase. In some cases, software will only run on a computer produced by a particular manufacturer and the choice is easy. More likely, there will be a variety of computer options, each capable of running the desired software. Obvious factors to consider are cost and processing power, but equally important are capabilities for expansion and obtaining a good fit with your institutional and user environments.

Constraints in the selection of computer hardware may dictate that some software choices be reexamined to improve integration with the hardware and to maximize how well the components function as a system. The evaluation and selection process thus becomes an iterative process. Appendix E provides more detailed guidance for hardware selection and a list of software products which were popular with workshop participants. Some products that emerged in the year following the workshop are also listed.

Selecting GIS Systems

Selection of software and hardware to implement a Geographic Information System (GIS) is an extremely challenging task. The task is complicated by the diversity of software products and approaches and by the complexity of GIS software. The term GIS covers a wide range of activities, ranging from computerized cartography to spatial analysis. No one system is best at all types of GIS work. As with selection of a general purpose data management computer system, a "top-down" approach is recommended. The first step, task identification and needs assessment, is covered in Appendix B and will not be addressed here. Selection of software must be based on the ability of specific packages to perform the required tasks, the potential for expansion, the ability to interface with existing digital data sources, ease of use (which has a major impact on the amount of training required), the initial cost of the software package and the cost of continuing support and licensing.

Once a package has been selected, decisions must be made about the hardware configuration of the system. Will the GIS be directly accessible by station researchers or only by station personnel? How many "seats" will be provided and how will they be implemented? For some systems single-user workstations (typically personal computers) may be an economical choice. In order to balance computationally intensive tasks (e.g., producing polygon overlays or network analyses) with display intensive tasks (e.g., digitizing and editing data layers) it is recommended that single-user workstations be provided with sufficient memory and software to support multitasking. For other stations, multiple single-user personal computers sharing peripherals over a LAN, or multiuser computers and terminal workstations may make more sense and facilitate centralized administration of data and computer systems. Local area networks may also play a critical role in permitting sharing of large data layers, reducing redundancy and simplifying system administration.

Obtaining sufficient online storage for large data layers is often a problem. The large size of GIS files, the number of intermediate files generated by GIS processing, and the cumulative nature of GIS data acquisition combine to strain the resources of all but the largest systems. For this reason, provision for very large data storage and backup capacity is a firm requirement for GIS computers. Inclusion of high-capacity off-line storage for backup and archival storage is strongly recommended.

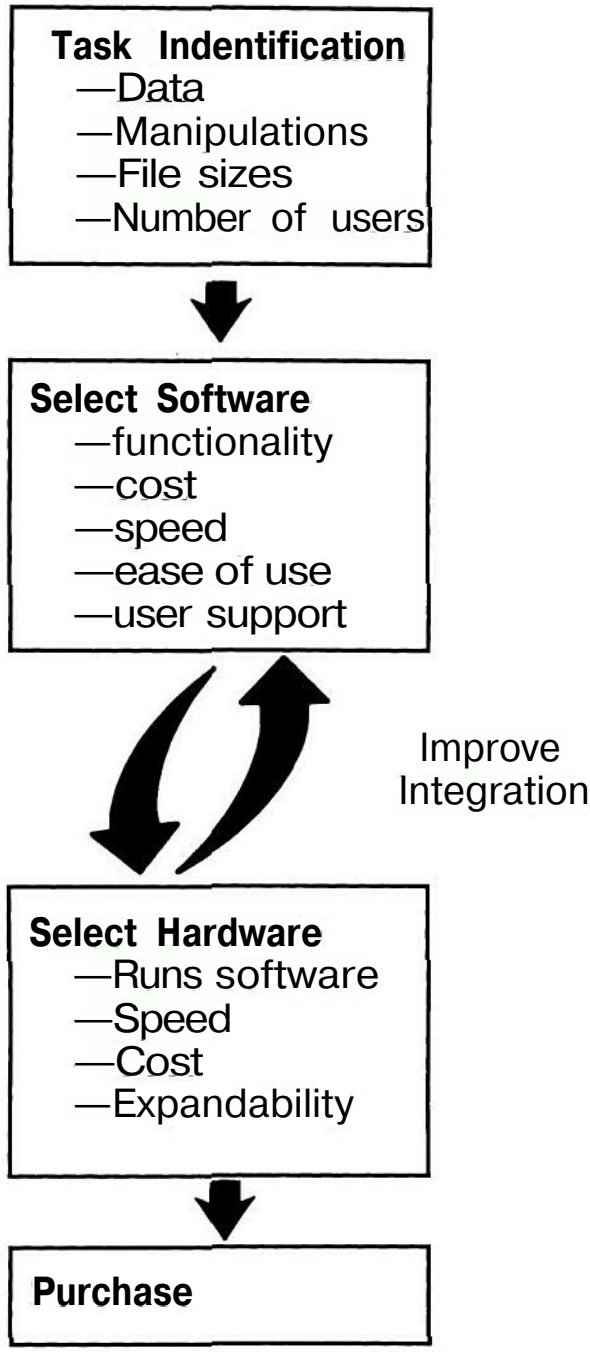


Figure 2. Selecting a computer system.

In selecting peripherals (e.g., digitizing tablets, scanners, video frame grabbers, plotters and film recorders) it is necessary to make sure that they are supported by both the hardware and the software vendor. In some cases the software vendor will bundle a computer and peripherals with the software. However, given educational discounts that are available to the research community (but not to the GIS vendor), it is often less expensive to purchase the computer and the peripherals separately.

NETWORKS

Connection to one or more networks can greatly enhance opportunities for scientific collaboration and help reduce the isolation that field stations often experience.

Local-Area Networks

Local-area networks (LAN's) can facilitate efficient use of computer resources by permitting sharing of data sets, software and computer peripherals, such as printers, disk drives and plotters (Figure 3). In a university environment, LAN'S are often integrated with campus networks that are in turn connected to the wide-area networks.

A LAN can take two forms. In its most basic form, it links individual computers. Using Telnet (a program which allows you to log onto computers across a network) and FTP (a program which allows you to rapidly and accurately transfer files between computers), the LAN can be used as the avenue for accessing multiuser computers on the network or for transferring files between computers at speeds orders of magnitude faster than with a modem. In its more advanced form, "server" computers running networking software are added. This permits direct sharing of peripherals, programs and data in a way that is virtually transparent to the user. Most LAN programs support add-ons for electronic mail and automated backups. Sharing of disk drives across a LAN permits the sharing of data files (subject to security restrictions) and greatly facilitates keeping current backup copies of all data on the network.

LAN'S can be used to eliminate redundancy of software and data at facilities with large numbers of individual computers. A single copy of a database or software product can be used by all the computers on the network, eliminating unnecessary duplication. Use of shared software and databases also reduces time spent installing updated software or modifying databases because only a single copy need be altered.

LAN'S can take a variety of forms, but the most common consists of an Ethernet (a cabling system and electronic protocol capable of 10 Mb/s data transfer rates) running one or more types of networking software (e.g., NFS, 3Com, Appletalk, Novell, or TOPS). Network software for personal computers is usually designed so that a user (although not the network administrator) need know almost nothing about how a network operates. He or she simply operates as though using his or her own stand-alone computer, but with the benefits of larger disk capacities, better backups and a larger variety of peripherals accessible through the LAN. An additional advantage of microcomputer LAN software is that it typically supports add-ons that make checking electronic mail as easy as turning on a computer.

Although some LAN software is specific to particular types of computers and operating systems, other software supports many different types of computers. This capability can be used to fully integrate different data management activities across computers. For example, on a network running TOPS (Transcendental Operating system) netorking software, Macintosh, IBM-PC and UNIX computers can share data files regardless of which machine the data actually reside on.

Wide-Area Networks

In the past decade, the availability of personal computers has put data processing power on the desk tops and in the briefcases of many researchers. The availability of affordable data processing capability has led to a decentralization of research-related data processing. At the same time, there has been extensive growth of wide-area electronic networks that connect computers on a national and international scale.

Wide-area computer networks exist in a variety of forms with many different capabilities, including:

- Easy to access, reliable, and fast electronic mail
- Rapid and reliable transfer of text and graphics (e.g. proposals, manuscripts)
- Rapid and reliable long-distance data transfer
- Archival storage of data on distant university computers
- Better access to researchers at other institutions
- Access to mainframe computers
- Access to supercomputers
- Access to files, programs, printers and similar resources on other networks
- Access to national information and software repositories
- Access to mailing lists and mail forwarding systems

The most widely used network that supports all the functions listed above is the Internet (an association of high-speed, high-capacity, wide-area networks, including NSFnet, a network established and funded by the National Science Foundation). According to the pre-workshop survey, 22 percent of the stations who responded to the survey presently have access to the Internet. Connections to the Internet can take two basic forms. In the first form, a local area network and its computers are linked to a node on the Internet via a high-speed telephone connection. Such a link fully supports high-speed file transfers (depending on the number of network links traversed and amount of message traffic transfer speeds can range from 1,000 to 20,000 characters per second). In its second

File and Print Servers for the LAN

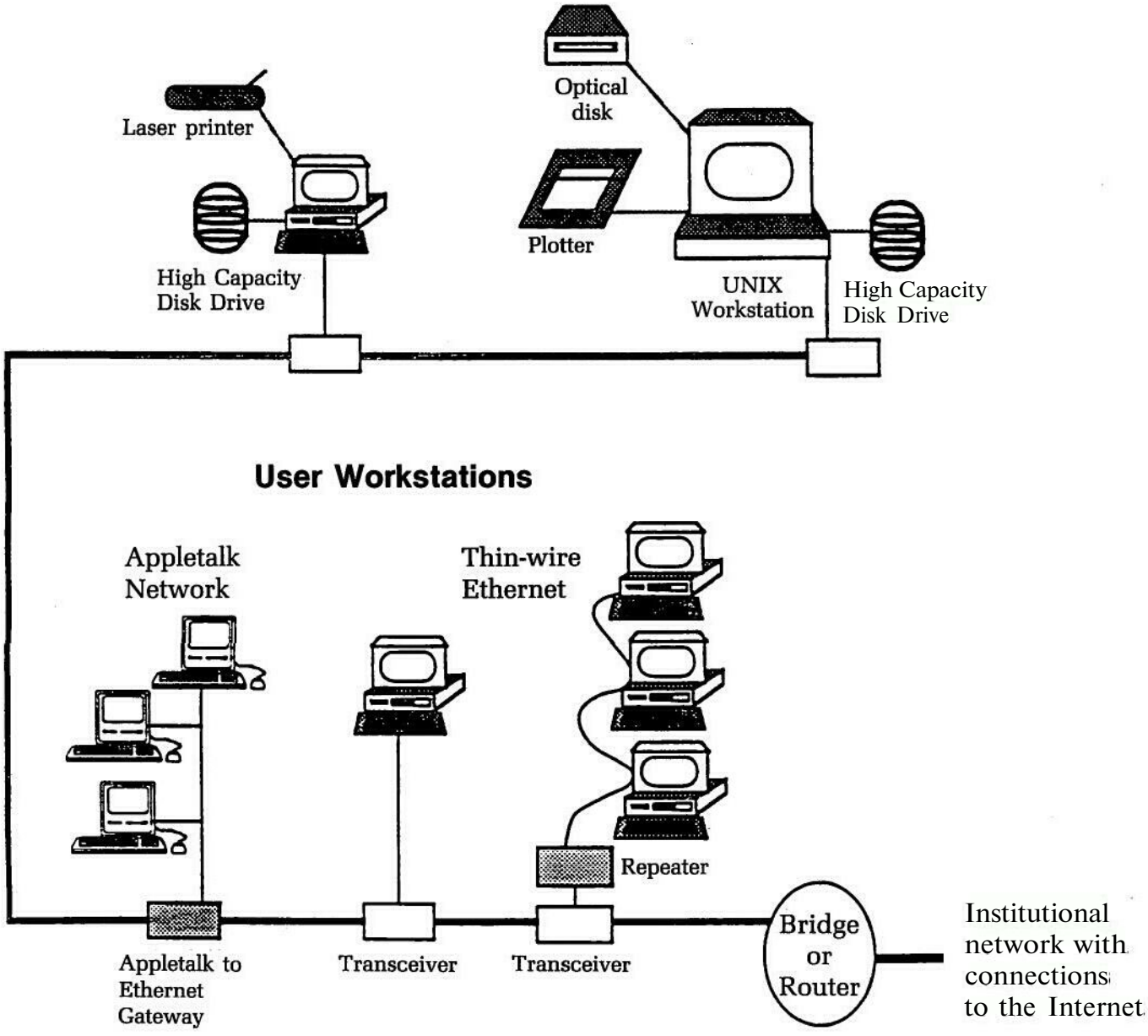


Figure 3. Typical LAN configuration

form, a modem is used to connect to a computer that is in turn attached to a LAN on the Internet (Figure 4). The speed of transfers is limited to the throughput capacity of the modem connection.

Commercial networks support subsets of the capabilities listed above. Typically these include electronic mail, access to bulletin boards and (limited) file transfer capabilities. Connections are made via modem and thus are limited in speed to the capacity of the modem. Thirty-eight percent of the field stations and marine labs surveyed support access to various electronic mail services (such as Bitnet, Omnet, and MCI mail). Unfortunately, sending mail between different electronic mail services is often difficult (electronic mail addresses become long and cumbersome) and occasionally impossible. Forty percent of the stations surveyed have no access to any kind of wide-area electronic network.

Establishing and maintaining network access entails installation costs, recurring costs (for operation and maintenance), and personnel time. Each of these costs varies widely depending on the network chosen and the location and facilities of the field station or marine lab. For example, installation costs for an Omnet account for an existing microcomputer might cost only $300 (for modem, communications software, and Omnet fee), while installation of a full Internet connection might cost $30,000 or more (for routing computers, cabling, network software, and network charges). An Omnet account is easily managed, whereas maintenance of a full Internet connection requires substantial time on the part of a networking expert.

A major concern for field stations is the recurring costs. The phone charge is a large part of such costs, because in many cases the remote location of the field station requires a long distance phone call or a dedicated phone line. Table 3 outlines approximate recurring costs for various network connections, each providing different levels of service. (Actual costs will be site-specific.)

The majority of the stations that do not have access to a full Internet connection through an existing institutional affiliation will find the cost of establishing their own full Internet connection outside the range of their budgets (particularly with regard to the high recurring costs). Provided that phone service is available, those stations can acquire an electronic mail box with one of the commercial mail services, the most common of which are Telemail, Omnet, MCI and CompuServe.

The process of deciding to which network a field station or marine laboratory should be connected must be guided by consideration of the needed capabilities and the cost of providing them. Identification of user needs is a critical first step. It does little good to provide a mail connection to one network when the majority of potential electronic mail correspondents are on another network. Similarly, a network that supports only limited data transfer capabilities is of little use when large data sets are to be transferred. In selecting an electronic mail system it is important to find out what "gateways" exist for transferring messages to other networks and how difficult they are to use. The next step is to compare capabilities of individual networks to user requirements. This will yield one or more candidate networks for which cost analyses may be performed. A final network may then be selected.

The utility of electronic mail to the field biological community could be significantly enhanced by access to a mail forwarding system for field stations similar to the one recently implemented by the LTER network. This system solves several practical problems by:

- Creating simple, uniform network addresses for all users
- Sending and receiving group mailings
- Disseminating information on request by automatic reply
- Routing mail between different networks (acting as a mail gateway)
- Integrating mail and bulletin board services

This sort of service could perhaps be provided by LTER for a wider set of field stations, given appropriate funding.

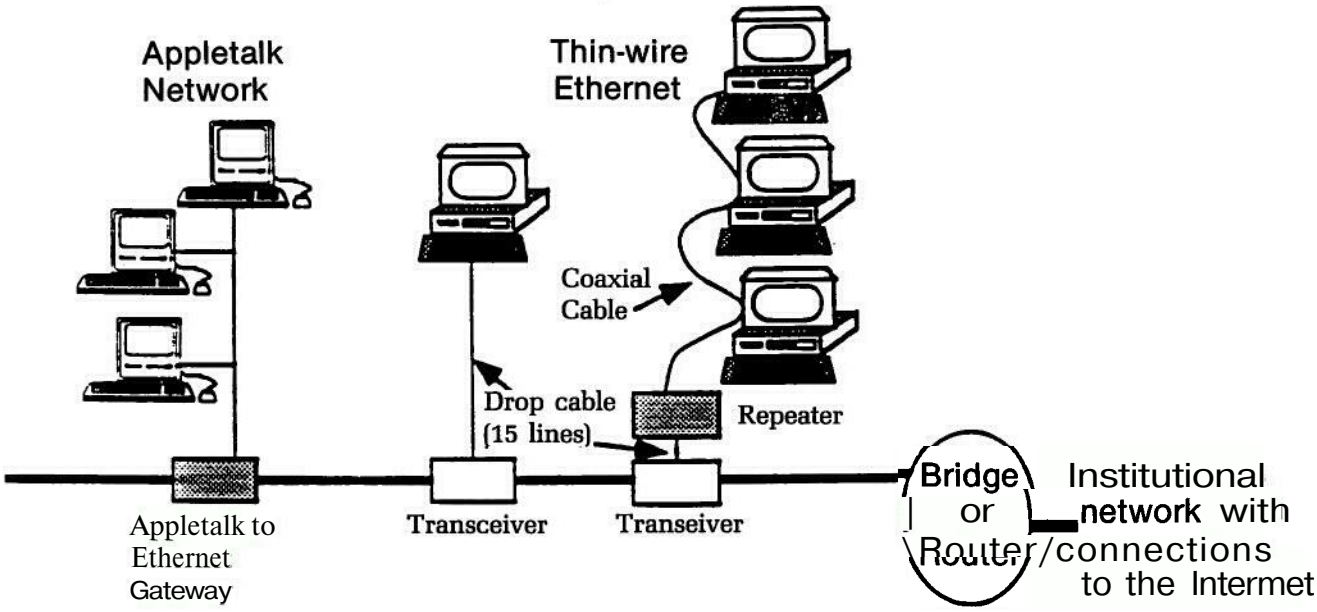
Archival Storage

A major objective of data management at field stations and marine labs is the archival storage of data. Data sets are prone to a variety of mishaps that can result in damage or loss. Data stored in printed form can deteriorate if exposed to excess moisture or heat, or spontaneously deteriorate if they are recorded on paper with a high acid content. Data stored on magnetic media can be lost because of equipment failures, power surges, extreme temperature fluctuations or simple deterioration of media over time. It is important to note that the "standard" magnetic tape or floppy disk has a recommended lifetime of only five years. It is said that there are no valuable data stored on 25 year old tapes simply because there are no readable data on 25 year old tapes.

A crucial component of archival data storage is making sure that backup copies of data are maintained. These should be kept current and stored in a location physically separated from the original data so that location specific calamities (e.g., floods, fires, hurricanes) are unlikely to damage all copies of the data. Off-site backups of data are facilitated by access to computer networks. By using transfers across a network, data can be backed up on another computer or device at a distant location without needing to transport physical media.

Data may also be lost through technical obsolescence. Optical storage devices are capable of retaining recorded data for many decades. Although optical storage reduces some of the problems associated with deterioration of media, access to data

A) Direct connections to LAN (ethernet)



B) Indirect connection to LAN via mainframe

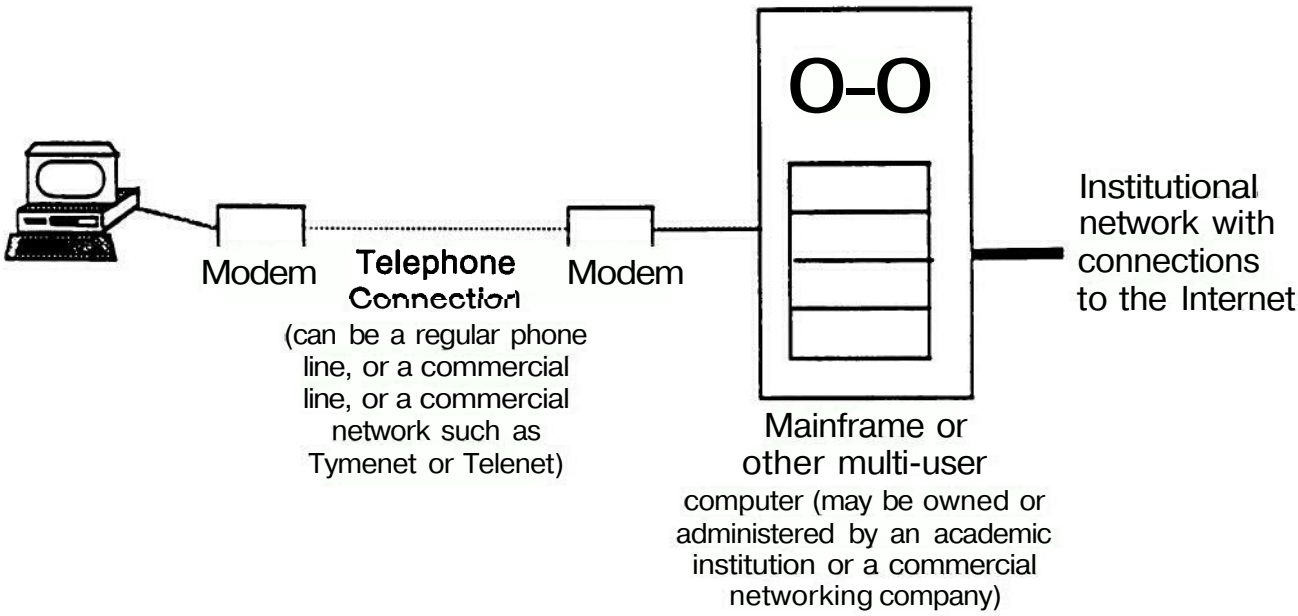


Figure 4. Connecting personal computers to the Internet.

Table 3: Approximate costs for network connections broken down by functionality. Units of cost are $/s/y = dollars per site per year and $/u/y = dollars per user per year.

Electronic mail only

Type of Connection	Recurring Cost	Cost of installing Network connection	Cost of on-site Equipment
Commercial (Telemail, Omnet, MCI)	600/u/y	(Phone connection)	2,000 (PC + modem)
Institutional (through a university, etc., Bitnet*, Internet mail, ...)		(Phone connection)	2,000 (PC + modem)

(*) Bitnet
Bitnet "membership" is no longer free. Organizations can join Bitnet for a fixed annual fee ($750 - 10,000, depending on the size of the institution). This fee is usually passed through to individual users in the form of administrative costs (such as overhead costs). Bitnet is now operated together with CSNET by an organization called CREN.

Full Internet connections (electronic mail, file transfers, remote login capabilities), do-it-yourself (no network administration services provided)

Type of Connection	Recurring Cost (*)	Cost of installing Network connection	Cost of on-site Equipment
Internet direct	8,400/s/y	30,000	10,000 direct
Internet dial-up IP (SLIP)**	5,000/s/y	10,000	10,000

(*) recurring costs exclude the cost of on-site personnel

(**) IP refers to the Internet Protocols used to transfer data. SLIP is Serial-Line Internet Protocol and is a subset of IP that can operate over low-speed **connections**

may be lost due to rapid technological changes that render storage media obsolete and unreadable long before the end of its service life. Optical disk systems depend not only on the disk itself, but also on the disk drive which has a much shorter service life. Without a suitable drive to read it, a disk is useless even though it still retains the data.

Specialized data formats can also result in a loss of data through obsolescence, but of software rather than of hardware. Data stored in format that is readable only by a single software package can be lost if that package becomes unavailable. One way to avoid this problem is to store archival data in a simple standard format such as ASCII (American Standard Code for Information Interchange). Although this solution works well for numerical and character data, is not adequate to protect binary data, such as images. This is because binary data typically are stored in any one of a number of specialized formats. For such data it is crucial that documentation on the format be kept with the data.

Technological Innovations

Computers and software are among the most rapidly changing technologies. Innovations having direct impacts on data management include automated data capture technologies, computer networks, improved data storage media, portable data entry systems and improved operating systems and software.

Automated data capture includes the use of optical scanners, image processing techniques, satellites and automated data loggers. Global positioning systems (GPS), which use radio signals from satellites to accurately calculate their current position, can also be used to automatically enter locational information.

Both local and wide-area computer networks will continue to increase in speed and can be used to integrate different types of computers into a single system. Additionally, groupware (sometimes taken to mean software which allows multiple individuals to see and edit the same data simultaneously) can facilitate collaborative projects over networked computers.

Improved data storage media take the forms of optical disks and digital audio tapes (DAT). Optical disks come in a variety of sizes and capacities, all of them capable of storing hundreds of millions of characters. DAT permits the storage of gigabytes of data on small magnetic cassettes and is extremely useful for making backup copies.

Portable data entry systems, in which a researcher types data directly into a small portable computer in the field, are increasingly popular and, when properly programmed, can help to reduce data errors by alerting researchers to apparent problems while they are still in a position to correct them.

Operating systems for microcomputers are becoming increasingly sophisticated and are converging with those on mini- and mainframe computers. For the user, this will result in an increasingly "transparent" computing environment where it will not be possible to tell what type of computer is being used. Reusable computer program modules generated by object oriented languages can help to simplify data management tasks that require specialized attention. Despite significant increases in capabilities, software packages are growing increasingly easy to use. The trend towards sophisticated (yet easy to use) graphical user interfaces is playing an important role in this process.

Keeping pace with technology requires the obvious investment in hardware, but more importantly, it requires a significant commitment to personnel, planning, and training. The speed with which the technology used for data management evolves makes it difficult for individual data managers to keep abreast of potentially important developments. These difficulties can be ameliorated by enhanced communication among data managers. Several opportunities exist for facilitating exchanges on technology-related issues. A periodic newsletter addressing data management issues at field stations would disseminate information. An electronic bulletin board or electronic mail group lists could serve as a forum for exchanges of information between data managers and would permit a more rapid response to questions.

Facilities for Visiting Researchers

The ultimate purpose of research data management is to facilitate and improve research. For a data management system to be successful, it must be used by researchers. Field stations exist in a variety of settings and circumstances and with a diversity of missions. Computing equipment and services are as varied as are locations, and a visitor does not always have free access to facilities. Some stations provide no common-use software or hardware, whereas other stations provide aid in every facet of research activity, from data entry to data analysis.

In some cases computers and computer access by visitors to resident data bases are critical to the success of scientific investigations. Many stations have data bases that represent the only historical information available. In the absence of replicates, these data are the only way to validate many models. Expensive duplication of previous work can be avoided by identifying and using extant work.

The utility of a station or laboratory environment to visiting researchers can be enhanced by:

1. A common pool of hardware and software of a type that is currently in wide use (e.g., Word Perfect or SAS in the DOS environment), or that is very easy to learn (e.g., MacWrite or Delta Graph in the Macintosh environment).

2. A variety of materials to orient researchers to database facilities. Person-to-person interaction is the preferred mode for starting the educational

process. Electronic, video and audio material can be made available to help visitors learn more on their own in a self-paced mode. Teaching and demonstration programs that come with software are useful for this purpose. Short interactive tutorials can also be produced locally.

3. Access to electronic mail. Electronic mail is useful for both administrative and research purposes. It is widely used for pre-arrival arrangements, for access to data bases at other facilities, and to keep in touch with colleagues.

In contrast to the campus data management environment, seasonal field station users have a relatively short period of time in which to enter data. Where staffing and circumstances permit, a discussion of the proposed work between the researcher and the data manager can result in a data catalog entry that anticipates the integration of the data into the site database managed by the station. Assistance can extend to suggested data entry forms, quality assurance, portable computers and even appropriate analytic procedures.

CHAPTER IV—SUMMARY OF THE WORKSHOPSURVEYQUESTIONNAIRE AND PRE-WORKSHOP DEMONSTRATIONS

John B. Gorentz
W.K. Kellogg Biological Station
Michigan State University

and

Michael P. Hamilton
James San Jacinto Mountain Reserve
University of California, Davis

INTRODUCTION

Several months before the workshop, in November 1989, the workshop planners sent a survey questionnaire (Appendix F) to about 200 inland field stations and coastal marine laboratories. The purpose was to assess the state of data management at field stations and gather information to use for selecting representative sites to invite to the workshop.

On April 22, 1990, in a pre-workshop symposium, a series of demonstrations was presented in order to inform workshop attendees of some of the data management systems and technology in current use.

Purpose of the Questionnaire

The sponsors of the workshop (the Organization of Biological Field Stations and the Southern Association of Marine Laboratories), although having many common interests, are a diverse group with a correspondingly diverse array of data management systems. Their activities range from seasonal summer school sessions to year-round programs in research and education with large resident faculties and staffs. Their computer systems range from the personal computers of few individual investigators to networked systems and large mainframes. Their data management systems range from non-existent to just getting started to large systems with a separate staff and budget.

It was important to have representatives of the various types of field stations and their data management systems at the workshop, so some of the survey questions were designed to elicit information on relevant site characteristics. A balance was sought between experienced participants and those whose data management systems were in early stages or limited by modest resources.

In assessing the state of data management, we were less interested in examining the technology in use than in discovering what kinds of data sites have seen fit to manage, and how these data are being managed. Assuming that the best recommendations the workshop could produce would facilitate goals and objectives already adopted by these sites, we asked questions designed to find out what was important to researchers. We also wanted to learn about common concerns and problems.

Background Issues

Although there is widespread agreement on many data management issues, perhaps much more so now than at the time of the 1982 workshop, there are also unresolved issues which influenced our choice of questions, and our interpretation of the responses. These background issues relate to the best use of limited resources, acceptable degrees of centralization, and the relative importance of technology vs. human resources.

Use of the term "data management" is usually accompanied by some unstated presuppositions. For some people, data management is whatever must be done with data, usually using computers, in order to analyze them for publication. Another view, perhaps less common now than at the time of the 1982 workshop, is that data management includes almost anything that has to do with computers and technology. For still others, data management means caring for certain data so that, whatever their original purpose, they are preserved and made available for more general use, now or in the future. This latter view was the premise of the 1990 workshop.

This workshop was based on the assumption that at field stations and marine laboratories there are historical data records worth preserving to enhance the value of the habitats for research, to provide background data, and to make long term studies possible. Some of these data sets are gathered for general use, others are the fortuitous by-product of specific research.

Without proper care, these data resources will be lost. This care entails a cost, and although there is widespread agreement that efforts to preserve data are worthwhile, there is not universal agreement that already scarce resources should be spent on data management.

Nor is it certain that the data sets compiled or otherwise preserved for general use have been used, or will be used, to advance science. Science builds upon previous work, including that represented in previous databases, and scientists have a responsibility to preserve data for those who will follow after them. There is, however, some disagreement over whether resources should be spent testing hypotheses rather than on preserving data without a clear hypothesis to be tested.

Even those who maintain that data need to be managed as a research resource will acknowledge that historical data are not yet being utilized as they could be. There are several possible barriers limiting the availability and use of existing data:

- The existence of these data sets is not commonly known. The scientific literature may serve as a partial index, but additional means are needed to make these data sets known.
- Physical access to data is difficult. Better means of electronic communication would make data sets more widely used.
- Some data sets are known to exist, and are accessible, but are too poorly documented to be useful. Better systems of documentation are needed.
- Data sets are sometimes not worth bringing together for comparison because they are too dissimilar in format, representation of data, methods, and meaning. Standards are sometimes proposed to resolve these problems, but there may be resistance to standards as being too restrictive for open-ended inquiry.

These barriers are not mutually exclusive, but disagreement as to their relative importance leads to disagreement over funding priorities.

Many data management solutions assume a certain amount of centralization. Long term care of data often implies a responsibility for data that goes beyond that of an individual investigator, and this in turn implies a degree of relinquishing control which is sometimes in conflict with the basic tradition of independent inquiry.

Perhaps most controversial is the issue of standards. A lack of standards makes comparative analysis of data sets from different sources very difficult. Researchers currently use and benefit from standards at many levels (e.g. ISO units of measurement). But there is a fear that any push toward standardized computer systems or data formats at any level will be too restrictive or too unwieldy and will interfere with research.

Finally, there is a question as to whether the greatest data management need at this time is for technology or for human resources. It is often easier to obtain funding for computers and software than for personnel, but it is possible that at the current stage of development, personnel are the limiting factor.

SURVEY METHODS

In November, 1989, questionnaires were mailed to about 200 sites that constituted the membership of the Organization of Biological Field Stations and the Southern Association of Marine Laboratories. Questionnaires were also sent to additional sites in the Long Term Ecological Research program, and to several National Marine Laboratories.

103 responses were received. Several of them were received too late to be used in selecting invitees to the workshop, but those responses are included in the results presented here.

In the questions, we did not ask about facilities and technology so much as about goals, priorities, and personnel. We tried to get the respondents to distinguish between institutional operations and those of individual research programs. Some respondents were more sensitive to the distinction than others.

The questions were open ended, because we felt that the most useful information would not necessarily fit into neat categories. In analyzing the responses, we did attempt to categorize the responses, and in the process made subjective judgments. Even though some information is presented quantitatively, the tallies were a matter of considerable interpretation on our part.

WHAT DATA ARE BEING MANAGED?

The first set of questions (Question 3a-3c) was designed to find out about the data that sites are managing as a general resource, or which could be made more available if resources permitted.

In doing this, we wanted to distinguish between those databases managed as part of a single research project, and those that are being managed for long term general use as a site responsibility. We also wanted the respondents to take a broad view of the term "database," including those data managed without sophisticated database tools or without computers, as well as non-traditional forms of data such as audio and video recordings.

The hope was that the responses to these questions would help define the subject-matter of the survey and workshop and give some idea of sites' goals and priorities.

We asked three questions about databases.

3a. Does your site have databases that have been compiled specifically for general use (e.g. species lists, meteorological data)? If so, please list some examples.

3b. Does your site have databases originating in individual research programs, that are or could be developed into general use resources. If so, give a few examples.

3c. Does your site have computerized records consisting of non-traditional forms of data, e.g. acoustic records, maps, visual images.

In response to question 3a, 90 percent of the respondents listed one or more databases; only 10 sites said they did not manage any general use databases at all. We categorized the responses into a few arbitrary, non-discreet, non-orthogonal groups, which we tallied as follows:

Climate data: A large, clear cut category was climate databases, with a little over half of the 103 respondents listing this among their general use databases. These databases appear to include everything from records kept on paper, to automated data collection systems and electronically networked databases.

Species lists: Almost as many sites listed one or more types of species list among their general use databases. They were variously described as species lists, species inventories, and species checklists. They were usually compiled for specific taxonomic groups, such as birds, mammals, vascular plants. Two sites pointed out that they have developed taxonomic keys for their lists. One site has made its species lists a key part of its data management system, by setting up a system of standard species codes to be used in data sets.

Hydrography and hydrology: Twenty-seven sites listed one or more types of hydrography or hydrology database. This category includes databases variously described as hydrological measurements, tide measurements, stream flow, water quality, water level data, bathymetry, sedimentology, physical and chemical limnology, groundwater levels, pond levels, sea levels, hydrology, stage/discharge data, seawater temperature, salinity, stream chemistry. (We did not include precipitation databases in this category.) These databases range in scope from (for example) a modest database of stream level measurements, to a large scale information center for Galveston Bay.

Bibliographies and project lists: Sixteen sites listed some sort of bibliography or project list among their general use databases. These appear to range from simple lists to elaborately indexed computer databases. Some of them appear to be stand-alone databases. Others appear to be integrated into a larger system, serving as an index or access point to the site's other data resources. That is, a data user can search the database for a particular subject or organism, and find not only the pertinent literature, but be directed to other databases as well. Four respondents described the contents and organization in more detail. Their databases are sorted or indexed by one or more of the following categories of information: author, location, research topic, funding source, species, sampling dates, and keywords.

Miscellaneous long term monitoring: Almost all of the databases listed by the respondents could be categorized as containing long term monitoring records. But we also noted some miscellaneous other types of long term databases, listed by 18 sites, which do not fit the above categories. They include records on flowering phenology, secondary succession, fish capture, sequences of plant surveys, vegetation on permanent plots, annual bird counts, bird migration, nesting, land use history, photo monitoring, and fire history. According to the responses to question 3c, a few sites keep databases of 35 mm slide photographs taken on a calendar schedule.

Maps and geographic data: Because of the current high level of interest in geographic information systems, we created a separate category for geographic data and map-type data. By including some responses to question 3c about non-traditional forms of data, we counted 32 sites that either have GIS systems, or have map-type databases now represented on hardcopy maps and aerial photos that could utilize GIS software or other spatial data management systems. It could be argued that this category is the largest of all, if one considers that all data that reference particular spatial locations on the earth's surface are potential GIS data. Most of the general interest data at field stations and marine labs fit this description.

In response to question 3b, which asked whether there are other data that could potentially be managed as a general resource, 76 respondents said yes, and 74 gave examples. These examples consisted of additional long term records of the types listed above, as well as point-in-time data sets. These include those resulting from (for example) three year projects, but are distinguished from continuous long term projects.

Based on these responses, it can be seen that most field stations and marine labs are in the business of data management. Even among those 10 sites that said they do not manage any general use databases, some plan to do so soon. These include field stations which are relatively new, or which have only recently adopted any data management goals.

However, among the ten are sites that have no intention of managing general use databases. These sites responded that their databases were all investigator-specific, and/or they do not think managing general use databases is an appropriate undertaking for their sites. In fact, some questioned the validity site-sponsored data management. This issue is discussed further under "Site Self-Evaluations and Recommendations."

It could be argued that our tallies under-represent the number of sites managing general use databases and the number of databases. The questionnaire asked only for examples, not a complete list. Although we intended the term database to be used in a general sense, not just applying to those data being managed in some formal DBMS, it is possible that some sites did not consider their more casual,

people can build on the work and data of those who have gone before them. Scientists are often engaged in breaking down technical and political barriers that limit collaboration with others. But scientists also guard their data in order to ensure proper recognition of their work through the publication process.

Many field stations are involved in data management on the assumption that data sharing through the traditional system of publications is not adequate, and that there are unpublished data, never-to-be-published data, and raw data behind publications that need to be made available as a resource for others. They therefore need to reconcile legitimate proprietary rights with the goal of greater accessibility.

In question 5b we asked,

"How do you weigh investigators' proprietary rights to data against the goal of wider availability? Is there security against unauthorized use of data?"

Ninety-four of the 103 respondents addressed the question. The responses represented two fundamentally different attitudes. Many sites view proprietary rights as a necessary evil, while others perceive the protection of proprietary rights as an important objective. This was perhaps expressed most strongly by a site which reported, "Proprietary rights are protected to the fullest."

Twenty sites reported that proprietary rights are not an issue, or at least are not yet. Some reasons given were that proprietary data are not involved, or that there is not a centralized system. Four reported that they have no policies yet, but that it is an issue that needs addressing and is being addressed.

Thirty sites reported that they have no policy, that the issue is left to the investigator, and that the data can be accessed only through the investigator anyway. One of these respondents simply said, "Data is uninterpretable to non-investigators." This is probably a common state of affairs. Some reported that data are indeed shared by these bilateral arrangements. Six sites had systems of central access to data, but left the issue of outside access up to the investigator. In some cases the investigators exercise control by deciding whether or not their data are to be added to the central database. At others sites they exercise control over data residing in a central database through a security and authorization mechanism.

Eleven sites reported some sort of policy to limit proprietary rights, but did not have a centralized database. Most commonly the policy consisted of a limit on the time during which investigators have complete control of data; after this time they are required to make their data more accessible. These policies were usually related to site-use requirements and responsibilities for visiting researchers. It would be interesting to know how effectively these policies are enforced, or whether any enforcement mechanisms are necessary.

Seven sites had policies in place that emphasized security and confidentiality, while complying with any regulations regarding open access. These tended to be some of the larger marine labs with highly centralized systems, at Which researchers' rights are subservient to other purposes. These sites reflected a strong sense of ownership of and responsibility for data, with policies in place and mechanisms to enforce them.

Three sites emphasized the protection, rather than limitation, of proprietary rights.

Thirteen sites emphasized central, general purpose databases that are open to all. However, the data they contained may not have included much that was investigator-specific.

ADMINISTRATION AND PERSONNEL

Administrative and personnel factors are possibly more limiting to progress in data management than are technology and equipment. We wanted to explore the magnitude of data management tasks by determining the level of committment to data management, including personnel resources committed.

We also wanted to learn the degree to which a field station's data management is a distinct activity, distinguished from related areas such as computer management or investigator-specific data management. We assumed that clear data management goals would be reflected in distinct data management budgets and personnel assignments.

We asked the following four questions:

4a. Where does the impetus for data management arise (e.g. site administrators, interested faculty members, research programs, technical staff)?

4b. Does your site have a data manager, or other person(s) with designated responsibility for data management?

4c. What personnel are involved in data management (number of persons, positions, training, experience, fraction of time)?

4d. How is data management funded? Is there a specific budget for data management? Is it funded at the site/institution level, or on individual grants?

Impetus for Data Management

We asked the first question, about who is pushing data management, to detrmine the extent to which it is research driven or technology driven. We also wanted to learn whether there was top-level administrative commitment.

We grouped the responses into categories and tallied them as follows: Researchers (68 sites), Administration (61), Technical staff (21), Long Term Ecological Research Program—LTER (9), and External (5). Two sites did not respond to this question. Many sites fit into more than one of these categories.

It does appear from the responses that data management is largely research driven. Two thirds of the sites cited researchers as the driving force, and almost as many said that administrators, who presumably have research interests foremost, were the impetus. Of course, a researcher or administrator can be overly enamored of technology for its own sake, but presumably most have not fallen into that trap. Of the 21 sites listing the technical staff as a driving force, only one listed it as the only group leading data management, and sixteen of those 21 also listed researchers.

The responses are probably a good sign, indicating that sites have their priorities in order. Research is driving data management, rather than vice versa. Data management has a supporting role, albeit an important one. As such it is not likely to take on a life of its own, unresponsive to research needs.

The number of sites listing site administrators as a driving force indicates that at a majority of sites, there is active support at the top level, and not just passive tolerance.

The five sites that indicated an external impetus were mostly government labs whose supervising agencies mandate their data management activities. The nine sites that cited the LTER program also are responding to an external impetus.

Designated Data Manager

It may be that people rather than technology are the limiting factor in successful data management. But money to fund personnel is often harder to come by than money to purchase equipment. Before making recommendations on personnel for data management, we needed a clear picture of the current personnel situation.

In response to question 4b, about a designated data manager, 41 sites reported having none. Twenty-three have a full time data manager. Thirty-six have someone doing data management part-time, including six sites at which the director is the data manager, four at which the data manager is the librarian, and two at which a laboratory manager performs this function. Two sites were unclear in their responses and are not included in the tallies.

Several of the 23 sites with full-time data managers reported that other computer-related duties besides data management are included in the manager's workload. If the issue is data management in a strict sense, the count of 23 full-time data managers is misleadingly high. Even many of the 36 part-time data managers also do computer management as well, with data management getting a fraction of the person's attention.

The six sites with site administrators serving as data managers are generally small sites with appropriately modest goals. The four sites with a librarian-data manager suggest a route for sites to follow when it is not desirable or necessary to develop a computer and data management infrastructure; data management can be made an adjunct to library rather than computer operations.

Staff Qualifications and Background

In response to question 4c about the number of persons involved in data management and their backgrounds, one site stated, "too irregular to tabulate." This telling comment is a good summary of the overall situation. Even though most respondents did attempt to provide numbers and descriptions of those in data management, the responses taken as a whole were too irregular for us to tabulate.

This is partly because of a confusion between data management and computer management. The two types of work are often confounded, and even where distinct, are often done by the same persons. Given this situation, we could not tell which qualifications listed by the respondents were relevant to data management.

Another barrier to tabulation was the lack of comparable functions for the data management portion of the work. Combinations of staff, duties, organization, and infrastructure varied greatly. Data management is done by site administrators, faculty members, graduate students, secretaries, statisticians, librarians, and sometimes even by specially designated data managers. Various "coordinator" positions (e.g. research coordinator, scientific coordinator, site coordinator, data coordinator) have responsibility for data management among their duties. Educational levels of data managers range from high school degrees to the Ph.D. level, with many in between.

Data management is commonly done by people who are self-taught. A few data managers have backgrounds in computer science, but data managers with backgrounds and training specifically in data management are perhaps non-existent. Those whose training is primarily in computer science are uncommon.

It is not possible to determine from the responses which personnel configurations are the most successful.

Table 7. Cross-tabulation of responses to Question 4D. Responses regarding a specific budget for data management are arranged horizontally, and those regarding funding at the site/institutional level are arranged vertically.

		Specific Budget?		
		No	Yes	Total
Funding at site/institutional level?	No	48	5	53
	Yes	37	12	49
	Total	85	17	102

Data Management Funding

We asked question 4d about funding to evaluate sites' commitment to data management. We wanted to know whether data management per se is an objective distinct enough to have it own budget (whether it gets at least some funding at the institutional level, rather than exclusively from individual grants), so we could judge whether it is getting support from the top institutional level.

Of 102 sites responding to this question, only 17 had a specific data management budget. However, nearly half (49 of 102) did have at least some funding from their institution. The responses on these two issues of a specific budget and of institutional support are cross-tabulated in Table 7.But even among the 17 sites with a specific data management budget, in many cases the budget appears to be more of a computer budget than a data management budget.

Similarly, the level of institutional support is probably not as high as it might seem from the raw numbers. In some cases, the amount of support is small, as small as a bit of funding for a weather station. The fact that only one fourth of the sites with some institutional support have a specific budget is an indication that such support does not involve serious money.

SITE SELF-EVALUATIONS AND RECOMMENDATIONS

To sum up, we asked sites to evaluate their accomplishments, resources, and needs. We asked a series of five questions, starting with:

8a. What have been your most important data management accomplishments?

The data management accomplishments that the respondents listed fell mostly into three categories: data, computer systems, and administration.

The data-oriented accomplishments included: assembling historical data sets (cited by 7 sites); setting up continuous, long-term databases (3 sites); other computerization of large databases, with emphasis more on the data than on computerization (5); establishing systems of baseline, site-characterization, or geographic data (16); the development of site bibliographies (7); and specimen databases (3 sites).

The administrative accomplishments included: coming to grips with the need and identifying the problem (cited by 4 sites); developing an overall plan (2 sites); getting started (2); getting organized (8); establishing policies regarding the responsibilities of investigators (3); compiling data catalogs and indexes (14); a methods manual (1); establishing data archives (2); establishing standards for data entry, documentation, and format (8); establishing data quality protocols (3); development of a data management staff (4); obtaining high level support for data management (1); getting funds (5); and establishing a training program (1 site).

The computer-oriented accomplishments included: implementing database management and geographic information system software (cited by 6 sites); developing database management software (2 sites); and data entry systems (2). Five sites developed computerized databases, with an emphasis more on the computer systems than on the data. New or improved computer systems, including storage systems and networks, were cited by 13 sites. While three of these sites decentralization their computer systems, moving from mainframes to microcomputers, one site centralized its database system. Five sites installed instrumentation for automated data acquisition.

8b. What things would you now do differently, if you had them to do over? What suggestions would you give to other sites?

Not all sites responded to the above question, and some of those who did stated that they were not far enough along to answer it. But those who responded listed the following types of items:

- Take time to plan, instead of just letting things happen.
- Implement policies regarding researchers' responsibilities.

- Make sure of researchers' support, and involve them in oversight.
- Get organized sooner; catching up is hard to do.
- Start baseline data collection sooner.
- Link all data sets by location.
- Do quality control.
- Set and enforce standards to ensure consistency; set standards earlier in the process.
- Keep it simple; do not try to do everything at once; do one data set at a time.
- Spend more time on documentation of everything.
- Consult with outside experts.
- Provide training.
- Avoid mainframes.
- Use networks to keep decentralization from going too far.
- Buy commercial database software rather than developing it in-house.
- Use relational database software technology.

The last three questions of the series asked about resources for data management:

8c. What personnel resources do you think are needed to meet your data management goals? Are these resources available?

Some sites said they had adequate personnel resources. These were mostly sites that apparently had just recently received funding for new positions. Those who stated a need for additional personnel listed everything from data entry personnel to skilled professionals. Many sites with no data manager stated the need for a part-time data manager "dedicated to data management." Some who had part-time data managers emphasized the need for a full-time person. Some that had a full-time person needed more persons.

Although some sites lacked highly-trained personnel with specialized technical skills, more of them pointed to the sheer amount of time rather than skill that was needed to do the work.

A few respondents also emphasized the need for researchers to take part in data management or exercise an oversight role, or to take an interest in the sometimes mundane gathering of baseline data.

8d. What additional facilities crucial to your goals (hardware, software, etc.) are lacking?

The following were listed:

- Data collecting instrumentation, e.g., for data loggers
- Computers and computer equipment
- New or upgraded mainframe computers
- More, better, or upgraded microcomputers dedicated to data management
- Computer systems or upgrades for GIS
- Computer systems and equipment for video analysis
- Database management software
- Personnel
- Bricks and mortar, e.g., office space, physical storage space
- Local area networks, network upgrades, or communications, including links from remote sites to university campuses
- Equipment located at the field sites to reduce the need to use equipment at distant campuses
- Equipment for long-term, reliable archives

8e. Where do you think additional funding is most needed?

This question was intended to elicit the most important priorities among all the items mentioned, The need for personnel topped the list of concerns, with 52 respondents citing it, as opposed to 21 who listed computers and hardware, and 13 who listed software. The raw count understates the strong emphasis that was placed on personnel, as well as on the strong concern, expressed by 13 sites, for the stability of long-term funding for recurring costs for personnel and for the maintenance of computers, software, and data. Other needs were buildings (cited by 1 site), instrumentation (3 sites), computer network links (2), and training (2 sites).

It should also be noted that a few persons stated here and in their additional comments that their top priorities were outside the realm of data management. Some were frankly skeptical about the feasibility of managing data for general use, or the appropriateness of diverting research resources to data management, preferring to focus attention on the immediate needs of individual researchers.

Some of the skeptical comments were as follows:

"To do it right at each lab might have prohibitive costs."

"Given the extremely diverse nature of the research and the individual approach (30-50 basically unrelated research projects/yr)...I have serious questions about the potential utility of centralized data bases."

"...We often wish we had much better base line data, but given our mission, it would be difficult to justify the diversion of resources from other goals."

"...A useful topic for...discussion might be 'How do we maximize the benefit, or judge the eventual benefit, of data we collect now for future use?'"

"...I have yet to see a data mgt. system (for ecological labs) that really worked and was actually used by scientists publishing papers based on the data..."

"What is the purpose? Most of our researchers believe that maintaining long term records without specific research goals is a waste of resources. Once they answer a question their data is useless and just takes up space in a filing cabinet (after publication)."

"...keep it basic...let the researcher who wants the data do all the work."

CONCLUSIONS AND SPECULATIONS

Diversity

Although it might seem almost too obvious to mention, one of the most significant characteristics of field stations and marine laboratories is their diversity. They do have some common interests and objectives, but there is such a myriad of missions, institutional arrangements, facilities, and types of data that great care is needed in developing standards, guidelines, and recommendations for them. It is important to analyze every assumption and conclusion from the perspectives of the full range of sites. Unlike other scientific disciplines in which, for example, the issue is how to deal with huge quantities of satellite imagery, the challenge for field stations is in dealing with the great diversity of data.

Long Term Data

Most of the data that need to be made available for wider use are long term records. Managing these data requires a sustained effort that is not likely to be funded by project-specific grants.

Descriptive Data

Sites' initial efforts at data management are in the area of descriptive data, such as climatological data and species lists, rather than in the area of experimental data. Possibly this is because organized data management is like many scientific disciplines, which need to start with descriptive work before moving on to the experimental. An alternate explanation is that the main purpose of long term data management is to provide descriptive background data which can serve as a context for experimental studies, and that this will always be the focus.

Access to Data

In addition to managing descriptive data, many sites have in the past decade embarked on the development of catalogs and directories to data, sometimes in the form of publication lists. These ventures will not only serve to make data more accessible to others, but help sites take inventory and evaluate priorities.

Dissenting Views

Although there are those who question its value and feasibility, most of the respondents took a positive view of the necessity and possibilities of data management, as indicated by their accomplishments, plans, and commitments. But it would be good for those who are committed to data management to keep the skeptics' comments in mind, because they lay bare the criteria by which data management should and will be evaluated.

Commitment to Data Management

Using the number of sites managing general use databases and those developing access mechanisms as a measure, it would appear there is great enthusiasm for data management. But judging from personnel, budgets, and other comments, data management might seem an indistinct activity, commonly confounded with computer management and short term exigencies. However, the situation has greatly improved since the time of the 1982 workshop. The survey results show a much greater agreement and understanding of the possibilities and needs than would have been found earlier.

PRE-WORKSHOP DEMONSTRATIONS AND PRESENTATIONS

By way of information and introduction to the major concept of the workshop, a day long pre-session symposium was held on April 22, 1990, highlighting examples and demonstrations of data management systems by 20 of the workshop's participants. Ten demonstrations of laptop, PC and Macintosh-based systems were presented, and discussions of other station-based capabilities were described.

Demonstrations had been pre-selected to provide a sample of diverse approaches in use at marine and inland field stations as of April 1990. Examples of the following categories of data management were demonstrated:

- field entry of data using portable computers
- automated acquisition of environmental data
- geographic information systems (microcomputers and workstations)
- relational database management for research project management
- microcomputer access to large SQL relational database
- hypertext (HyperCard) and interaction multimedia databases

- multimedia networking over Internet
- access to databases over networks using electronic mail

The participants listed below provided informal overviews of the status of their stations' data management approaches:

- John Briggs, Konza Prairie, Oracle SQL database demo
- Vie Chow, Bodega Marine Lab, MOMS, Paradox demo
- Steve McNeil, UC NRS, FileVision Database/GIS**demo**
- Robert Moeller, Pocono Comparative Lakes, Reflex, Paradox demo
- Jim Brunt, Sevilleta LTER, Overview of programs
- Mike Hamilton, UC James Reserve, Hypermedia GIS demo
- Paul Montagna, Marine Science Institute, Overview of program
- Grady Cantrell, Hancock Biological Station, Overview of program, Dbase III
- Fred Lohrer, Archbold Biological Station, Overview of program
- Lance Risley, Institute of Marine and Coastal Sciences, Overview of program
- Rudolph Nottrott, LTER Network Office, "ANDREW" Internet System
- Warren Brigham, Illinois Natural History Survey, Overview of GIS applications
- Bill Seitz, Texas A&M, Galveston Bay, Macintosh-based demo
- David Nebert, Institute of Marine Science, Overview of programs
- Craig Staude, Friday Harbor Labs, Mac-based demo
- Deborah dark, La Selva Biological Station /OTS, Overview of programs
- John Heuer, Savanna River Ecology Lab, PROGRESS Database Language
- Bill Michener, Baruch Institute, Easy Entry demo
- John Porter, Virginia Coast LTER, database entry demo
- Jim Beach, Michigan State University, Demonstration of network for herbarium label data exchange

The following brief descriptions of the workshop pre-session demonstrations do not necessarily represent the range of data management approaches undertaken at marine and inland biological field stations. They do, however, reflect the diversity of ways in which scientific data management can proceed and is successfully being implemented at marine and inland biological stations.

1) ARC/INFO, Warren Brigham, Illinois Natural History Survey

The use of the ARC/INFO geographic information system running on Prime minicomputers and workstations was described. The system serves 300 users to provide a state-wide database for biodiversity, including occurrence records for distribution mapping, land use features, etc. The use of CIS to begin predicting potential habitats was demonstrated, including examples of how certain museum specimen label locations were biased by non-biological parameters such as road access. Also demonstrated was a study using GIS to increase the spatial accuracy of museum records by determining the probability surface for location descriptions on museum specimens.

2) Research projects database, John Porter, Virginia Coast Preserve LTER

A DBASE IV relational database of research project descriptions for the Virginia Coast LTER was demonstrated. The database could be sorted by date, place, location, investigator, and topic, and provided text descriptions of each project (historical and on-going).

3) Stebbins Cold Canyon Reserve GIS, Steve McNeil, University of California, Davis

A Macintosh GIS based on the program Business FileVision was demonstrated. The database links relational files about land use and environmental features to graphical display of points, lines and polygons. Query of the relational files can generate unique maps for visual display and hard copy. This program is used as an alternative to a written management plan for the University of California Stebbins Cold Canyon Reserve.

4) The "Andrew" multimedia bulletin board system, Rudolph Nottrott, University of Washington, Network Coordinator for LTER

An overview was given of the electronic network structure connecting the 17 ecological research sites which, along with the Network Coordination Office comprise the Long-Term Ecological Research network (LTER). Particular consideration was given to the national Internet and its NSFnet backbone. There are communications needs resulting from the wide geographical distribution of the LTER sites (Continental U.S., Alaska and Puerto Rico) and the dispersal of the 425 LTER researchers affiliated with over thirty institutions.

Electronic networks at three different levels are providing to ecological researchers: local-area networks (LAN), campus networks and wide-area net-

works (Internet). The functions at the highest level, the wide-area network level, include instantaneous and reliable electronic mail, access to supercomputers, access to national information and software repositories (including electronic bulletin boards), access to the LTER network office information system (mailing lists, mail forwarding system, LTER core data set catalog) and rapid long-distance transfer of data and programs, as well as text and graphics.

An overview was given of the electronic information system at the LTER network office. A detailed description was given of the LTERNET electronic mail forwarding system and a prototype installation of a multimedia electronic bulletin board (ANDREW) to be integrated with the mail system. The mail forwarding system can be reached from most major networks (Internet, Bitnet, Telemail, OMNET, UUCP, DialCom, MCI and others), and forwards messages to a user's "home" mail box on any of these networks. On request, an automatic reply function will return help information and various files stored on LTERNET. (To get initial help, send any message to forQuick@lternet.washington.edu (Internet) or forQuick@lternet (Bitnet).)

Plans for further development of the LTERNET information system were outlined, including the installation of an on-line catalog of LTER core data sets and development of this catalog into a distributed database system with local maintenance, administration and access control of all catalog entries and data sets, but with network-wide access for authorized researchers. Further development of this distributed database should include information already available at the LTER network office, such as the personnel directory, and data to be acquired in the near future (satellite images and other remotely-sensed data for all LTER sites).

5) HyperCard Bibliographic Database, Bill Seitz, Texas A & M, Galveston Bay

A bibliographic information database developed using HyperCard was demonstrated. This database runs on a Macintosh and is used for indexing maps, books, and articles. An optical scanner was used to read abstracts, and an optical character recognition program to convert bit-mapped images into ASCII characters. Approximately 2,000 records were entered in a short period of time with untrained staff. The HyperCard program can be linked to an Informix (SQL) database for rapid, relational search. Also described was a new service called MacSat which allows satellite images to be accessed directly via antenna, image processed, and displayed graphically on the Macintosh in color or 8-bit grey scale.

6) The PC-based FIS Database, John Briggs, Konza Prairie LTER

The FIFE project developed with NASA to study multi-scale remote sensing was discussed. The 100 gigabyte + image database is accessed using FIS software written by NASA and accessed by PC. The data is stored on a mainframe in an ORACLE database, and is accessed over a NOVELL network using the FIS software run by a PC. The database is accessed by about 200 users/month. The database will eventually be published on CDROM.

7) Macroscope Ecology Laserdisc Demo, Michael Hamilton, University of California, James San Jacinto Mountains Reserve

The "data" collected at biological field stations often consists of a wide variety of types and formats, ranging from paper-based tables of numbers and text, photographs, films, illustrations, and tape recordings of sounds, to many forms of machine readable information. Computerized techniques which allow multiple forms of information to be integrated and accessed from a single microcomputer require the use of a class of tools loosely called "interactive multimedia" or "hypermedia." Hypermedia systems generally consist of 32-bit microprocessors, hard disk mass storage, videodisc or optical disk storage, digital signal processors for audio files, and appropriate software. The most widely used hypermedia platform is the Macintosh computer running software called HyperCard (tm).

The James Reserve data management program uses a hypermedia approach as an index and database integration tool to many of the Reserve's information resources. A Macintosh hypermedia database was demonstrated using HyperCard to control access to laserdisc images, record and retrieve sound files, access GIS software and display text fields which can be queried using words or phrases. This database is used to organize a time-series photomonitoring study of plant succession and vertebrate census records. Spatial fields are calculated using an ARC/INFO and displayed through HyperCard. The database is used primarily as an ecological inventory system for the field station and for teaching at the station and campus.

8) HyperCard Demo, Craig Staude, Friday Harbor Labs

HyperCard (a programming environment for Apple Macintosh) is suited for many tasks at field stations that require a short learning time, ease of use, and flexibility. Several examples were offered

demonstrating these merits, and one which currently falls short of expectations. The Friday Harbor Labs Information Program was originally developed for public relations to a general audience (e.g., open house and county fair booth). It was subsequently adapted to advertise the facilities of the station at a scientific meeting. It is a series of screens of graphic-rich information, including scanned images and simple animation, which are linked by mouse activated buttons. The demo startup stack (Macintosh jargon for "program") is used to alert new users to the peculiarities and capabilities of our Mac IIci. It is automatically displayed whenever the machine is restarted, by means of the "Set StartUp" feature of the Mac operating system. The Research Sites stack is essentially a mini-GIS. Invisible buttons overlay symbols or features on a map of the local region. When clicked, these buttons call up additional, small-area maps or text fields that describe each site in greater detail. Craig's Amazing Crustacean Database is a prototype database for storing species-specific taxonomic and collection information. Craig's Amazing Tab Inserter is a utility program that edits an imported comma-separated or space-separated ASCII file (e.g., modem accessed temperature data from a NOAA/NOS sensor) and converts it into a tab-separated ASCII file that can be printed or exported to other applications (spreadsheets, databases, etc.). The most ambitious project to date is the FHL Housing program, which finds vacancies in housing units for visiting researchers, but it has not been implemented due to its slow response and the large number of query exceptions in the search arguments. Future versions might utilize streamlined script or add XCMDs to speed up the search process.

9) Easy Entry, Bill Michener, Baruch Institute

A database generation program called EASY ENTRY was demonstrated which is used to format data entry forms for inputting data while in the field. This system allows for the rapid uploading of data into other relational database programs running under MS-DOS.

10) SAS for database management, Paul Montagna, University of Texas, Marine Science Institute

Most users are familiar with the statistical features of SAS software (Statistical Analysis System version 6.03). However, SAS is an entire system with surprising data handling features. FSP can be used for database entry, checking and reporting. Base SAS has powerful manipulation features. Where data are maintained primarily for users who are familiar with and use SAS, it may be easiest for them to enter data directly into SAS. This eliminates the need for additional training and porting of data.

APPENDIX A—WORKSHOP PARTICIPANT LIST

The letters in brackets indicate working group participation:
[A] = Data Administration
[B] = Data Standards for Collaborative Research
[C] = Computer Systems for Data Management

Representing Field Stations and Marine Labs

Gary F. Anderson, Virginia Institute of Marine Science, College of William and Mary, Waterman's Hall, Gloucester Point, VA 23062 (Internet: gary@ches.cs.vims.edu) [A]

Michael A. Bowers, Blandy Experimental Farm, University of Virginia, P.O. Box 175, Boyce, VA 22620 [A]

Mary Bythell, West Indies Laboratory, Fairleigh Dickinson University, Teague Bay, Christiansted, USVI 00820 [A]

Barbara A. Carlson, Motte Rimrock Reserve, University of California, Riverside, Biology Department, Riverside, CA 92521 [A]

Victor Chow, Bodega Marine Laboratory and Reserve, University of California, Davis, P.O. Box 247, Bodega Bay, CA 94923 (Bitnet: UCDBML@UCDAVIS) [A]

Deborah A. Clark, La Selva Biological Station, Organization for Tropical Studies, Apartado 676, 2050 San Pedro, COSTA RICA (Internet: 3279995@mcimail.com) [A]

Philippe S. Cohen, Granite Mountains Reserve, University of California, Riverside, P.O. Box 101, Kelso, CA 92351 [A]

Robert W. Hastings, Turtle Cove Biological Research Station, Southeastern Louisiana University, P.O. Box 814, Hammond, LA 70402 [A]

Dean Kettle, Kansas Ecological Reserves, The University of Kansas, 2041 Constant Avenue, Campus West, Lawrence, KS 66047-2906 [A]

Robert Moeller, Pocono Comparative Lakes Program, Dept. of Biology, Lehigh University, Bethlehem, PA 18015 [A]

David L. Nebert, Institute of Marine Science, University of Alaska Fairbanks, AK 99775-1080 (Omnet: d.nebert) [A]

Janet Webster, Hatfield Marine Science Center, Oregon State University, 2030 S. Marine Science Drive, Newport, OR 97365 (Bitnet: hmsc@orstate) [A]

John M. Briggs, Konza Prairie Research Natural Area, Kansas State University, Division of Biology/Ackert Hall, Manhattan, KS 66502 (Bitnet: Konza@ksuvm) [B]

Walt Conley, Department of Biology, Box 3AF, Foster Hall, New Mexico State University, Las Cruces, NM 88003-0001 (Internet: wconley@nmsu.edu) [B]

Kenneth W. Cummins, Pymatuning Laboratory of Ecology, Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260 [B]

John H. Heuer, Savannah River Ecology Lab, University of Georgia, Drawer E, Aiken, SC 29801 (Internet: heuer@srel.edu) [B]

Fred E. Lohrer, Archbold Biological Station, P.O. Box 2057, Lake Placid, FL 33852 [B]

Lance S. Risley, Institute of Marine and Coastal Sciences, Rutgers Pinelands Field Station, Rutgers University, P.O. Box 206, New Lisbon, NJ 08064 [B]

Bill Seitz, Moody College of Marine Technology, Texas A&M University at Galveston, P.O. Box 1675, Galveston, TX 77553 [B]

Emery R. Boose, Harvard Forest, Harvard University, Petersham, MA 01366 (Internet: eboose@lternet.washington.edu) [C]

Grady Cantrell, Hancock Biological Station & Center for Reservoir Research, Murray State University, Murray, KY 42071 [C]

Michelle Georgi, Forbes Biological Station, Illinois Natural History Survey, Box 599, Havana, IL 62644 [C]

Linda May, Horn Point Environmental Laboratories, University of Maryland P.O. Box 775, Cambridge, MD 21613 (Internet: may@umdc.umd.edu) [C]

Steve McNiel, Stebbins Cold Canyon Reserve, University of California, Davis, 144 Walker Hall, Davis, CA 95616 [C]

Paul A. Montagna, Marine Science Institute, University of Texas at Austin, Box 1267, Port Aransas, TX 78373 (Bitnet: paul@utmsi) [C]

Rudolf Nottrott, LTER network office, College of Forest Resources AR-10, University of Washington, Seattle, WA 98195 (Internet: rnottrott@lternet.washington.edu) [C]

Bob Vande Kopple, University of Michigan Biological Station, Pellston, MI 49769 (Bitnet: userhcyb@umichub) [C]

Craig Staude, Friday Harbor Labs, University of Washington, 620 University Road, Friday Harbor, WA 98250 (Bitnet: 96680@uwacdc) [C]

Nicholas Wolfe, NOAA/NMFS Beaufort Laboratory, Beaufort Laboratory Beaufort, NC 28516-9722 [C]

Workshop Pre-session Coordinator

Michael P. Hamilton, James San Jacinto Mountains Reserve, University of California, Davis, P.O. Box 1775, Idyllwild, CA 92349 [C]

Rapporteurs—Data Administration:

William K. Michener, Baruch Institute, University of South Carolina, Columbia, SC 29208 (Bitnet: a299050@univscvm. Internet: bmichener@lternet.washington.edu) [A]

Ken Haddad, Florida Marine Research Institute, 100 8th Avenue SE, St. Petersburg, FL 33701 (Omnet: r.burkhart) [A]

Rapporteurs—Data Standards for Collaborative Research

Warren Brigham, Illinois Natural History Survey, 607 E. Peabody Dr., Champaign, IL 61820 (Bitnet: brigham@uiucdenr) [B]

James W. Brunt, Sevilleta LTER, Department of Biology, University of New Mexico, Albuquerque, NM 87131 (Internet; jbrunt@sevilleta.unm.edu) [B]

Rapporteurs—Computer Systems for Data Management

John H. Porter, Virginia Coast Reserve LTER, Department of Environmental Science, Clark Hall, University of Virginia, Charlottesville, VA 22903 (Internet: jhp7e@virginia.edu, Bitnet: jporter@lternet) [C]

Jeff Kennedy, University of California Natural Reserve System, 300 Lakeside Drive, 6th Floor, Oakland, CA 94618 [C]

Organization of Biological Field Stations Representative

George Lauff, Kellogg Biological Station, Michigan State University, 3700 E. Gull Lake Dr., Hickory Corners, MI 49060 (Internet: lauff%kbs.decnet@clvaxl,cl.msu.edu)

Southern Association of Marine Labs—Representative

James Alberts, University of Georgia Marine Institute, Sapelo Island, GA 31327 (Omnet: j.alberts)

Host Site Representatives

John B, Gorentz, W.K. Kellogg Biological Station, Michigan State University, 3700 E. Gull Lake Drive, Hickory Corners, MI 49060 (Bitnet: gorentz@msukbs, Internet: gorentz°/okbs.decnet@clvaxl.cl.msu.edu)

Stephan Ozminski, W.K. Kellogg Biological Station, Michigan State University, 3700 E. Gull Lake Drive, Hickory Corners, MI 49060 (Bitnet: ozminski@msukbs, Internet: ozminski%kbs.decnet@clvaxl.cl.msu.edu) [B]

Lolita Krievs, W.K. Kellogg Biological Station, Michigan State University, 3700 E. Gull Lake Drive, Hickory Corners, MI 49060 (Internet: krievs%kbs.decnet@clvaxl.cl.msu.edu) [B]

Workshop Consultant

Patricia Rich, Patricia Rich Associates, 115 Lake Forest, St. Louis, MO 63117

Michigan State University—Speakers & Observers

James H. Beach, MSU Beal-Darlington Herbarium, Michigan State University, Department of Botany and Plant Pathology, East Lansing, MI (currently: Herbaria and Museum of Comparative Biology, Harvard University, 22 Divinity Avenue, Cambridge, MA 02138 (Internet: beach@huh.harvard.edu) [B]

Paul M. Hunt, Academic Computing and Technology, Michigan State University, 400 Computing Center, East Lansing, MI 48824 (Bitnet: pmhunt@msu)

National Science Foundation • Speakers & Observers

Robert Robbins, National Science Foundation, Room 312, 1800 G St. Washington, DC 20550 (Internet: rrobbins@note.nsf.gov)

James L. Edwards, National Science Foundation, Division of Biological Research Resources, 1800 G St. NW, Washington, DC 20550 (Internet: jledwards@note.nsf.gov)

APPENDIX B - GEOGRAPHIC INFORMATION SYSTEMS/ADMINISTRATIVE ISSUES

William K. Michener
Baruch Institute
University of South Carolina
and

Ken Haddad
Florida Marine Research Institute

INTRODUCTION

Management of spatial data relevant to a site is an important component of that site's data management system. A Geographic Information System (GIS) is a data management system that allows the capture, synthesis, generation, retrieval, analysis, and output of spatial data and, by some definitions, non-spatial data. Although this particular definition of GIS can be argued, there is general agreement that it is a rapidly evolving technology which is revolutionizing geographical analysis and has applications in many fields of science and resource management.

Parker (1988) and Cowen (1988) attempt to put into perspective the definitions and characteristics of a GIS as well as some of the fundamental operations. Some additional references which deal with all facets of GIS include: Burrough, 1986; Goodchild and Gopal, 1989; GIS/LIS'89; ASPRS, 1986; PE&RS, 1988; Michener, et al., 1989. In addition, almost every field of science and resource management now includes published articles and workshops related to GIS technology.

The applications of GIS technology at biological and marine field sites are numerous but can be approached through two broad and interrelated perspectives: (1) accomplishment of site management goals, and (2) accomplishment of research goals.

GIS related site management goals can range, for example, from cataloging and maintaining information generated at the research level, to conducting an integrated analysis of data collected by individual researchers, supplemented by data considered generic to a site, for management of the site's natural resources.

GIS related research goals can range from the use of spatial data for choosing research sites and the visual presentation of a researcher's data, to the use of GIS as an analytical tool for drawing scientific conclusions. In reality, the use of this technology as a research tool has only minimally been explored and in limited fields of science.

The interrelated applications of GIS technology as a tool for management and research can provide both opportunities and conflict at a field site. All aspects of GIS development at a site should be considered prior to implementation.

BLUEPRINT FOR A GIS

It is likely that, if not already implemented, many inland and coastal biological field stations are or will be considering the design and implementation of a GIS. At the site level, GIS should be considered a structured form of data management. The decision to design and implement GIS is an immediate step into a sophisticated level of data management. A site immediately goes beyond information cataloging and archiving and must be concerned with all aspects of data management and administration. All of the discussions on data management in previous chapters are relevant to GIS. In fact, particularly at the smaller sites, GIS may be the core for data management implementation.

Depending on the site and its functions, the individual researcher can have varying influences on GIS development. A concern often voiced at the research level, when administrative structure is imposed, is that science is being stifled. It is important to include the researcher in GIS design and implementation to assure that a rational data administration structure is applied and the user support base developed.

It should be recognized that GIS implementation at the site level may not be of benefit to all sites. Addressing other aspects of data management may better meet a site's needs. A given site should determine the need for a GIS from an administrative and research perspective and not assume its benefits. Individual researchers may provide the impetus for implementing a single user GIS as part of a research program. That is a site-specific issue. These observations are directed primarily at the site-initiated GIS.

There are avenues for GIS implementation outside existing data management operations. Successful GIS development and administration can occur as a parallel entity connected to data management efforts but not governed by the "data center." In fact, traditional data management administration can conflict with GIS evolution even though the principles of data management need to be applied.

If design and implementation are to occur at the site level, a GIS needs assessment should be conducted (Guptill, 1988). A GIS needs assessment is

not a trivial process, even for small or low activity sites. If the knowledge base to conduct a proper GIS needs assessment is not on site, then off-site expertise must be consulted. Because a needs assessment requires time and resources, it is often considered an impediment and consequently ignored. However, proper understanding and design are critical for long term data applications and research support, and should not be construed as an impediment to GIS implementation,

GIS IMPLEMENTATION CONSIDERATIONS

While a needs assessment should be a prerequisite to GIS implementation, elucidation of some of the important management considerations can provide a site administrator with some useful insight. Understanding the people, data, and cost considerations can facilitate successful GIS implementation.

PEOPLE AND TRAINING

Implementation at the site level should relate to the intensity of staff use and needs. This is a people consideration and will have major impact on successful implementation. Access by both the managers and researchers should drive the entire GIS development process. Technically, hardware and software play a role in implementation, but planning for longterm success must focus on the user, GIS operators, and their interactions.

Training should be considered key to GIS use and accessibility. Accessibility to the GIS can be accomplished by the availability of a skilled translator who can work with the investigators to build their understanding of the capabilities of the GIS and assist in operation of the applications software. This person should have site knowledge, a science (includes geography) background, and be well trained in the GIS applications software.

The researcher may prefer to be the analyst with hands-on skills. In this case the researcher must be trained not only in the applications software, but also in GIS concepts and principles. As with any technology, improper use and lack of understanding of the equipment can lead to error. A combination of the availability of a skilled translator and researcher training may be the best solution for optimum accessibility and effective utilization.

DATA

The use of GIS technology is dependent on the availability of data. Acquisition of hardware and software does not mean that a site has, or will ever have, a functional GIS. Selection and prioritization of data sets for entry and general access should be determined by the site administration in consultation with the site researchers. The needs of outside users should also be a consideration.

Although user needs drive the prioritization of data acquisition and maintenance, some additional issues which impact prioritization and determine successful GIS implementation are:

Spatial Resolution: Spatial resolution is probably one of the most important aspects of a GIS and one of the least understood. Spatial resolution can be divided into two components. The first component consists of the positional accuracy of an entity in the database. For example, if the location of a bald eagle nest is not accurately located it may show up in the middle of a lake, when compared to a database depicting land cover. The second component of spatial resolution is related to the user's need for detail and contains the elements of positional accuracy. Does the user need an accurate location and description of each tree in a forest, or will the location and description of the forest suffice

The subject of resolution is complex, and, if not properly addressed, could lead to wasted effort and be a major source of error in GIS analyses. Goodchild and Gopal (1989) put the question of the accuracy of spatial data, relative to GIS technology, in perspective. They suggest that the statistics do not even exist to define the error when spatial data are analyzed. It should not be concluded that GIS implementation is error bound, but that it is necessary to proceed with caution and with knowledgeable planning.

Coordinate System; Selection of an earth coordinate system is important. The three common coordinate systems are Latitude/Longitude, Universal Transverse Mercator (UTM), and State Plane Coordinates. Most coordinate systems are interconvertible, but commonality at a site may be advantageous for general communication of the data.

Quality and Documentation: Data quality and data documentation are two major issues in the development of a GIS. Variations in data quality can be amplified when analyzed in relation to other data. For example, when analyzing the relationship of soil data types (90 percent accurate) to the location of earthworm colonies (50 percent accurate), the resulting data may be only 45 percent accurate relative to hypotheses being tested. It becomes extremely important to have adequate documentation defining the source and lineage of the data and an assessment of the quality and accuracy of that data. The individual researcher or user can then determine the utility of that data set relative to the analyses they wish to conduct.

Proprietary Rights: Proprietary rights to data can often be the first controversial issue to arise when a site-initiated GIS is implemented. This issue should be anticipated and settled prior to implementation.

COST

As with any data management effort, the cost of the process and ability to support that effort should be a deciding factor in implementation. From an administrative perspective, can the site bear the long-term costs? Can a site finance common database generation and data maintenance and updating, and do so at the spatial resolution and update frequency necessary to make it useful to the researchers and other users? These are tough questions that are often ignored. The alternative to dealing with these questions in the planning process is to buy the hardware and software and hope that funds will become available for development and implementation. However, this approach has a high rate of failure.

Depending on the site, costs can be partitioned into the following:

Hardware: In addition to the initial purchase, the need for hardware evolution and expansion may be more accelerated for CIS needs than for traditional, non-spatial data management. Needs for computer hardware peripheral devices go beyond traditional printer and hard drives.

Software: Both operating and applications software can be a significance expense.

Maintenance: Hardware and software (operating and applications) often require maintenance and upgrades if a fully functional GIS is to remain operational. Routine expenses to support daily operations and replenish depleted supplies are far more than those associated with standard data processing.

Personnel: Depending on the site, CIS personnel functions can require the attention of more than one full time person for each function. The primary personnel functions are: (1) Data administration and coordination. This involves the development and maintenance of support for the GIS and all the basic administrative functions associated with data management. (2) Data capture. This refers to the identification, evaluation, and preparation of data to be entered into the GIS. This process is critical to the success of the GIS, particularly in understanding quality and accuracy of the data. (3) Data entry. This can be one of the more expensive functions, particularly for the common databases to be supported by the site. (4) Data analysis and output. This function requires technical skills and is the critical measure of successful implementation.

The costs associated with the above functions are variable and are often linked to the magnitude of the site operations.

Training: Training should include not only the technical aspects of GIS software and operations but also the development of an understanding of the theory and algorithms applied during GIS analyses. Frequently, the GIS is treated as a black box and one is led through a process which, if not understood, leads to erroneous conclusions. Training must be a continual and scheduled process.

Data: This is frequently the most costly portion of the GIS operation. Costs include data acquisition, quality control, data maintenance and updating, data analysis and output, and archiving and security. These operations can easily cost at least four times as much as hardware and software. For example, it may be possible to acquire hardware and software for $10-20,000, but the data necessary for implementation of an operational GIS could cost an additional $40-80,000.

LITERATURE CITED

ASPRS (American Society for Photogrammetry and Remote Sensing). 1986. Proceedings of Geographic Information Systems Workshop, American Society for Photogrammetry and Remote Sensing, Falls Church, Virginia. 264 pp.

Burrough, P.A. 1986. Principles of Geographic Information Systems for Land Resources Assessment. (Oxford: Clarendon).

Cowen, D.J. 1988. GIS versus CAD versus DBMS: What are the differences? Photogrammetric Engineering and Remote Sensing 54(11) 1551-1555.

GIS/LIS '89. 1989. GIS/LIS '89 Proceedings. American Congress of Surveying and Mapping, American Society of Photogrammetry and Remote Sensing, Association of American Geographers, Urban and Regional Information Systems Association, AM/FM International. Volumes 1 and 2. 836 pp.

Goodchild, M. and S. Gopal. 1989. The Accuracy of Spatial Databases. Taylor and Francis, Bristol, PA. 290 pp.

Guptill, S.C. (ed.). 1988. A process for evaluating geographic information systems. U.S. Geological Survey Open-File Report 88-105. 136 pp.

Michener, W.K., D.J. Cowen, W.L. Shirley. 1989. Geographic Information Systems for Coastal Research. Proc. of Sixth Symp. on Coastal and Ocean Management/ASCE: 4791-4805.

Parker, H.D. 1988. The unique qualities of a Geographic Information System: a commentary. Photogrammetry Engineering and Remote Sensing 54(11): 1547-1549.

PE&RS (Photogrammetric Engineering and Remote Sensing). 1988. Special GIS Issue. 54(11): 1-167.

APPENDIX C—CLIENT/SERVER DATABASE ARCHITECTURE, NETWORKS, AND BIOLOGICAL DATABASES

James H. Beach
Herbaria and Museum of Comparative Biology
Harvard University
22 Divinity Avenue
Cambridge, MA 02138

INTRODUCTION

The development of the academic research networks, in particular, the NSFnet or Internet and the forthcoming National Research and Education Network (NREN), will provide the potential to make myriad biological data resources available to scientists and students around the world. Although electronic mail, interactive sessions with remote applications, and high-speed file transfer are now integral to many research programs, the development of database systems which will bring biological data resources to the networks is in its infancy.

DATABASE ARCHITECTURE

The NSFnet/NREN permits several types of long-distance access to biological data sets. The traditional and still commonplace form of communication with remote databases is one where users connect over the network to establish a terminal session with a remote host. Remote users log onto the computer and operate database application programs in the same way a local terminal user would. This is an example of "host/terminal" database architecture.

A technical characteristic of host/terminal database systems is the logical cohesion between the database manager software, which stores and manages user data files, and the application programs interacting with it (Figure C-l). A major benefit of the logical integration of the layers is ease of database system development; application programs can be tailored to fit like a glove around the features of the database manager. Remote access to data in host/terminal systems is exclusively through the host's application programs.

"Client/server" database architecture in contrast, uncouples the application programs from the database server software (Figure C-l). Client/server databases sandwich an additional logical layer to handle communication between the server and the applications, through the use of a go-between, standard query language.

The importance of the client/server model, in the context of network access to information, is that it allows the application layer programs and the database server software to reside on different machines. Because the two layers communicate through discrete, structured messages, the conversation can be carried out between machines connected across the room, across the country, or across the globe. The development of the high-speed, high-capacity research networks strengthens the importance of client/server systems for biological databases, because institutional data servers could be queried at any time over the network by any number of applications at remote sites. A particular application could rapidly access multiple institutional servers over the network channel.

A functional difference between the client/server and host/terminal database architectures has far reaching implications for access to scientific information. In the host/terminal model, a remote network user running a (virtual) terminal session from a local computer, e.g., a desktop PC, only receives screen images; information is visually presented but there is no mechanism to download data records for local use. Capturing data from a foratted screen display, one screen at a time, is usually an imperfect process at best. As a result, access to remote information is essentially limited to the duration of the virtual terminal session. A client/server database, in contrast, transmits actual data records to the remote user's system. The records (in a standard exchange format) are then available to local programs for further processing or formatting. Note that with client/server architecture, remote users are not constrained by the application interface or program logic of the server system, but work with a familiar local application to query and obtain records from network database servers. An additional distinction of the client/server approach, in a computing environment characterized by autonomous institutions in a collaborative enterprise, is that it allows organizations to control the ongoing development of their hardware, database, and application software, while at the same time presenting a standard and stable network interface for remote client access.

STANDARDS

The scientific disciplines will need to resolve various types of data format, application and data

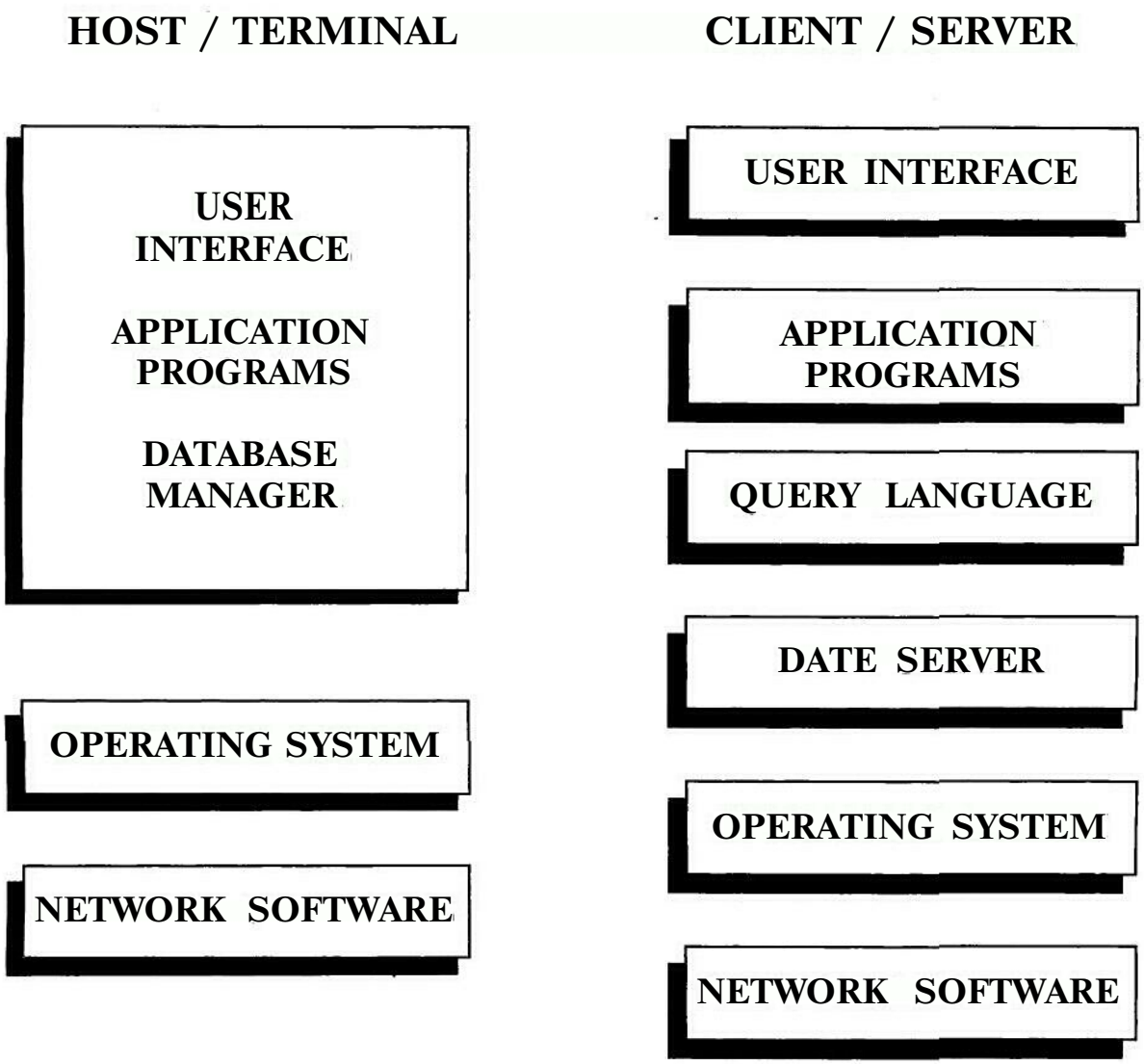


Figure C-l. Database System Architecture

communication standards in order to establish network client/server systems. Database systems developed in isolation, on small, single-user computers or on large un-networked machines may be elegantly customized for local needs, but biological databases intended to inter-operate with remote applications will need to be specified, designed and implemented on much technical common ground. The most important computerization standards for network client/server systems are:

A common set of core data definitions

Discipline or community-wide standards for core data type definitions, coding, and cataloging rules are essential for the biological information stored in systems designed for network access. These standards comprise formal descriptions and specifications for data types currently in use in non-computerized or non-networked databases. Ecologists will have an especially difficult task, due to the breadth of ecological research, but certain ecological data types are already fairly well standardized.

There are several ongoing ecology, systematics, and museum community efforts in this area, including: the LTER data catalog project, NSF-sponsored, discipline-based data workshops, and various projects of the International Working Group on Taxonomic Databases in the Plant Sciences, the Association of Systematics Collections, the Museum Computer Network, as well as several additional society and institutional efforts.

A standard exchange record format

The results of a search on a remote data server must be returned to the requesting application in a standard record format. Without such a format, result sets would not be understood by the client process, and client/server data exchange would be impossible. Included here are standards for data representation, syntax and structure specifications for records and fields. Data definition and encoding standards (above) can be applied as a part of the exchange record format definition.

The library community has standardized the definition of data elements and data record exchange formats in its highly successful MARC record format. The MARC format standards are only applied to records intended for exchange and not to database design. They have provided tremendous stability and have greatly facilitated information interchange between diverse library database systems.

Some of the organizations mentioned above have begun to investigate a MARC approach for museum data and there is growing interest in MARC formatting of biological information for record exchange purposes.

Standard network request and response protocols

For client applications to communicate with data servers, there must be a well-defined language and syntax for the interaction. Such standard protocols specify the structure and to some extent the content of the messages passed between client and server machines as part of a data request/response dialogue. They also specify how control and state information will be communicated and under what conditions diagnostic messages and result sets will be transmitted to the originator of a query.

The best example of standard protocols for the retrieval of information in a client/server architecture again comes from the libraries. That community sponsored the development of the ANSI/NISO standard Z39.50 (NISO, 1988) which specifies network session protocols for library information retrieval. The Z39.50 protocols are being used for library data exchange as part of the multi-institutional "Linked Systems Project" (Fenly and Wiggins, 1988). A thorough examination of the libraries' computing and standards infrastructure would assuredly be profitable for nascent data standards efforts in biology.

IMPLEMENTATION

Biological client/server database systems can be implemented over networks today, and they will become increasingly common as discipline, national, and international communication standards are completed. There are numerous engineering options for implementing client/server systems, but an overriding design objective is ultimate compliance to network communication standards, particularly those of the International Organization for Standardization (ISO), the U.S. National Information Standards Organization (NISO), and to the data definition standards developed within the scientific disciplines.

As a prototype example of client/server architecture, a biological client/server database system using two networked computers was demonstrated at the Data Management Workshop. A Digital VAX/VMS system functioning as the client was located at Kellogg Biological Station (KBS), while the server, a Sun Microsystems workstation, was about 70 miles away on the Michigan State University campus in East Lansing. An Ingres client application at KBS, using a query-by-form screen, recorded a query specification based on user selections and then mailed the query to an Ingres herbarium specimen data server in East Lansing. That computer parsed the contents of the network mail message and applied it as a query against the database. The result set was stored in a file, then mailed back to the KBS client application within a few minutes. The client process reported the

arrival of the result set to the user and imported the records into a local database table for further processing.

A mail-based client/server system, although in some ways a "low-tech," approach, uses a universally supported network application and is relatively easy to implement. Limitations include delays caused by network mail routing, limits on the content and length of mail messages imposed by network mail programs and the inherent difficulties of managing state information and a request/reply process with network mail. Due to short-term exigencies, the Ingres QUEL query language was used, but SQL (Structured Query Language), which is the industry standard query language for client/data server communications in relational database systems (Date, 1990; Tucker, 1990), could have been employed.

A more standard and sophisticated client/server design is a "connection-oriented" approach, whereby client and server processes enter into a real-time network dialogue. In this case, a precisely defined protocol is required for the client/server communication (e.g. ANSI/NISO Z39.50) which specifies a predictable sequence of back-and-forth control and data messages while a client is requesting or receiving information from a server. Connection-oriented, client/server, network database protocols function on top of lower-level network communication standards to form an integrated, layered stack of protocols. In contrast to the "connectionless" mail-based, client/server approach, a connection-oriented system requires sophisticated system and network-level programming to implement, but, in addition to speed, it offers numerous technical advantages for monitoring client/server sessions and for returning useful status information to the user.

SUMMARY

Client/server database systems can provide a direct and powerful method for biological database access over the NSFnet, the NREN and the international extensions of those networks. When implemented for open access, they have several advantages over host/terminal systems. The most notable are:

1. Users would not need to obtain an account on each target system they wish to query, and they would not need to learn the logic and design of each institutional database application.

2. Data records can be acquired and processed locally in the client/server model, as the data server actuallyopies data records and not just a refreshed screen image to the remote user. Only result sets meeting the user's query criteria are returned over the network.

3. Institutions could provide open, read-only, server-level, access to their biological data resources with limited risk or loss of administrative control.

4. Once network, client/server, interface standards are in place, institutional database system hardware, server software and applications can evolve independently and still provide open, long-term, network access.

REFERENCES CITED

Date, C. J. 1990. An Introduction to Database Systems. Vol. 1. 5th Ed. Addison-Wesley, xxv+ 854 pp.

Fenly, J. G. and B. Wiggins. 1988. The Linked Systems Project: a networking tool for libraries. Online Computer Library Center, Dublin, Ohio. xii + 138 pp.

National Information Standards Organization. 1988. Information Retrieval Service Definition and Protocol Specification for Library Applications [Z39.50-1988]. Transaction Publishers, xii + 52 pp. (Available from: Transaction Publishers, Rutgers University, New Brunswick, N.J., 08903, USA)

Tucker, J. T. 1990. The inevitable merging of SQL. Unixworld *7(2]:* 68-70, 72, 74.

APPENDIX D—INTERSITE ARCHIVAL AND EXCHANGE FILE STRUCTURE

(Excerpt from an article submitted to Coenoses)

Walt Conley
New Mexico State University

and

James W. Brunt
University of New Mexico

An Intersite Archives File structure has been defined in order to facilitate the need for an orderly approach to the design and implementation of a data manipulation capability. The manipulations to be done are alterations on the shape and/or the content of original (archived) data files, and communication of original or descendent files to remote sites.

The Intersite Archives File (Figure D-l) is a generalized data structure that contains full documentation and comments. It is intended that the test of adequate documentation is that these files should stand alone, and that the file itself should contain sufficient information so that a future investigator who did not participate in collecting the data can use the information for some scientific purpose. The Intersite Archives File structure is intended to be used across cooperating research sites that, taken together, represent the ultimate heterogeneous computing environment. The intent is to define a generic data structure that can be useful on any hardware and software system, and that can be sent on any electronic network or file transfer system. A companion effort provides an Intersite Toolkit for obtaining information from the files; there are also tools for manipulating and screening the files. Manipulations include stripping an Archives File of various categories of information to produce descendent files that can by read by any application package.

The basic Intersite data structure is a generic ASCII flat file that contains categories of information that define the data, as well as the data itself. Intersite Archives Files can be of any basic structural type, including statistical data, text data, graphics data (e.g. files that you can write to a graphics plotter), gene sequence data, or bit map image data. Other file types will no doubt be required. Note that file type refers to the general nature of the data in the file, and not to data typing such as floating point, integer, or character. All of the data in the Intersite Archives Files are ASCII characters, and provision is made in the Intersite Toolkit for handling files containing non-printing ASCII characters which make file transfer difficult on some networks and impossible on many of the file transfer protocols.

The general categories of data in an Intersite Archives File is as follows.

\ log; A record of the history of the file; when it was initiated, updating, changes entered, locations and dates of copies of the file. Any ASCII characters with any format may be included.

\ doc: Documentation—as detailed a description as is necessary of the data contained in the file. Any ASCII characters with any format may be included. An ABSTRACT may be included here to allow automatic extensions of data dictionaries from Archives Directories. The abstract is simply a paragraph beginning with ABSTRACT and ending with a blank line; it may appear anywhere in the documentation section.

\ type: File type refers to the basic nature of the data. Statistical files are typically rows by columns tables of numeric or character data. Text files include bibliographic data, abstracts, or any prose. Graphics data refers to files which can be written to a plotter or a printer. Genome data refers to long sequences of base pairs that require line delimiters and other embedded information. Image data refers to bit map images.

File typing currently includes statistical, text, graphics, genome, and image. Other file types are possible and can be added as necessary. The only operation anticipated on file type is identification for sorting.

\ header: Header refers to labels for the columns of data in a statistical data file, or a list format text file. This allows for automatic building of data dictionaries from Archives Directories. For files of other types, the header can contain keywords that describe the data. Labels or keywords in the header are automatically retrieved for the development of data dictionaries in Intersite Archives data directories. The Intersite Toolkit provides tools that do this work.

\ data: Data refers to the actual data of the archives file—the numbers, text, etc. The data section may contain embedded comments that further describe individual records of the data.

\ log

A log of activity for the datafile including names, dates, etc.

\doc

All the documentation needed to accompany the datafile in free format

ABSTRACT

Includes the option for an extractable abstract

\ type

A one word descriptor of the data ie., statistical, image, list, etc.

\ header

A description of the attributes for statistical data

\data

The Data

(Includes comments)

Figure D-1 Intersite Archive File Structure

The Intersite Toolkit contains programs that manipulate the Archives File data structure, making the files ready for applications programs such as relational data management systems, statistical or graphics packages, and reporting systems such as text formatters. Any combination of categories of information in the Archives Data Files can be extracted for further use. Thus in a statistical file it is possible, for example, to quickly extract only the column labels and the table of numbers, only the ABSTRACT, only the documentation section. The Toolkit also contains compression and decompression filters (useful for disk maintenance and some communication applications), an encryption and decryption algorithm (useful for converting files with non-printing characters to files that can be sent over networks that do not handle binary data or via dial-out modem transfers), and a suite of programs that automatically build and reference a data dictionary that contains various presentations of labels, keywords, and abstracts.

For statistical and text file types, there are 2 additional formats that are useful to consider. "Table" format is the typical row X column format of statistical data with a label at the top of the column. "List" format is a transposed table, where the labels are on the left margin, giving unlimited category width but with a single column of data. List format is useful for text data such as keyworded bibliographic citations, or any similar kind of text. Note embedded comments can be included anywhere in the \ data section simply by enclosing the comment in curly brackets. The only restriction is that comments and other data cannot be mixed on the same line. (This preserves the positioning of tabular data, and serves the goal of keeping these files "readable" by humans.)

The general structure of an Intersite Archives File (type is "statistical") in Table format is shown in Fig. D-2. Note that the category indicators (\ log, \ doc, \ type, \ header, \ data) occupy a separate line but do not need to begin in any particular column. The suggested categories are optional, although deletion of any category limits the usefulness of the file and the use of the Intersite Toolkit for manipulating the files. The structure of an actual Intersite Archives File in Table format is shown in Fig. D-3. The general structure of an Intersite Archives File in List format is shown in Fig. D-4.

In the log and doc sections, there are no format requirements, and free-form text can be entered as you choose. In the header section and the data section, some structure is necessary. In the Table format, the header labels provide searching tags for the data file manipulations (and serve as handy reminders), and the dashed lines indicate the maximum width of each column of data (which is used for subsequent manipulation of the data columns. The dashed lines are not necessary for many applications; they are useful for providing information for manipulation routines. To include them requires little, and adds considerably to the potential for cross-site data manipulation. In the List format labels appear at the left of the field, and the dashed-lines indicator for column width is not necessary. In Tables, data columns conform to the labels in that they are in the same order, and in the Table format, the data must fit within the number of columns indicated by the dashed lines.

A Table format has one or more columns, and a List format has only a single column. Columns may be of arbitrary width. The labels in each case provide for data abstraction in good applications packages in that the researcher may refer to variables by name (i.e. labels) rather than, for example, as column 3. Archives Files are specifically intended to be browsed by human researchers who want to become familiar with the data and the circumstances involved in the collection of the data. Once converted to the descendent files that will be manipulated via available relational operators (etc.), data files are not designed to be read by humans, and will be confusing to look at.

In practice, any numerical data set can be put into a rows by columns table format, and the only restriction is that the columns have some white space between them. This is the format that is typically used when recording data in the field, or when reporting data, and the Intersite data structure simply provides a computerized version of what you probably do anyway. There is a utility in the Intersite Toolkit called "extract" that can subset the standard Intersite Archives File structure (Figure D-5). This utility can create a new file with any combination of the various elements of an Archives File stripped from the original; the original is, of course, preserved intact. Other programs in the Intersite Toolkit provide manipulation and screening of the Intersite Archives Files, building of a data dictionary based on labels and keywords, extracting and sorting Abstracts, and generally obtaining information from the Archives directories.

Once the documentation has been stripped from the chosen archive files, and the files are ready for some serious work, the descendant files can be read into any applications package of your choice. A next obvious choice is entering the filtered data into a database system for further manipulation. If you use a relational database system is being used the

labels can be used without further change. Some statistical packages can also use such labels. If an application can make use of short explanations of the labels (e.g. SAS), such information can be included in the doc section. If the only thing needed is a table of numbers to read onto a graphics or statistics package, only the data should be extracted and not the header or the comments embedded in the data.

The Intersite Toolkit currently contains utilities that convert descendent files into a common relational data base format. Additional utilities can be easily added.

For more information about the Intersite Archives File Structure or the Intersite Toolkit contact:

Walt Conley
Department of Biology
New Mexico State University
Las Cruces, New Mexico 88003

\log
A history of the **file.**
\doc
Any amount of explanatory text in any format.

ABSTRACT Title of the data set followed by a paragraph of text. You will also want to put the name of the responsible researcher and a phone number or E-mail address. The abstract can appear anywhere in the section.

(NOTE: blank line under ABSTRACT allows automatic extraction.)
\type
statistical
\ header
coil_label col2_label ... coin_label
---------- ---------- ... ----------
\data
DATA .... in column format as described in the header.
DATA ....
DATA ....
Comments: contained by row within DATA and referring to specific portions of data. Any ASCII characters are allowed, and no format is imposed other than comments occupy an entire line, and must be enclosed inside curly brackets. By convention, a comment follows the record being described.

DATA ... ...
DATA ... ...
Comments may occur anywhere in DATA.
DATA ... ...

Figure D-2: General structure of an Intersite Archives File in the Table format.

\log
********** **BEGIN CHANGE LOG** **********
23 December 1987. Data entered and documentation established. MAUhl
********** **END CHANGE LOG** **********
\doc
ABSTRACT Ant Total Density on the Jornada. This file, ant/ _ total.density, is monthly mean densities of new colonies grouped into zones, pooled for all species. The last 5 columns represent the monthly densities by year, and the first column describes the area ("zone") where the colonies were located. Data were collected by Marsha R. Conley 1982-86.

These 5 species were pooled to create the file:

Code:	Scientific name:	Common name:
PODE	Pogonomyrmex desertorum	Desert Harvester Ant
PORU	Pogonomyrmex rugosus	Red Harvester Ant
MYDE	Myrmecocystus depilis	Honey-pot Ant
MYMI	Myrmecocystus mimicus	Honey-pot Ant
NOCO	Aphaenogaster cockerelli	

\ header

Zone	1982	1983	1984	1985	1986
\data					
Playa	0.0	0.0	0.0	0.0	0.0
Mesquite Fringe	2.7	3.0	3.3	3.3	3.3
Basin Slope	6.7	8.1	8.8	7.9	6.8
Bahada	0.5	0.6	0.8	1.3	1.5
Lower Piedmont	2.8	3.1	3.2	2.4	2.6
Upper Piedmont	0.8	0.9	1.1	1.1	0.9

(Only Pogonomyrmex were found in the Upper Piedmont)

Figure D-3: Structure of an Intersite Archives File in Table format.

\log
Records of the history of the datafile. When it was initiated, changes entered, locations and dates of copies of the file. Any ASCII characters with any format may be included.

\doc
Documentation: As detailed a description as necessary of the data contained in the file. Any characters with any format may be included. An ABSTRACT of 1 paragraph may be included anywhere in this section.

\type
Typically List files are of type text.

\ header
Nothing needed here for List format. Note that the \ intersite data dictionary tools will pick up the Labels at the left margin of the first record and will automatically treat them similarly to the column labels from the Table format.

\data
This is a comment. Note that the new line below is required to automatically identify the List format.

\ begin (verbatim)

```
labell      line of text      ....
label2      line of text          :
label3      line of text          :     ->      record 1
   .                              :
labein      line of text      ....
labell      line of text      ....
label2      line of text          :
label3      line of text                ->      record 2
   .                              :
labein      line of text      ....
   .
   .
```

Figure D-4. General structure of an Intersite Archives File in List format. Note that the Labels are simply the first unbroken string of characters in each line.

APPENDIX E—SYSTEM SELECTION OVERVIEW

John H. Porter
University of Virginia

and

Jeff Kennedy
University of California Natural Reserve System

Advice on choosing a computer and software is always short-lived. Changes in systems and prices occur almost daily. Nonetheless, such advice is valuable to a person setting up a new data management system. The following sections attempt to provide needed information to new data managers on how to choose a PC computer (running MS-DOS) or a Macintosh (running the Apple operating system). What is not included is guidelines on whether to choose a PC or a "Mac." This is because the general capabilities of the two computers overlap so greatly. Choosing between them will depend on relative costs, the computing environment and the preferences of users.

SELECTING AN MS-DOS COMPUTER

The type of computer that is "best" for you depends entirely on what you want to do with it. Critical questions to ask are:

1) What sorts of activities do you want to use the computer for? Different uses have different requirements. Here is a brief table of uses and minimum desirable configurations for each.

Use	Processor	Memory	Numeric Coprocessor	Hard Disk
Word Processing	8086	640K	N	20MB
Spreadsheets	80386SX	>1 MB	Y	30MB
Statistics	80386SX	640K	Y	40MB
Database	80386	640K	N	40MB
Programming	8086	640K	Y	20MB
Communications	8086	640K	N	20MB
Graphics	80386	>1MB	Y	80MB
Multitasking	80386	4MB	Y	40MB

Because data management activities tend to be both computationally intensive and storage intensive, a minimum configuration for a primary data management computer would be an 80386, 80486, or 80586 central processor, with a numeric coprocessor and a large disk drive (>40 MB). Some form of high-capacity backup system (tape cartridges, Syquest or Bernoulli removable hard disks, or DAT tape cartridges) should also be added. Everything listed requires a hard disk and at least 640K of memory (RAM), which will let you run 98% of all MS-DOS programs. Skimping on memory reduces costs in the short term, but increases frustration in the long run.

The 80286 processor is not listed, because 80386SX-based machines approach the price of 80286 machines, and they have the potential for expansion and for support of the OS/2 and UNIX operating systems. 80286 machines lack these capabilities. Not listed in the table is the "clock" speed of the machine. The venerable IBM-PC used 4.77 MHz, but you do not want anything that runs slower than 8 MHz. For really intensive tasks (such as graphics or multitasking) higher speeds (33 MHz and above) may be desirable. Keep in mind that disk-intensive tasks, such as using databases and statistics packages, benefit much less from a higher clock speed than from a fast disk drive or a RAM disk. The "width" and speed of the data bus will also affect the effective speed of the computer for data management.

2) Where do you want to do your computing? If you do your work in a fixed location, a desktop machine with video monitor (preferably VGA color) is a better value. If you need to compute in the field, it may be worthwhile to pay the 30% extra for a portable computer.

3) How long will it be before you buy a new computer, and how much do you plan to spend on software until then? If you plan on keeping your new computer for several years, adding new software as it becomes available, purchasing an 80386-based machine may be important. The next several years will see increasing numbers of programs that require the 80386 chip. Most of these programs will be for specialized applications (spreadsheets, graphics and multitasking) rather than word processing.

4) How much help will you need in setting up your computer and how much "down time" can you tolerate? This really affects where you buy your computer and what brand of computer you buy more than what type of machine you buy. If you feel comfortable installing boards and disk-drives, mail order can be the cheapest place to buy. If you need someone in

town to help with system setup and maintenance, it makes sense to pay a little extra to establish a relationship with a local dealer.

The brand of computer is important in determining how long it will take for computer repair. Most major domestic computer companies make their own computers with standard main boards. However, some cheaper imported computers actually come from a large number of different sources, each with variant main boards. Getting main boards for such computers can take a long time (even domestic computers may take a month or more). On the positive side, hardware failures are rare and are usually confined to individual add-on boards (not the main board), making replacement easy on all brands.

Choosing software is an art in itself that is highly dependent on the scope and difficulty of the computing tasks in question. Surveying the computer magazines for software reviews and consulting with user groups is the best source of detailed, current information affecting software selection. These software packages were recommended by attendees at the Data Management Workshop.

Word Processing:	WordPerfect, Microsoft Word
Statistics:	SAS, SYSTAT, STAGraphics, PSS-PC
Database:	SAS, DBASE (III and IV), Paradox, Foxbase
Graphics:	Sigmaplot, SAS, Harvard Graphics
Spreadsheets:	Lotus 1-2-3, Excel, Quattro
Utilities:	386Max, Norton Advanced Utilities, XTREE

SELECTING A MACINTOSH SYSTEM

As with MS-DOS machines, the type of Macintosh you need depends on your computing needs and your working environment. Critical questions include:

1) What tasks will you be using your computer for? Different uses have different requirements (suggested minimums are shown):

Use/task	Processor	Memory	Co-processor	Disk
Word Processing	68000	1-2 MB	N	20 MB
Desktop Publshng	68030	2 MB	N	20-40 MB
Spreadsheets	68030	2 MB	Y	20-30 MB
Statistics	68030	2 MB	Y	40 MB
Database	68030	2 MB	N	40 MB
Programming	68030	2 MB		20-30 MB
Communications	68000	1-2 MB	N	20-30 MB
Graphics	68030	1-2 MB	Y	40-80 MB
Multitasking	68030	2-5 MB	Y	40-80MB
Image processing	68030	4-8 MB	Y	>80MB
GIS	68030	4-8 MB	Y	>80MB

Apple's release (at the time of this publication) of its System 7 operating system will require a minimum of 2 MB of random access memory. Upgrade to System 7 is not essential for simple computing, but if you have two or more Macintosh computers connected to a LAN, all must have the same System 7.0 printer drivers. Multitasking is possible using System 6.0X with Multifinder and 1 MB of RAM, but 2 MB is the practical minimum. System 7 has multitasking built-in. Both operating systems may coexist on machines connected to the same local area network.

Given that data management tasks at field stations tend to be computationally and storage intensive, the recommended minimum configuration for a primary data management computer would be a 68030 machine, such as a MacSE30, a Mac Usi, or Mac lici, with 4-5 MB of RAM, a built-in numeric coprocessor, and a 40 MB hard drive. RAM costs have dropped to the point where 4 MB of RAM from a mail order house can be added to a 1 MB machine for approximately $175, installed, resulting in a 5 MB machine. The extra RAM can significantly speed processing by reducing hard disk read/write cycles. Forty-five MB Syquest removable cartridge drives are ideal for backing up and archiving files. Image processing or GIS work will require a 25 MHz Mac lici, and preferably a 40 MHz Mac IIfx. Accelerated video display boards will vastly increase the speed of display and data analysis.

The MacClassic and MacSE, with their 68000 and 68020 processors may be fine for word processing, student use, or data entry—as opposed to data analysis—but the 68030 machines will enjoy a longer time to obsolescence. As with MS-DOS machines, disk-intensive tasks, such as data base and statistical analyses, will benefit from a fast disk drive.

2) Where do you want to do your computing? If you work in a fixed location, or you need a larger monitor than the 9-inch built-in a MacClassic or a MacSE, you will need a desktop machine with a video card and external monitor. The selection of Macintosh portables is much smaller and the costs much higher than in the MS-DOS world. For simple word processing, spreadsheeting or data logging

in the field, consider an inexpensive MS-DOS clone portable, a Radio Shack portable, or a Z88 used in conjunction with Laplink or MacLink Plus file transfer and cable packages. Data and graphics analysis can then be done on your office machine with the uploaded data.

3) How long do you plan to keep your computer before upgrading and how much do you plan to spend on software until then? In general it is cheaper in the long run to buy a more sophisticated machine initially than to upgrade at a later date. Buying a 68030 machine will give you a longer usable lifetime for the machine. An SE30 is the cheapest 68030 machine, but it has only one slot for an add-in board such as an external video board. The Mac lisi currently provides the greatest combination of low cost, expandability, functionality and ease of access and repair. The 68030 is also compatible with Apple's version of the UNIX operating system, AUX 2.0.

4) How much reliability, service, and support do you need for your system? Buying your CPU (Central Processing Unit) and peripherals from an Apple authorized dealer gives you one stop shopping, and subsequent service, but Apple limits its warranty to one year, and the quality of post-purchase service and support varies significantly from dealer to dealer. Buying your peripherals from third-party vendors can earn you 2-5 year warranties and often improved support, but at the expense of having to deal with multiple manufacturers and/or dealers. Research the support and repair programs of each purchase carefully. Local Macintosh user groups and bulletin boards are good sources for this information.

Software selection is highly dependent on the scope and difficulty of the computing tasks in question. Again, user groups, bulletin boards (such as ZMAG on the CompuServe Information Network) and magazine reviews are excellent sources of current information. The following software packages were recommended by attendees at the Data Management Workshop and/or were given high rankings among 1000 Macintosh products evaluated in MacUser 7(8):135-220.

TASK	SOFTWARE
Word Processing:	Word, MacWrite II, WordPerfect, TeachText, WriteNow
Page Layout & Desktop Publishing:	PageMaker, Framemaker, QuarkXPress, Fast Forms
Desktop Presentation:	Persuasion, PowerPoint, More
Multimedia:	MacroMind Director, Media Tracks, MacRecorder Sound System, Audiomedia
Hypermedia:	HyperCard, Reports
Spreadsheets:	Excel, WingZ, Works, Parameter Manager Plus
Statistics:	SYSTAT, DataDesk, SPSS, Statview II, IMP
Graphing & Charting:	DeltaGraph, KaleidaGraph, Igor, MacSpin (see also statistics & spreadsheet programs, above)
Mathematical Equation Writing/Solving/ Modelling:	Mathematica, Theorist, Expressionist, Stella, Extend
Data Acquisition & Lab Instrument Interface:	LabVIEW 2, MacADIOS, MacLab
Flatfile Database:	FileMaker Pro (quasi-relational), Borland Reflex Plus (quasi-relational), DAtabase
Relational Database:	4th Dimension, FoxBASE+/Mac, Omnis, Panorama, Double Helix
Bibliographic Database:	EndNote & EndNote Plus, EndLink
Communications & Multi-platform Connectivity:	Microphone II, White Knight (Red Ryder), SmartCom II, VersaTerm Pro, Kermit, TinCan, ZTerm, MacTerminal, Timbuktu, MacLinkPlus/PC, LapLink Mac III
Networking & E-Mail;	AppleShare, MacTOPS (small networks, primarily), Novell Netware, Microsoft Mail, QuickMail
Multitasking:	System 7.0, MultiFinder (System 6.0x)
Paint/Draw Graphics:	Canvas, Illustrator, Freehand, MacDraft, Mac Draw, MacPaint, Studio/I & Studio/32, Super 3D, Swivel 3D
Image Processing:	Pixel Paint, Photoshop, Image (National Institutes of Health shareware). Digital Darkroom, Spyglass View/Transform/Dicer
CAD:	Claris CAD, MiniCad+3.0, VersaCad, Ashlar Vellum
GIS:	MacGIS (U. Oregon), MacGIS (Cornell), Map II, ESRI ArcView (for download & display of ArcInfo data & images on a Mac), Business File Vision/File Vision IV (a poor-man's quasi-GIS)

APPENDIX F
WORKSHOP SURVEY QUESTIONNAIRE

Workshop on Data Management for Inland and Coastal Field Stations
April 1990

Pre-Vorkshop Survey Questionnaire
November 1989

Use the enclosed envelope to send your responses to:

Data Management Workshop
V.K. Kellogg Biological Station
Michigan State University
3700 East Gull Lake Drive
Hickory Corners, MI **49060**

Please reply by December 11, so we can make your responses available to the planning group which expects to meet later in the month.

The questions are somewhat open ended, based on the assumption that the most useful information you can give us won't fall into neat categories. Please feel free to add explanatory comments, using additional sheets of paper if necessary.

If you have questions, please contact John Gorentz at the above address, or at 616 671-2221, or by electronic mail at gorentz@msukbs (Bitnet), jgorentz@Iternet.cfr.washington.edu (Internet) or J.GORENTZ (Omnet).

1. Questionaire respondent(s)

   Institution:

   Date:

   Name(s) and posltion(s):

   Hailing address:

   Phone:

   Electronic mail:

2. Your site and institution

   To inforo workshop planners who are unfamiliar with your site, please include copies of any brochures or materials describing your site, it's facilities, habitats and ecosystems, types of research, level of activity, and other programs,

3. **Data bases**

   a. Does your site have databases that have been compiled specifically for general use (e.g. species lists, meteorological data)? If so, please list some examples.

   b. Does your site have databases originating in individual research programs, that are or could be developed into general-use resources. If so, give a few examples.

   c. Does your site have computerized records consisting of non-traditional forms of data, e.g. acoustic records, maps, visual images.

4. Administration and Personnel

   a. Where does the impetus for data management arise (e.g. site administrators, interested faculty members, research programs, technical staff)?

   b. Does your site have a data manager, or other person(s) with designated responsibility for data management?

   c. What personnel are involved in data management (number of persons, positions, training, experience, fraction of time)?

   d. How is data management funded? Is there a specific budget for data management? Is it funded at the site/institution level, or on individual grants?

5. Availability of data

   a. How can a person find out what data are available at your site? Is there a catalog or directory of data? If so, what information is kept, and how is it organized?

   b. How do you weigh investigators' "proprietary" rights to data against the goal of wider availability? Is there security against unauthorized use of data?

   c. Does your site have standardized quality control procedures for data?

6. Goals and Objectives

Below are items that could represent data **management** goals for your Institute or field station. Following each item, circle the status code(s) that best describe your site in relation to the goal.

Status Codes

ACF ° An accomplished **fact**	VIP a Work is in progress
HPG = High priority goal	SKF = Seeking funding for this goal
HPG = Medium priority goal	
LPG = Low priority goal	NPL = No plans to do this

a. Implement a central catalog or directory describing all data sets on natural habitats (i.e. data about data, computerized or not).

Status: ACF **HPG** HPG LPG VIP SKF NPL

b. Implement a central catalog or directory of data about data that is electronically searchable, "on-line".

Status: ACF HPG MPG LPG VIP SKF NPL

c. Manage selected databases for general use as a site/institutional responsibility.

Status: ACF HPG HPG LPG VIP SKF NPL

d. **Implement** a standard format for all research data.

Status: ACF HPG MPG LPG VIP SKF NPL

e. Manage working copies of all data in a unified, on-line database. (This doesn't necessarily mean a **"centralized"** database.)

Status: ACF HPG MPG LPG VIP SKF NPL

f. **Implement** an archive or repository for all historical data on natural habitats.

Status: ACF HPG MPG LPG WIP SKF NPL

7. Facilities

a. What facilities (computers, hardware, software) are the most important to your data management **system?**

b. List any electronic mail or other network links your site has to the outside world.

8. Evaluation

a. What have been your most important data management accomplishments?

b. What things would you now do differently, if you had then to do over? What suggestions would you give to other sites?

c. Vhat personnel resources do you think are needed to meet your data management goals? Are these resources now available?

d. What additional facilities crucial to your goals (hardware, software, etc.) are lacking?

e. Where do you think additional funding is most needed?

9. Other comments on data management not covered in the foregoing:

APPENDIX G

DATA MANAGEMENT AT BIOLOGICAL FIELD STATIONS

Report of a Workshop
May 17-20, 1982
W.K. Kellogg Biological Station
Michigan State University

(reprinted)

Prepared for
National Science Foundation
Directorate for Biological, Behavioral, and Social Sciences
Division of Biotic Systems and Resources
Biological Research Resources Program

TABLE OF CONTENTS

PREFACE

This report presents the results of deliberations at a workshop held in May 1982 to address what is perceived as a general problem of omission at field research sites—that of data management. Data management has not had a very high priority at most established field research stations and only recently has there been a coordinated effort to develop data management systems among sites identified in the NSF-supported Long Term Ecological Research network.

Field stations and ecological reserves have some common problems regarding data management and could benefit from joint efforts. This is not to suggest there be identical data management systems at the sites, or that there be centralized management of data. Rather, data management systems should be compatible. It is particularly desirable that there be certain standardized features which would make it easier for researchers to access and use data bases at field sites. An effective data management system can contribute to research efficiency and is deserving of more attention if field stations are to be effective in support of ecological research.

The concern for development of data management systems at field stations was communicated to the Biological Research Resources Program of the National Science Foundation in June 1981 together with the suggestion that a meeting be organized to discuss the general problem. Encouraged by a favorable response, a small ad hoc planning group was convened during the 1981 AIBS meetings at Indiana University. The elements of a draft proposal for support of a data management workshop were developed. These were amplified and finalized by a coordinating group from the Kellogg Biological Station with the continuing counsel of a formalized Planning Committee.

Participants in the workshop were selected to include data managers and research scientists representative of biological field stations of the United States. These included university facilities as well as those operated by private institutions and federal agencies. Participants also included representatives from The Nature Conservancy, the Association of Systematics Collections and the National Science Foundation.

The workshop was organized around four general topical discussion areas (cataloging of data, administration of data, computers and software for data management, and intersite exchange of information) that were addressed in some detail in site reports from selected stations. The members of the Planning Committee assumed responsibility as co-leaders and discussants for the four working groups that were established. These working groups developed preliminary materials that were integrated in the draft report. The report went through a lengthy process of editing, review, and re-writing, being sent out twice to all workshop participants for review. The co-leaders continued to provide counsel and further inputs as the report was finalized.

John Gorentz is deserving of particular recognition and thanks for his diligence in coordinating the report through its various revisions and his overall efforts that have resulted in this publication. Also, Steve Weiss provided some especially thorough critiques of each draft of the report.

George H. Lauff
Director for Education and
Biological Science Programs
Kellogg Biological Station
Michigan State University

Planning Committee and Working Group Co-Leaders

Cataloging of Data

John Gorentz
W.K. Kellogg Biological Station

Greg Koerper
H. J. Andrews Experimental Forest

Computer and Software Systems

Marvin Maroses
Belle W. Baruch Institute

Steven Weiss
W.K. Kellogg Biological Station

Data Administration

Paul Alaback
H. J. Andrews Experimental Forest

Michael Farrell
Oak Ridge National Laboratory

Exchange of Information Between Sites

Melvin Dyer
Oak Ridge National Laboratory

G. Richard Marzolf
Konza Prairie

INTRODUCTION

Biological field stations and their habitats are a unique and valuable resource for ecological research and education, especially so because of the wealth of data on those habitats. Many field stations want data management systems that will make those data more widely available to other researchers at their sites, as well as to the entire ecological research community, and thus make their facilities, habitats, and data even more valuable.

We are now at a crucial point in the development of those systems. Some field stations already have data management systems in use, albeit undergoing much further development. But most stations are in the initial stages of planning or development, and are looking to those with experience for guidance. It is desirable that all systems be able to work together in a compatible manner to serve the entire ecological research community, and it is desirable that field stations take advantage of each other's experience.

To foster the development of coherent data management systems, the National Science Foundation (Biological Research Resources Program) sponsored a "Workshop on Data Management at Biological Field Stations," held May 17-20, 1982 at the W.K. Kellogg Biological Station of Michigan State University. This workshop brought together data managers, researchers, and site directors from university affiliated biological field stations and other sites and agencies (listed in the appendix) with a similar interest in data management. These persons developed guidelines and recommendations for data management systems of high quality that could be compatible among the many field stations. Their work began prior to the workshop, when many of the participating sites prepared a written report of the current status of their data management systems and their plans for the future. These reports served to familiarize the participants with each other's activities. At the workshop itself, presentations and discussions were grouped into four categories: 1) administration of data, 2) cataloging and documentation of data, 3) computers and software for data management, and 4) intersite exchange of information. A working group for each of these four topics was formed, each participant joining one of the groups. This work at the workshop and subsequent to it, as well as material from the site reports, is the basis for this report.

Data management means different things to different people, so some comments on the scope of this report are- in order. It places emphasis on computerized data management, but much of it also deals with a degree of data management that should take place at every field station, whether or not computers are used. All data, computerized or not, should be made known and accessible to the research community. It is, of course, the increased use and accessibility of computers for research that has stimulated interest in data management. There are now tools that make it practical for a researcher to amass large amounts of data, which in turn necessitate greater attention to orderly means of care for them. Also, technological developments now make it possible to develop efficient information systems to help researchers locate and obtain existing data sets. However, it is also possible for sites to do some types of data management with very modest computing resources (at least to get started) and such possibilities are also considered.

Systems to provide for greater sharing of data call for a certain amount of coordination among field stations, and were the primary motivation for the workshop. However, they cannot be properly developed without also devoting attention to the more general topics of research data management and other uses of computers in the research environment. Secondary use of data will be most successful where data are managed well for their primary purposes. This report considers data management issues unique to biological field stations as well as some more general data management topics.

Because needs and resources differ from site to site, strategies of data management rather than tactics are emphasized. For example, it is not possible, nor even desirable, to recommend particular computers and software. Decisions about computers and software cannot be made until objectives are clear. Therefore, this report gives guidance in drawing up objectives. Then, assuming that some objectives are common to most biological field stations, recommendations and guidelines are given. These are explicit where appropriate, but on some topics the recommendations take the form of lists of factors and features that ought to be considered when designing procedures and databases, and selecting software. Some distinction is made between the essential and the desirable. It is expected that through a discussion of rationale, this report gives more practical guidance than if specific products were named.

Data management goals are described in Chapter 1. Three common perspectives are discussed, so that

with an understanding of the sometimes differing viewpoints, we can build systems of mutual benefit to all researchers. Chapter 2 presents several types of databases and their data management needs, ranging from individual researchers' data sets to comprehensive databases of all data and supporting documentation at each site. Chapter 3 is devoted to software tools that deal with these databases. Chapter 4 discusses administration of data, although this topic is also addressed elsewhere throughout the report, especially in Chapter 2 where administrative issues specific to certain types of database are treated. Chapter 5 considers several types of exchange of data between sites. These chapters correspond roughly to the four working groups at the workshop, but because the issues are so interrelated, the contributions of all the working groups (and especially the group on administration) appear throughout the report.

SUMMARY OF RECOMMENDATIONS

The recommendations of this report are summarized below. They are addressed to biological field stations and institutions that manage ecological data, and to the National Science Foundation and other funding agencies. They are grouped into four sections: A) perspectives and major conclusions, B) managing databases for primary and secondary use, C) computing facilities and software, and D) methods of continued cooperation. (These sections do not necessarily correspond to the chapters of the report. The numbers in parentheses refer to pages in the report where the issues are discussed.)

Section A

The following recommendations serve to define the perspective of this report, and summarize the major conclusions.

A1 **Data as a resource:** Existing data on habitats at biological field stations should be treated as a valuable, irreplaceable resource. Biological field stations should make these data known and readily accessible to the ecological research community. (7-8)

A2 **Data management perspectives:** Data management systems should be planned so as to benefit both primary and secondary users of data. They should serve not only to improve research support at biological field stations, but also to make data more usable and accessible to secondary users of data at other field stations and institutions. The sometimes conflicting viewpoints of these different types of researcher and institution should be reconciled, so that their data management practices complement and reinforce each other. (7-9, 34)

A3 **Data** management network: Biological field stations and other institutions that manage ecological data should be viewed, not as isolated entities, but as nodes in a data management network. This network should provide efficient means of: 1) communicating information about data sets, and 2) exchanging data. Although it need not consist of computer links, it should be a distributed database. That is, data should be stored and cared for locally, but accessible from every node. (8, 37-39)

A4 **Data management agencies:** Data management at two types of institution warrants financial support: 1) All biological field stations should be supported in their efforts to care for data about their habitats, and 2) a small number of central, secondary institutions should be supported to manage data and/or information about data that originates at other field stations. These secondary institutions (so named because they deal with secondary use of data) should be designated on a regional or topical basis. They might be biological field stations, federal agencies, or other organizations that already have responsibilities related to environmental problems or biological disciplines. (7-8, 37-39)

A5 **Types of data management:** To avoid confusion, plans and proposals should distinguish among four different types of data management, dealing with: 1) research data analysis, 2) compilation of databases for general use, 3) data directories and catalogs, and 4) data banks. (9-22, 38-39)

A5a **Research data analysis:** Computing and data management facilities for research data analysis at field stations should be given strong support, since good data management practices by primary users are a necessary precursor to secondary use, both within and among field stations. (7-12, 14-16, 18-19, 23-30, 32-34)

A5b **Databases for general use:** Biological field stations should compile data for general use, such as comprehensive species lists, lists of research sites, and meteorological databases. Some of these should also serve as directories or indexes to study sites, data sets, and publications, and as the basis for merging related data. Other databases, more general in scope, should be compiled by selected secondary agencies. They include databases covering large geographic areas, comprehensive taxonomic databases, and ecological thesauruses. (12-14, 19-22, 37)

A5c **Data catalogs and directories:** Information services that help researchers locate and obtain data sets should be developed. At each field station there should be, at minimum, a directory to the data sets. Selected secondary institutions should serve as central sources of information about data available at field stations (and elsewhere). (8, 17-18, 25-26, 34, 37-39)

A5d **Data** banks: Data banks should be established to maintain (at least) those data sets that have no other means of long term care. Each bio-

logical field station, whenever feasible, should have such a repository. In addition, secondary agencies should be designated as repositories for data that cannot be cared for at the local level. (8, 18-19, 37-39)

Section B:

The following recommendations pertain to administering data for maximum usefulness for both primary and secondary puposes.

B1 **Data managers:** Each field station should have a data manager responsible for the care of those data to be managed as a station's resource. Data managers should have expertise in ecological disciplines, data management, and computer technology. To ensure coherence and continuity, a data manager should be funded directly by the field station and should report to the top administrative level of the station. (31)

B2 **Support for relevant data sets:** Field stations should identify those research data sets that have a potential for secondary use, and provide researchers with tools, services, and incentives to maximize their usefulness to others. (7,14-17, 18-19, 32, 34)

B3 **Documentation of data:** All data available for secondary use should have full, easily accessible documentation. This documentation should include both the scientific and the technical details needed to decipher the data. It should be complete enough to permit the data analyses as well as the data collection procedures to be reproduced. (Specific recommended categories of documentation are listed in Tables 1-3.) (7-8, 11-12, 14-19, 25-26, 34, 37)

B4 **Integration of databases:** So that related data sets can be brought together for analysis, they should be made consistent and compatible with respect to (at least) site, taxonomic names, and topic. Consistent coding schemes, indexes, master site lists, master species lists, and other means should be employed. (7, 17-23, 37)

B5 **Centralization vs. decentralization:** Where possible, data management functions should be left in the hands of the owners and originators of data. At the same time, there should be centralized means of access to data (through centralized directories and information services). This principle should be applied to relationships between researchers and data managers at field stations, and to relationships between field stations and secondary data management agencies. (16, 18-19, 37-39)

B6 **Redundancy control:** Data management for secondary use should avoid redundant copies of data sets, since redundant copies tend to become inconsistent when additions or corrections are made. (The distributed database approach is preferred.) If copies of data are needed (e.g. for a repository), care must be taken to ensure that they are up to date and consistent with the copies in the hands of the contributing researchers. (13, 19, 27-28, 37)

B7 **Error checking:** Rigorous error checking of data should be encouraged and (where appropriate) enforced. The procedures used should be noted in a data set's documentation. (11-13, 19, 22-25, 32, 39-41)

B8 **Review procedures:** Data and documentation should be reviewed periodically to keep them up to date. (16, 19)

B9 **Documentation of data management:** To ensure continuity, a station's data management policies, decisions, and procedures should be documented (and publicized). (19, 34, 35)

Section C

The following recommendations pertain to computing facilities and software. Some will have to be treated as long range goals, since they might not be practical at present. They are all, however, consistent with current trends in computer hardware and software capabilities.

C1 **Software strategies:** Rather than build software systems "from scratch," biological field stations should, where possible, use software that is already available. It will often be necessary to use several software packages or components in order to meet all needs, but these should be made to work together consistently for ease of use. (28-30)

C2 **High level tools:** The high level data analysis tools that are available should be used to: 1) provide standardized methods for manipulating data, 2) make documentation easier, and 3) free the researcher from the need to deal with tedious details (12, 14, 18-19, 23-28)

C3 **Record keeping tools:** The tools that researchers use to analyze data should also help them to document those data. Record keeping tools should work consistently with data analysis tools, and should also assist researchers with other record keeping needs in addition to those associated with computerized data. (7-8, 12, 14-16, 18-19, 23-26, 28)

C4 **Data entry systems:** Data entry systems should be used that 1) capture supporting documentation at an early stage of data set development, 2) help researchers use consistent, compatible coding schemes, and 3) enable researchers to use rigorous error checking procedures. (11-12, 23-25)

C5 **Data dictionaries:** Because of its central role in managing documentation and in linking related data, data dictionary software (whether or not it goes under that name) should 1) be able to handle textual as well as other types of data, and 2) be usable directly by researchers as well as have interfaces for use in data entry (and other) software, and 3) have indexing and cross-referencing capabilities. (25-27)

C6 **Computing facilities:** To support decentralized data management, computing facilities should be practical for use directly by researchers. They should be accessible, interactive, and easy to use. They should help to integrate data management with all other facets of research. Charging policies (where needed) should not discourage their use. Equipment should be selected in light of software requirements (not vice versa). (11-12. 18-19, 23-30, 32-35)

Section D

The following recommendations pertain to means of continued cooperation between field stations, and between field stations and secondary data management agencies.

D1 **Data exchange protocols:** Researchers can make data known for secondary use via directories and catalogs. Researchers who make data available can stipulate that their data can be obtained and used by permission only. Co-authorship or prominent acknowledgment should be given for the use of data. Channels of communication should be developed by which researchers can receive feedback on the use and utility of their data for secondary purposes. (7-8, 17-18, 37-41)

D2 **Compatibility:** Field stations and other data management agencies should strive to be compatible with each other in all areas affecting intersite exchange of data. Examples are the organization and indexing of documentation, catalogs and directories, and the identification of taxonomic groups within data sets. Also, all field stations and secondary agencies should have facilities that permit them to send and receive data in a "normalized" form, with standardized documentation. (9-11, 14-16, 17-18, 20-22, 37-41)

D3 **Informal communication:** To achieve consistency and standardization, field stations should advertise their successful projects to each other, through newsletters or other such means. Data managers should keep informed so that they can consider systems in use at other field stations when developing their own. (18, 42)

D4 **Formal communication:** In addition to informal communication, some formal means of communicating data management ideas and develop ing compatibility standards warrant support. These include 1) a national newsletter, 2) conferences and workshops (perhaps in conjunction with meetings of professional societies), and 3) consulting services and courses in scientific information management. (41-42)

CHAPTER 1
VIEWS OF DATA MANAGEMENT

THE PERSPECTIVE OF BIOLOGICAL FIELD STATIONS

Data management, as an activity supported by biological field stations, is a means toward furthering their objectives of education, research, and habitat protection. By making the existing data on their habitats accessible and usable, their facilities and habitats can become more valuable resources. Their wealth of existing raw data constitutes an irreplaceable record of habitats and populations. Many of these data form long term records, and if preserved, can be used for novel applications in the future. Bringing together all the information on a site in a coordinated fashion can foster the further development of ecological science by making the site more useful for new research. Researchers can make plans with the confidence of knowing that they have available all information about a site. New research can proceed without getting bogged down in the collection of background information. A station's data and habitats can become a resource available for studies on a regional and national scale.

It is not enough that the data exist. They must also be accessible, but the current state of affairs is such that they usually are not. There exist few good systems to help researchers find all the data sets about a given habitat or taxon at a site. There are few systems that help researchers locate habitats on the basis of ecological characteristics, even though the data that could form the basis for such searches often do exist. Sometimes data sets can be located, but they usually do not have the necessary documentation to make them useful. Sometimes poor data management practices on the part of researchers make it difficult for others to use their data. And even if researchers organize their data well, there is no systematic means to care for the data past their lifetimes. These are all obstacles to greater and more efficient use of habitats and associated data.

In order to remedy this situation, many biological field stations wish to develop systems for maintaining information in an accessible form. They wish to compile species lists, meteorological information, and other databases for general use. They are concerned that the data collected by individual researchers be available to a wider audience, so that their sites are also useful to a wider audience. Also, in many cases, field stations wish to provide computer services, both to assist data management and to enhance capabilities for research.

In a sense, there is already a well established and systematic data management scheme in place for ecological (or any other) research—in the form of the scientific literature. However, biological field stations can bring together not only the published data, but also the unpublished data and the data behind publications on particular habitats for more efficient research on those habitats. It is this link between data and habitats that makes data management at biological field stations a unique concern.

A RESEARCH PERSPECTIVE

Researchers already at work on a site tend to view data management in a somewhat different light. While a biological field station's primary concern is facilities and habitats, a researcher's primary concern is his or her own research program. Researchers view data management as a means to more efficient data collection and analysis. They give high priority to tools such as statistical and graphics packages which help them analyze data efficiently, and a lesser priority to systems whose purpose is to make their data accessible to other researchers. This is not because they oppose the furthering of ecological research by this means, but because limitations of time and money force them to set other priorities.

This attitude is not a complete hindrance to data management. On the contrary, data management should always be a servant to data analysis. Data are managed to make them accessible and usable, but a system is of little use if it only enables good organization of data, but does not permit analysis of data. Whether for secondary or primary users of data, data management is a means to better data analysis.

The data management tasks done in the course of a researcher's own analyses have much in common with the tasks necessary to make data available to a wider audience. While it is sometimes possible for a lone researcher faced with the pressures of publication to do data analysis without good data management practices, in general, poor practices and tools waste time and money. If one takes a large volume

of data and multiplies it by many complex analyses, the result is the need for a lot of record keeping. Researchers need to record information about data items, data files, updates of data files, procedures, and results, so that they can know exactly how each data file, variable, and "piece of output" came about, and what its current status is. In short, they must be able to reproduce every analysis they do.

Researchers need to keep track of these things, but it is extremely time consuming and clumsy to record all the necessary details manually. If they are able to get by without complete record keeping, it will be to their future disadvantage. However, a secondary user needs to know these details just to get started. Efficient ways are needed to keep this documentation.

Researchers at field stations value their time in the field. They do not want to waste time with clumsy data processing systems, whether the clumsiness results from having to go through human intermediaries, from inaccessibility, from poorly designed computer systems, or from good computer systems that work together poorly. They want to spend their time doing research. Data management systems which help them be more efficient will also make their data more accessible.

THE PERSPECTIVE OF SECONDARY USERS

Some research investigations, such as those on a large temporal or spatial scale, can benefit from, or must rely on, data obtained from other research at their own or other sites. Three types of data exchange are 1) simple personal communication of data between two researchers, 2) collaborative research among sites as in the Long Term Ecological Research (LTER) program, and 3) research on problems of a regional or national scale requiring data from a large array of sites.

The first type of exchange is a horizontal information transfer across (or within) sites driven by the interests of individual scientists. It has and will continue to be served by the scientific literature, meetings and symposia, and personal contacts among researchers, but it can be made more efficient through good data management practices and by computer aided methods which can increase researchers' awareness of data available at field stations.

The second type of exchange is done on a larger scale. It differs from the first in that it involves not only data management, but cooperation in making data sets compatible through common or comparable measurement techniques. Whereas the first type of exchange involves data which happen to be comparable or otherwise useful, an expressed intent of the second type of activity is to do comparable research. Efforts to bring multiple data sets from multiple sites to bear on particular topics has sometimes been prompted by common interests among individuals and groups. On the other hand, there have been quite formal studies launched by (for example) the U.S. Forest Service, the Department of Interior, the National Science Foundation, and the National Academy of Sciences. These modes of study will continue, and can be aided in the future by computer aided data management, analysis, and communication.

The third type of exchange is driven by the need to research environmental problems of public concern on a large geographic scale. These problems include national and regional issues such as air pollution, acidic precipitation, and water quality. Such research relies on data from a wide array of geographic, biotic, economic, and political provinces. An expedient mechanism is needed to locate and obtain from field stations such existing data sets as might contribute to this research. Such a mechanism might also focus the attention of ecologists at field sites on these problems, and might stimulate research in theoretical and applied ecology which will assist in the management of natural resources.

In the past decade various large databases on environmental subjects on a large geographic scale have been developed. Examples are the Oak Ridge National Laboratory (ORNL) Geoecology Database (Olson, et al. 1980) information systems on fish and wildlife species developed by the Departmentof Interior and information about ecological and environmental data summarized under The Institute of Ecology's ACCESS program for the Department of Energy. Within reasonable time limits and with reasonable resources, study teams can assemble moderate to low resolution assessments for regional and national issues.

Environmental issues drive research at both site and regional or national levels. Yet there are differences in the way data are acquired. Researchers doing site level work use their own data or sometimes data that are available from local repositories (e.g. data banks). These are instances where collaborative research between sites has motivated the exchange of data. However, research on environmental problems on a large geographic scale, most likely operating out of regional or national centers, requires the knowledge of the existence of data sets and information, and the ability to obtain such information. Exchanges among field stations and between field stations and national, regional, and topical agencies are all needed.

CHAPTER 2
DATABASES

A diversity of data is collected at biological field stations. These data are in many forms, such as maps, specimens, charts, field notes, microfiche, and computerized textual as well as numeric information. Some, such as climatic data, are of immediate, obvious utility to a great number of researchers. Others, while seemingly more esoteric, are still of potential value to other research in the future. Some data sets are applicable to a large geographic area, while others may pertain only to processes or species at one field station. They include both long term and short term records. The former obviously require long term management to be useful, but the latter do also if they are to be useeful beyond their original purpose.

The databases discussed in this chapter include data sets compiled by individual researchers for their individual use as well as those developed by field stations for general use. They include not only data in the usual sense, but databases of data about data, such as directories and catalogs of data, and documentation. They deal with some data that are computerized and some that are not.

This chapter first focuses on how to manage individual data sets for efficient analysis, and progresses to a discussion of how to manage them together as a coherent whole, with consistency and long term care.

DATA SETS

A data set carefully managed for its primary purpose will also be more useful to others. Thus, although the originator of a data set will place priority on immediate data analysis needs, this is not necessarily at odds with long term data management goals. The cooperation of researchers is essential to building up a complete, well documented database. Researchers can be more easily convinced to provide well documented data to a station's database if they, in exchange, can be offered tools and services that help them do data analysis and keep good records. The extra documentation and management needed to make data available to secondary users are simply an extension of what researchers need to do for their own purposes, not a different kind of data management. If a researcher's data set is well managed for him, it will take less extra work to incorporate it into a station's database. Therefore, this section discusses how to use data management to help researchers analyze their data.

Data Organization annd Structure

One of the first steps in managing a data set is deciding how to organize the data. Some organization is of course necessary in order to store data on a computer, but even before that point some decisions about data organization are needed to design data recording forms and data entry procedures. Time and money can often be saved by deciding these things as early in the project as possible.

By organization, we refer to that which is known in database technology as the "logical" structure of the data. For example, an animal behavior study might include various types of observations of behavior, as well as information about the different habitats and meteorological conditions under which the behaviors occurred. It is necessary to decide what all the behavioral, habitat, and meteorological variables are, how they should be organized into different types of records, and how the variables and records should be arranged with respect to each other. Hierarchies of data should be delineated.

A concept that is very useful in organizing any data set is "normalization." It is a simple, straightforward way of structuring data. It is also a "common sense" approach, in that many persons have by trial and error arrived at major elements of the scheme.

Although the steps of normalizations have exact definitions (e.g., Martin 1977), we will deal here only with a simplified version of it. We can normalize a data set by asking two questions: "What are the types of entities about which we have data?" and "What data do we have about each type of entity?" For each type of entity, a table (or file) is made. Each table is a two dimensional matrix of rows and columns, in which the data about an entity make up one row.

As an example, consider some of the data collected in the National Atmospheric Deposition Program (NADP). These data include pH and conductivity measurements, other chemical parameters, daily rainfall measurements, descriptions of each site, and information about instruments used. If these data were normalized, they might be organized into tables, one for each of the following types of entity: I) sites, 2) samples, 3) daily meteorology, 4) instrument use, and 5) instrument maintenance activities.

Each table has a row for each entity (e.g., for each site or each sample), and a column for each variable (e.g., for each parameter, type of observation, identification code). The table of "site" data has one row for each site, and a column for each variable that is specific to a site. The table name and its columns can be denoted:

SITES:

SITE NUMBER	site name	latitude	longitude	...

One additional concept is that of a "key" for each table. In the SITE table, the key is the variable SITE NUMBER, and is thus denoted in upper case. For SITE NUMBER to be a key, it must uniquely identify each row in the table. That is, there is one and only one row for each site number. We can use its key to tell what type of entity a table describes.

A variable such as pH is not included in the SITE table. A pH measurement is not specific to a site, but rather to a particular sampling interval at a site. The pH measurements are instead included in a table of SAMPLES, which has as its key the variable SITE NUMBER and two variables that define the sampling interval (TIME BEGUN and TIME ENDED).

SAMPLES:

SITE NUMBER	TIME BEGUN	TIME ENDED	PH	conductivity	calcium	...

Note that in this table, the key consists of three variables which, in combination, uniquely identify each row. Each sample is identified by a site number, time begun and time ended (where time consists of date as well as time of day).

The information recorded on a daily basis, such as precipitation amount, belongs in its own table:

DAILY METEOROLOGY:

SITE NUMBER	DATE	precipitation amount	precipitation type	...

The above three tables closely resemble the way the NADP data are actually organized. A central register of sites is maintained, with complete information about each site. The field forms are designed to accommodate some data on a per sample basis and others on a per day basis.

Data on the instruments used are not currently kept this way, but could also be represented by normalized tables. One table could describe each instrument and when and where it was used:

INSTRUMENT USE:

SITE NUMBER	TIME BEGUN	TIME ENDED	instrument number	instrument description

To be more systematic, and ensure that certain data are recorded for each instrument, the variable named "instrument description" could be augmented with others, such as make and model number. Another possibility, probably better (but not depicted in the diagrams), would be to have two separate tables, one describing instruments, and another telling when they were used, especially if a site often switches back and forth between different sets of instruments. For some instruments, such as rain gages, a maintenance log would be useful to record calibration and winterizing:

INSTRUMENT MAINTENANCE LOG:

INSTRUMENT NUMBER	DATE	person	description of activity and comments

Persons familiar with the NADP program will note that the actual data are somewhat more complex than presented here, and would necessitate some additional columns and tables. However, the general principles can be applied no matter how complex the data set: Define the types of entities about which there are data, and the data about each type of entity.

Representing data as normalized tables is of use in several ways: 1) It is a simple scheme, yet general enough for data sets of any degree of complexity, 2) it is helpful for designing data bases no matter what database management software will be used, 3) it is useful for organizing data that will be kept on paper, 4) it is compatible with the data formats required by most data analysis software, and 5) it can be a framework for a system of data documentation.

The simplicity of normalization derives from its single, uniform structure for representing data. The concept of a table of rows and columns is readily understood, even by those unfamiliar with database technology. While a hierarchical notation might be better for hierarchical data, the normalized scheme is more general. It can represent any data set, no matter how complex.

No mattter what type of software is used, normalization helps to organize the database. In a relational data base, the data are viewed (and usually sotres) as a set of normalized tables. For a network database it can help one to determine its "entity types" and "relationships." If the data are to be stored as a hierarchy,

normalization can be used to determine what hierarchical levels there are, and what data should be stored at each level. No matter what type of database management system is used, the data should be grouped according to the entities arrived at by normalization.

Normalization is even useful in designing databases to be kept on paper. It can help in developing proper forms for recording the data. For example, if each NADP site kept instrument logs, it would point out that some data will remain constant for each instrument, but that there may be several periods of use for each instrument. The forms should be designed so that constant information need only be recorded once, and so that there is room to record several periods of use.

Dealing with normalized data is also easy if one is going to use a statistical package or other data analysis software. The data formats required by data analysis software once were quite varied, but now are rather standard. They usually require data to be in the form of the familiar table of rows and columns, with one row for each observation. (The number of rows is the familiar "n" of statistical tests). Although for purposes of analysis, several tables may need to be merged to form a large table (admittedly with some redundant information), we are still dealing with a single, uniform structure, the table.

Normalization also provides a framework for a system of documentation. It can clarify just what needs to be documented, and the documentation itself can be normalized. In the preceding example, each table and each variable should be documented. And some of the tables, such as those describing instruments and instrument usage, serve mainly to document the precipitation analyses.

A familiarity with normalization is recommended for all persons who have to manage data sets. It can help avoid some common mistakes in developing data structures.

Data Coding

Another part of organizing a data set is deciding how to represent and store variables that must be coded. The different treatments, methods, species, or sites in a data set commonly need to be represented by codes. A set of codes may be chosen to simplify the writing of data on a field sheet, to minimize keystrokes during data entry, to minimize data storage requirements, or to make for faster processing by a computer. It is less confusing if codes are consistent within a data set and between data sets. (Sometimes the consistency among codes will make a difference as to whether or not comparing two data sets is practical.)

Schemes for storing code definitions in a data set are sometimes overly elaborate. A simple approach is to store them as normalized tables. The NADP data set includes a code called site number, which occurs in several of the tables. The table of SITES lists these codes, one per row, and the other variables in that table serve to describe just what each code represents. At least one statistical software package stores codes and printable labels in normalized tables, similar to any other table. It is a conceptually clean approach to a task that sometimes has been made more complex than is necessary. Even if the available software does not lend itself to dealing with data sets that include separate code definition tables, they can at least be used as a simple, easily understood way of storing some necesssary documentation with a data set.

Codes should be as straightforward and clear as possible. For example, a variable to indicate sex might be coded as (1 = male, 2 = female), but it would be better to use the codes "M" and "F," and even better yet to use "MALE" and "FEMALE." Clear, mnemonic codes can help make the data set self documenting.

Data Entry

It is important that transcription of data from field forms to computer media be efficient and error free. Data entry is best done by a person who is familiar with the data, and is best done during the data collection process, not afterward. Errors are more easily caught while the data are fresh in a researcher's mind. A person who was involved in collecting the data will tend to catch not only transcription errors, but also mistakes on the original data forms. (No matter who enters the data, someone familiar with the data must be involved in the error checking process.) Timely data entry can also make it possible for researchers to use preliminary results to make midstream modifications to data collection procedures.

While desirable, this sort of timely, personal data entry has not always been practical. It is not the best use of personnel, for example, for a busy researcher to enter large batches of data via a keypunch machine in a remote location. However, the growth of personal computing and easily used data entry software often makes it the method of choice.

Whatever the equipment and software used, error checking deserves much attention. First of all, the original records should be scrutinized carefully before any data are entered. During the actual data entry process, there are four types of technique that can be applied. We will call them the outlier, proofreading, double entry, and checksum techniques. They can sometimes be used in combination.

By the outlier technique we mean using software to check for values outside an expected range, or not in a list of legitimate values. It can also mean checking complex combinations of variables. It is a means of ensuring that certain types of errors do not occur in a data set.

The three other techniques, by contrast, are intended to ensure that each datum is correct. Even though data have been checked for outliers, proofreading will detect additional errors. An effective technique is to have it done by two persons. One person reads the numbers aloud from a printed listing of the data, and the other confirms each datum from the original data forms. While this technique may seem inordinately tedious, it will catch many errors that one person working alone will miss.

The double entry technique accomplishes a similar result in a more automated way. Two different persons each enter the same data. and the results are compared. It can be done mechanically on keypunch machines, or by software capable of reporting differences between two sets of data.

The checksum method is similar in that it also involves "entering" each datum twice. The data forms must be designed so that the person filling them out not only has to write down the raw data, but also compute a sum (or mean) and record it on the sheet. Typically a calculator will be used for the computation; it is here that the data are "entered" for the first time. Then, when the data are put on the computer, the sums as well as raw numbers are entered, and software is used to verify that the recomputed sum matches the one that was entered. This technique is especially appropriate if the sum (or other summary) is of immediate usefulness to the researcher.

The latter three methods of verification all are labor intensive, but additional time spent at this stage of data analysis usually saves time in the long run. Errors not found until the later stages of data analysis typically cause a great waste of time and effort because many of the earlier analyses must then be redone.

Facilities that make data transcription unnecessary can be especially efficient. It is often possible for a person to enter data directly via a personal computer or terminal while examining and measuring specimens. The transcription process, a source of errors, is omitted. In this mode, it is wise to produce a printed record of the data immediately, as insurance against a possible computer failure.

No matter how thorough the original error checking, some errors may not be found and corrected until much later. In this case, keeping a revision history, whether automatically or manually, can be important. This is especially true if more than one researcher is using the data, or if some results of the analysis have already been put to use.

Record Keeping

From one point of view, the term record keeping is almost synonymous with data management. It is a type of data management that has always been done in science. However, computerized data analysis poses some additional record keeping needs, and computerized data management can improve both the old and the new types of record keeping.

A basic aim of record keeping is to ensure repeatability, not only of experimental treatments but also of data analyses. Analyses often need to be redone, because of corrections or additions to the data base, or with slight variations to previous procedures. This requires that not only each datum, but each procedure used to derive data from data must be documented.

This task is made necessary and sometimes difficult by the ease with which a researcher can do a multitude of analyses, using a computer to generate data from data. It is easy to let the record keeping lag behind. Self-documenting systems can help by storing definitions of variables, definitions of procedures used to generate derived data files, and other such documentation. A more general purpose record keeping system can also be used for information about projects, data sets, methods, files, and variables, using a combination of database and word processing technology. Managing this type of information is the topic of much of the rest of this report, since it is also needed to make data usable by anyone else.

BIOLOGICAL INVENTORIES

While many data sets are gathered by individual researchers or research teams for their own use, there are others that should exist as general resources at all field stations. But they are not likely to be compiled unless supported directly as a field station's responsibility. Some of these databases can be thought of as "biological inventories" that describe ecological characteristics of the station. In addition to constituting research databases in themselves, they are useful in education and research planning, since they can serve as directories to the populations and localities of field station. Field stations are encouraged to assume responsibility, directly or indirectly, for developing such databases.

Two typical types of biological inventory, species lists and indexes to biological collections, illustrate some special data management needs.

Species Lists

A familiar sort of biological inventory is the species list. These often take the form of printed lists arranged in a taxonomic or spatial sequence. Some species lists are intended to describe a specific habitat by listing species of special interest, while others are intended to represent more exactly the distribution and abundance of all species in a geographic area. Some are compiled by a researcher or instructor directly from observations, and are kept up to date by the same person. Others are compiled indirectly from anecdotal data, published reports, class surveys, or research data sets.

In many respects, species lists can be managed just like any other data sets, but in cases where a species list is derived, in whole or in part, from other data, there are some additional data management issues. Such a list is, in effect, a summary of other data. A summary of data, by definition, does not include all the data from which it is derived, and if only the summary is saved, the raw data are, in effect, thrown away. Data with fine spatial, temporal, or taxonomic distinctions tend to get lost in summaries. Since not all taxa or localities are likely to have been treated with original thoroughness in the source data, a lowest common denominator is usually chosen for the summary.

For this reason, it is best to maintain a link between species lists and their source data. When a person wants to locate a site appropriate for detailed research on a population or habitat, the species list may be a good starting point, but it should also refer to the source information.

Since biological communities are dynamic, species lists should be dynamic, and reflect changes in distribution or taxonomic nomenclature. This is, of course, more easily done when the species lists and the source material are computerized. In the ideal situation, using computer database technology, there would not necessarily be a species list stored as an entity in itself. Compiling the species list would consist of establishing links to the source databases (which are dynamic) so that a computer program could extract the taxonomic information to create an up to date copy of the species list. For the present, most sites will have to use less automated techniques to achieve a similar result.

It should also be noted that a species list consists of several types of information, two of which might be best treated as separate databases (which can be merged or linked with species lists as necessary), because their utility goes far beyond use with species lists. The first type is information on taxonomic relationships, such as might be manifested by the hierarchical arrangement of a printed list. The second type pertains to detailed information about each of the locations covered in the species list. Treating this sort of information separately can avoid redundant data and effort. A later section of this chapter, "Integrating Databases," discusses this concept in more detail.

Collection Indexes

The sheer numbers of specimens in biological collections, and the care required in handling them, sometimes limit the ease with which they can be examined. Computerized indexes increase the utility of such collections by making it easier to locate specimens quickly and by making some of the data inherent in the collection available for efficient analysis.

An index usually contains, for each specimen, data such as the taxonomic name, locality from which the specimen was obtained, name of collector, date collected and other information describing characteristics of the specimen. This makes it possible to search the list of specimens on the basis of location, taxon, date collected, etc. Minimal data categories are reviewed in "Guidelines for Acquisition and Management of Biological Specimens" (Lee et al. 1982).

To design a computerized index, it is necessary to decide what information is to be included, and what procedures will be used for entering the information into the database. The two decisions should not be made independently of each other.

The procedure for entering data on specimens which are accessioned after the task begins may be different from that for specimens already accessioned. For already accessioned specimens, it may be prudent to first enter one taxonomic group and then make the database available to researchers, and later add other taxonomic groups as priorities, finances, and other resources dictate. In this way, the database is used (and tested) soon after onset of the project and before commitments waver.

Entry of specimen label information involves some redundancy. Typically, the specimen label is prepared first, and then the exact same information is put on a computer. When entering data about already accessioned specimens, the redundant labor is necessary, and error checking procedures are needed to ensure that the information is transcribed correctly. But when entering data about newly accessioned specimens, redundancy can be avoided by entering the information about each specimen only once. The person who accessions the specimen can enter the information directly into the database and have a specimen label printed out (computer resources and the physical

nature of the labels permitting). This avoids a "middleman," reduces errors, and makes the process as efficient and simple as possible.

Just as with species lists, it is best if the data about taxonomic relationships and sites are maintained in separate databases, and merged with the collection list as necessary. (To the casual user, it would be best if the collection list appeared to contain all these types of data, but from a data management point of view they should be separate.)

DOCUMENTATION SYSTEMS

Documenting of data is an elaboration of some already existing practices. For example, scientific publications require descriptions of methods and materials to ensure that research can be reproduced. Researchers generally keep detailed notes of all their procedures and results.

In addition to documenting the scientific aspects of research, it is also necessary to document technical aspects of data handling, structure, and content. Every researcher knows of data that were effectively lost, not because they had been destroyed, but because there was no documentation to explain what they represented. For a researcher to use a set of data (his own or anyone else's), he must know what the numbers and codes represent (e.g., how they were derived or measured), and how they relate to other numbers and codes in a data set (e.g., which values go with which sites and treatments). Publications do not usually include such technical details, and it is not always possible to match up publications with the data files on which they were based.

Careful record keeping is necessary, but for a variety of reasons, the traditional sort of record keeping is often not adequate. There is often a temptation for researchers not to bother recording the necessary information, especially when they can generate new variables, files, and other output much more quickly and easily than they can generate the accompanying documentation. Such records as are kept are often cryptic notes in a chronological log, mixed together with other notes, and are not intended to be used in that form by other researchers.

It is necessary that both sorts of documentation, scientific and technical, be available to secondary users, and it is desirable that they be handled in an integrated fashion. Primary users (contributors) and secondary users can deal more efficiently with documentation that is in a uniform format.

In keeping with the principles of normalization that were discussed earlier, we first need to decide on the types of entities that need to be documented. Typically, there might be many data sets at a site, each data set consisting of one or more files (or tables). Each data set and each data file should be documented. In addition, each file will have several constituent variables, and some of these variables might be contained in more than one file. The variables also need to be documented. We will focus our recommendations on these three entities: data sets, data files, and variables. Although elaborations will be required at some sites (perhaps because of the nature of the data management software or for other reasons), these three represent the most important documentation needs.

The documentation that ought to be maintained for each can be organized into categories (or "fields"). Tables 1, 2, and 3 list the categories of documentation needed for data sets, data files, and variables, respectively. Documentation organized into categories like those shown is much better than an amorphous collection of notes. The systemization imposed by this structure can ensure that no important details are omitted, and also makes it easier for a person to scan the information quickly.

These categories are a composite of those now in use at some field stations. A particular field station may choose to modify this list after weighing the value of each category against the cost of maintaining it, and it may choose to use a subset of these categories, add others, or merge or subdivide categories depending on its own needs and resources. The categories for data sets are rather general, and as such are appropriate for quite diverse research data. An example of much more specific categories that apply to a narrow range of research topics is presented by Altman and Fisher (1981).

The use of higher level database languages, in which one does not need to deal with low level details such as physical positions on cards, can make documentation easier. Stations that use such software will find some of the listed categories irrelevant. Some database management systems and statistical packages enable researchers to deal with data in terms of named variables and tables rather than physical locations of data, and enable them to express algorithms in a way similar to that used in scientific writing. This sort of software not only makes data management and analysis easier, but also makes documentation simpler.

Documentation requirements can also be simplified if "coding" of data is minimized. Low level systems often require researchers to refer to their data in terms of codes, for example, in dealing with a "species" variable where 1 represents Quercus alba, 2 represents Acer rubrum, etc. With high level systems, researchers can deal with data directly in terms of species names and treatment names (even though for internal efficiency, hidden from human view, codes may be used).

Table 1. Categories of documentation for data sets.

1. **Data Set** Name	A name or code that uniquely identifies the data set.
2. Data Set Tide	A title that describes the subject matter.
3. Data Set Files	A list of the data files that constitute the data set.
4. Research Location	Information that identifies the site of the research at a level of detail appropriate to the purpose of the data set.
5. Investigator	Name of the person(s) responsible for the research or other project that generated the data.
6. Other Researchers	Names of other persons responsible for various phases of data collection or analysis, especially those who could conceivably be consulted regarding use of the data.
7. Contact Person	Name of the person to contact for permission to use the data, and for help in locating and obtaining it.
8. Project	Description of the overall project of which this data set is a part (to place it in the context of other research and to describe its purpose).
9. Source of Funding	
10. Methods	Description of methods used to collect and analyze the data, including the experimental design, field and laboratory methods, and computational algorithms (via reference to specialized software where necessary). (This category is analagous to the methods and materials section of published papers. It could easily be subdivided into other categories. The experimental design, especially, could be put in a separate category, since it can help describe the rationale of the data set.)
11. Storage Location and Medium	Storage location and medium of the data set as a whole, e.g., magnetic tape, disk files, punched cards, etc.
12. Data Collection Time Period	A description of the data collection period and periodicity, and major temporal gaps or anomalies in the data set pattern.
13. Voucher Material	Site (institution, collection) where voucher material has been deposited.
14. Processing and Revision History	A description of data verification and error checking procedures, and of any revisions since publication of the data.
15. Usage History	References to published and unpublished reports or analyses of the data that could be of interest to a secondary user.

Table 2. Categories of documentation for data files.

1. File Name	A name or code that uniquely identifies the file.
2. Constituent Variables	A list of the variables contained in the file. This list (and the information about each variable, i.e. the categories listed in Table 3) is the most important information about the file.
3. Key Variables	A list of the hierarchy of variables that determine the sorted sequence of the data, or a list of the variables that constitute the file's "key."
4. Subject	An explicit description of the subject matter of the file. It should make clear what type of entity is described by the records.
5. Storage Location	A description of the location of the file (in terms of a computer system's file naming system, where appropriate).
6. Physical Size	The number of records and total number of characters, or other such descriptors.
7. File Creation Methods	A description or list of procedures or algorithms used to create the file, and the files from which the file was derived (if applicable).
8. Update History	A record of updates to the file (where those records might **help to reconcile** differences with previous versions of the data).
9. Summary Statistics	A brief set of summary statistics (means, sums, minima, maxima, etc.) for each variable. (These can be used to verify that the data file one is using is indeed the correct version, and to verify the accuracy of data transfers.)

Table 3. Categories of documentation for data variables.

1. Variable Name	The name of the variable (which should be unique within the data set), and any synonyms which a user might encounter.
2. Definition	A definition of the variable in ecological terms.
3. Units of Measurement	
4. Precision of Measurement	(Statements about precision should not only give error bounds, but explain what they refer to. The user should know whether the variance given is that of determinations by an instrument, or among replicate samples at a single location, or among locations within a given area, etc.)
5. Range or List of Values	The minimum and maximum values, or for categorical variables, a list of the possible values (or a reference to a file that lists them and any code definitions).
6. Data Type	A description of the variable, in terms like "integer," "date," "4-byte real," or whatever others are used by a database management system (DBMS) or statistical package. (This information is needed when dealing with data stored in the special formats of a DBMS or statistical package.)
7. Position and/or Format	Any information that will be needed by a program in order to read data from (for example) an ASCII file. (This information is typically needed in a non-DBMS environment and is almost always needed for data transfer between sites.)
8. Missing Data Codes	A list of codes that indicate missing data. If there are several types of missing data codes, they should be distinguished.
9. Computational Method	Algorithms that were used to derive this variable from others (if applicable).

Data that are not coded are much more self-explanatory, and require less additional documentation.

The degree to which documentation is computerized will vary from site to site. Much of the documentation of variables, and some documentation of files, is handled more or less automatically by some data analysis software. However, it is important that both computerized and uncomputerized data be documented.

Software to support documentation is discussed in more detail in Chapter 3, under "Data Dictionaries." However, it is not likely that complete documentation will always be stored in a computerized database, and it will also be necessary for computerized documentation to refer to supporting materials that are stored elsewhere, such as extremely lengthy and detailed descriptions of methods, original data sheets, maps of the site, photographs, and so forth. The computerized portion of the database of documentation should, for each category, either contain the necessary information, or explain to the user where it may be found. It may be best to at least include summary information in the computer database, in addition to references to supplementary materials stored elsewhere.

No matter how sophisticated the technical aids, effective documentation for secondary users requires some administrative policies and procedures. Researchers, on their own initiative, may maintain documentation about data structure for their own use, given efficient tools for doing so, but documentation of the origin of their data sets tends to be incomplete. All relevant information, including field notes, data abstracts, published articles, study plans, maps, and reference specimens, should be made available to secondary users. An ideal time for a data manager to obtain this information is when data are entered into a computer.

Documentation efficiency and uniformity can be fostered by developing forms for researchers to use to record the information. There should be both manual and computerized versions of these forms, so that information can easily be transcribed from paper to computer media, or so that researchers can use the computer directly as a note keeping device.

The field station's data management group should review all documentation of data supplied by researchers for incorporation into a data bank (or that are otherwise made available by a field station to secondary users), to ensure that minimal standards have been met.

Care should be exercised in developing forms and procedures, so that the recording of documentation does not become a burdensome extra task for the researcher. It is all too easy for a data management staff to become a bottleneck to efficient use of computing facilities.

DATA CATALOGS AND DIRECTORIES

Each field station that wishes its database to be a general resource for research and education should maintain some sort of directory or catalog of data. A data catalog or directory contains enough information about each data set 1) to enable a searcher to accurately locate a manageable subset of data sets of potential usefulness, 2) to direct the searcher to further information about the data sets, and 3) to direct the searcher to the data sets themselves. (A directory is simply a list, perhaps indexed, of data sets, while a catalog usually contains more complete information.)

In one sense, the information needed to enable researchers to locate and select useful data includes every bit of documentation down to the finest detail. The usefulness of a data set to a researcher may hinge on a fine detail of methodology, sampling schedule, or spatial distribution. However, the effort required to maintain all that information in a directory may be prohibitive. What is necessary is not that every detail be included in a catalog, but that the catalog direct the researcher to the relevant detailed information, whether it be in the hands of researchers, in a centralized data bank, or wherever. A data catalog can fit in quite nicely with a good documentation system. The information in a catalog is, in part, a subset of the documentation that ought to be kept for each data set.

The following is a list of the questions that a catalog should be able to answer, either by containing the information, or by telling the researcher where he or she can obtain it.

1. What do the data describe? (e.g., what organisms and parameters were studied?)
2. What was the purpose of the data? What hypotheses or questions were addressed?
3. What locations or habitats do the data pertain to? What is the spatial distribution?
4. When were the data gathered? What is the temporal distribution?
5. What persons were associated with collecting and analyzing the data?
6. What methods were used to obtain the data? (Experimental design, field and laboratory procedures, data processing algorithms, verification procedures)
7. How have the data been used? What publications pertain to the data? Do salient computer programs or printed versions of the data exist?
8. Where are the data, and in what form? How can they be accessed?
9. At what stage of activity is the data set? Is data collection ongoing or complete?

Data catalogs that contain this information can take on a variety of forms. They can be intended for browsing directly by interested researchers or via reference persons such as data managers or librarians. They can be kept on paper, or automated to varying degrees. Searching can be done via card indexes or through search commands issued at a computer terminal.

Although "paper" catalogs can be very useful, computerized catalogs have much greater potential. With an appropriate system, information can be entered and updated more easily, can be made more accessible, and can be searched more quickly and easily. It can also be more readily and clearly referenced with related information, so that, for example, it is easy to find the data corresponding to a publication, or vice versa. (However, as with all databases, computerization per se will not necessarily accomplish these objectives; a good manual system can be better than an inadequate computer system.)

Whether or not a catalog system is automated, it is best that its contents be organized into categories similar to those described in the previous section on data documentation. The following set of categories represents the minimum information that should be maintained for each data set:

1. DATA SET CODE, NAME, or TITLE-A unique identification for each data set.
2. DATA COLLECTION TIME PERIOD-A description of the data collection period and periodicity, and major temporal gaps or anomalies in the data set pattern.
3. PARAMETERS or VARIABLES—A complete list of the significant ecological variables contained in a data set.
4. INVESTIGATORS—Name of the person(s) responsible for the research or other project that generated the data.
5. CONTACT PERSON—Name of the person who is the primary contact regarding authorization to use the data, and access to the data.
6. BIBLIOGRAPHIC REFERENCES DESCRIBING THE DATA SET
7. DATA SET STORAGE LOCATION

8. RESEARCH LOCATION—Information that identifies the site of the research.

While this minimum information can alert researchers to relevant data sets, a catalog is much more useful if it is also indexed by (at least) taxonomic group, location, and general subject. These indexes might be in the form of card file indexes, or in a computerized catalog they might take the form of additional categories like the following:

1. KEYWORDS—Indexing terms that describe the subject matter of the data set.

2. TAXA—Indexing terms that describe the taxonomic groups that the data set pertains to.

3. RESEARCH SITE CHARACTERISTICS-Indexing terms that describe the habitat type or other ecological characteristics of the research site.

With some software, not only these, but any fields, are potential indices.

While data documentation can very well be the responsibility of individual researchers, a catalog must be centrally administered. As with much of data management, the development of a catalog is as much an administrative as a technical task, especially regarding its "input" aspects. There must be methods to get complete, up to date information from researchers. Giving researchers good documentation tools will make it especially easy for them to assist in the compilation of a catalog. If the catalog is a subset of a database of documentation that resides on a computer, it is conceivable that it can be compiled more or less automatically from the documentation.

. In order for data sets to be indexed consistently, a controlled list of indexing terms may be established. Developing these controlled vocabularies can be quite a task in itself. While a very large list of words may be needed to index a large bibliographic database containing hundreds of thousands of entries, it may not be necessary to index data sets with the same detail. If a field station has five hundred data sets, relatively coarse indexes might enable searchers to locate satisfactorily small subsets of entries.

To make future intersite access to data more efficient, catalogs should be compatible among sites. One sort of compatibility could be achieved through common indexing vocabularies. At present, sites are encouraged to exchange their indexes with each other, **to** promote an evolution of high quality, common indexing vocabularies. It would also be good for the information categories to be as similar as possible at all sites. The compilation of national, regional, or topical catalogs will be much easier for both compilers and contributors if the information is already maintained in a compatible format at the individual field stations.

Printed catalogs can serve a useful public relations function, but can also be misused. It helps to think of a printed catalog as only one view, or "subset" of a dynamic database. It may be sufficient to print relatively little information about each data set, perhaps only enough to call attention to the data catalog itself and to some of the grosser features of each data set. If the database is dynamic, a printed version will always be out of date, and should be treated accordingly. The same software that does ad hoc searching of a catalog can conceivably be capable of producing customized printed catalogs (for example, listing aquatic data sets for those persons researching aquatic habitats).

DATA BANKS

In order to preserve its total research database and make it more generally available, a site may choose to establish a data bank as a centralized repository for data. A data bank can be thought of as a database of databases. It provides researchers with a single source for all data pertaining to a site, and can ensure a degree of quality and consistency in the management of data and documentation. A data bank can ensure against loss of valuable data due to mismanagement, and provide a continuity of care for data, spanning researchers' careers and lifetimes.

Most of the work needed to develop and maintain a data bank pertains to the ways in which data are put into it. Although developing storage structures and search tools (the "output" system) for use by secondary users is an important task, it is even more important to develop methods for obtaining cooperation and data from contributing researchers (the "input" system).

Although it is desirable to have a central repository and access point for data, a station should have as a goal the decentralization of as many data bank functions as possible. Inadequate resources of hardware and software are likely to necessitate more centralization than is ideally necessary, but a station should work toward certain types of decentralization. For example, it is desirable for a data bank manager to ensure that certain standards for documentation are adhered to. One simple way for this to happen is for him or her personally to enter documentation into a database, or to supervise such activity, thus controlling what goes into the database. However, if the data management system is such that it can serve re-

searchers as a convenient note keeping device (a "super-notebook") and if subsets of their documentation can simultaneously be their own super-notebooks as well as part of the data bank, it is then possible for the researchers to maintain much of the documentation themselves.

If a data bank is a repository into which researchers put copies of their data after they have done their analyses, some potential problems must be dealt with. First of all, the process may mean an extra (redundant) step for the researcher if the data happen to be in a different form from that required for the data bank, or if they are in a different place. To avoid creating a barrier to cooperation by the researcher, a means of minimizing the extra effort, or of avoiding it altogether, is needed.

Secondly, if two copies of data are maintained, one in the data bank and one in the hands of the researcher, a means must be employed to ensure that any updates or additions to data or documentation are applied both to the researcher's copy and the data bank copy. It is better that there are not separate copies of active data sets, but rather that a single copy of data and documentation serve both the data bank and the researcher, especially in the case of active, long term data sets.

In the absence of more sophisticated, automated techniques for dealing with the problem of updating data and documentation, it is recommended that a regular system of review be set up. Each data set and its documentation should be scheduled for periodic review by the contributing researcher, who can be requested to note any updates or corrections that should be applied to the data or documentation. The period between reviews can be short when the data set is relatively active, and relatively long (on the order of years) thereafter.

Another issue that must be dealt with is quality control. The term means different things to different people. The types of quality control range from the scientific to the technical. They include the quality of research (e.g., quality of hypotheses and experimental design), quality of measurement (e.g., adequacy of instrumentation and methods, replication, confidence limits), and quality of recording and transcription of data (e.g., from field forms to computer).

The first type, quality of research, is of concern insofar as decisions must be made about what data are to be included in the data bank. For example, at many field stations operated by universities, there exist data resulting from student projects. These data may be useful for some purposes, but may not be of the same quality as those resulting from more rigorous studies by experienced researchers. Some selection criteria may be needed. The selection requires scientific judgment, and decisions by data management technicians should at least be subject to review (directly or indirectly) by the administrators of a field station. A simple way to handle the issue is to accept any data which an established researcher feels ought to be included.

In a sense, quality of research and of measurement can be "controlled" through rigorous documentation of data. If all data are thoroughly documented as to persons responsible, methods, etc., a secondary user can decide for himself whether a particular data set is of sufficient quality for his purpose.

The final type of quality control, regarding data recording and transcription, is particularly troublesome. Data entry procedures are prone to error. Much time is wasted when errors are found in data at advanced stages of analysis, requiring correction and reanalysis. Even worse from a scientific standpoint are the situations where errors are never detected. (Techniques for detecting errors are discussed under "Data Sets" earlier in this chapter.) Whatever data verificaton techniques are used, the documentation for the data should make clear to the user what procedures have (or have not) been used.

Whether or not it is done to ensure quality, there must be some control over what data are put in a data bank. Limitations on time and other resources require a station to at least set priorities on what data are to be included. A station may elect to include only data from certain habitats, or only data from "natural" habitats (as opposed to laboratory studies). A clear policy is necessary in order to maintain smooth relations with contributors, as well as to explain to secondary users the coverage of the data bank.

INTEGRATING DATABASES

In addition to managing databases such as species lists, data catalogs, and the individual research data sets within a data bank, a field station should consider how to manage them all as an integrated whole. These databases can be of much greater utility if they are linked together on the basis of related information, so that all data pertaining to a particular topic can be brought together for further analysis.

A data **catalog** itself provides an important degree of integration. While there may be disparate systems of data storage and coding among the different data sets, a data catalog describes them all according to a common set of indexes and information categories.

A special need at biological field stations is to link data on the basis of research locations and taxonomy. These two types of data deserve additional attention.

Research Locations

Almost all biological field data need to be identified as to the exact site to which they pertain. Data sets often contain a "site" variable, and even if all data in a set are from a single site, that location still needs to be identified in the data set's documentation. A field station may also maintain a database of land use information or land use plans that uses a coding system to identify sites.

It is desirable to tie all these data together, to make it possible to bring together all data pertaining to a particular site. However, inconsistent systems of coding or identification of sites are an obstacle. Research groups each tend to develop their own systems. A single scheme for labeling sites tends to be difficult to establish because different types of research require different sorts of spatial resolution, and because researchers tend to cling to time honored names for sites. One group may refer to its study area as Jones Field, another might refer to the same area as Plot 17C, while yet another might prefer to refer to it in terms of township, range, and section.

In spite of these inconsistencies, a great deal of compatibility can be achieved without requiring a rigid conformity by all researchers. A field station can achieve a good measure of integration by developing a master list (or database) of all its research locations. Some of the locations in a master list might be specific points (perhaps sampling stations in a stream], some might be small areas (study plots), and some might be large areas (an entire county or more). Some sites might be located within other sites, or might overlap. Locations at different levels of spatial resolution can be readily accommodated.

The master list can include complete, detailed information about each research site. Some possibilities are:

1. LOCATION NAME OR CODE-A standard name or code that uniquely identifies the site. It should be suitable for use as a code for the values of site variables within research data sets. All data sets should either use these codes directly, or else define a one-to-one correspondence between their codes and these.
2. SYNONYMS—Other names by which the site is known.
3. COORDINATES OR GRID LOCATION—The exact location of the site in terms of a common coordinate or grid system or equivalent. This information can serve as an index, and systematically identifies all locations.
4. GENERAL DESCRIPTION—A verbal description of the location and nature of the site.
5. TRAVEL DIRECTIONS—Instructions on how to travel to the site.
6. REFERENCE TO MAPS OR AERIAL PHOTOGRAPHS—References to maps or photographs on which the site is delineated.
7. ECOLOGICAL CHARACTERISTICS—A description of ecological characteristics, perhaps in terms of plant community types. This information can serve as an index to the master list.
8. CROSS REFERENCES—References to other locations in the list that encompass this site, or are included within it.

The use of such a master list does not preclude the use of disparate systems of identifying sites within the different data sets. Researchers can continue to use their own site naming systems. What is necessary is that all data sets and databases containing location data should either use the codes in the master list, or define their own codes in terms of the master list.

In addition to research data sets, some of the databases that should use the master list are data catalogs, species lists, land use databases, and publication lists. (See Figure 1.) The master list can serve as a de .facto index to all data at a field station, as well as serving as a common basis for merging or linking comparable data.

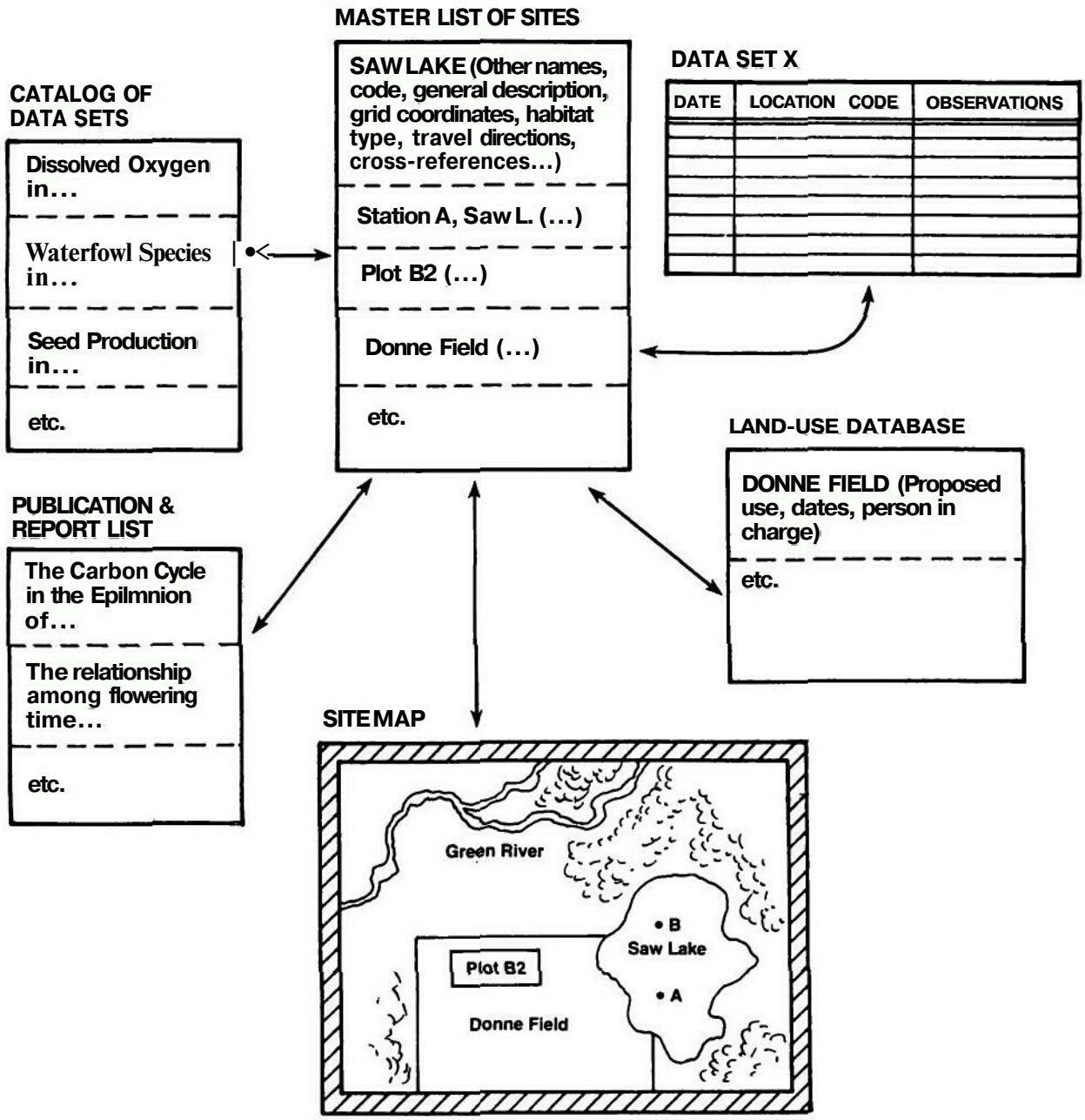
Taxonomic Data

Taxonomic information can also be treated as a separate master list. However, a complete taxonomic database for a field station can be a very ambitious project, since it should contain not just a list of names, but also show the taxonomic relationships. Ideally, it should reflect not only the current taxonomic nomenclature, but also should describe the sequence of changes that led to the current state, and should be periodically updated to reflect further changes.

Note that developing a taxonomic database is not just a matter of producing an all encompassing coding system; Linnaeus developed one in the eighteenth century which works quite well. Some form of more compact coding may be useful for computer efficiency, but is a relatively trivial part of the task.

There are several ways in which a taxonomic database can be put to use with other data sets. Sometimes researchers want to summarize, arrange, or aggregate data according to different taxonomic levels. It should be possible to merge the necessary information from a taxonomic database with that in their data sets so they can do so. Another use is as a basis or source

Figure 1. Role of a master list of locations at a field station. Sites ranging from large areas down to single points are accommodated, and cross-referenced to describe spatial relationships. General information about each research site is contained in the master list, rather than in the other databases. To ensure consistency, location codes used in data files are drawn from the master list. The general locations to which data sets or publications pertain are described by reference to the master list. The master list in turn serves as an index to the data catalog and to the publication list.



for a taxonomic thesaurus used to index data catalogs or publication lists. Even data entry software can use information from the taxonomic database to ensure that legitimate names are being entered into data sets. The use of a single taxonomic database for all of these purposes can avoid redundant data and effort. The use of a single coding system can make it easier to combine data sets for further analysis.

In contrast to a master list of locations which serves a single field station, a machine readable taxonomic database is of potential use to the entire ecological research community. The Association of Systematics Collections (ASC) is currently engaged in compiling such databases, and it is recommended that biological field stations look there for leadership and counsel. (Vertebrate species of the United States and mammal species of the world are presently available in hard-copy or magnetic tape, in whole or by selected subsets. Both data sets were compiled and verified with the assistance of specialists and will be updated periodically to reflect taxonomic changes.)

CHAPTER 3
COMPUTER SOFTWARE SYSTEMS

This chapter discusses several kinds of software for managing databases. These tools help us to manage both data and supporting documentation, and permit us to integrate data management with data analysis. There are many components to a complete software system, but this chapter begins by discussing two that are of particular importance to data management: data entry systems and data dictionaries. A third section discusses the more comprehensive type of software known as a database management system. The final section deals with ways of integrating the various tools into an easily used whole.

It is not possible or appropriate for every field station, given its resources and priorities, to implement all of the capabilities discussed in this chapter, at least not in the short term. However, these capabilities are becoming more commonly available, and short term planning should be done in the light of the longer term potentials.

DATA ENTRY

A data entry system should be given a high priority at each site. The data entry step is a crucial point at which standard procedures and protocols can be exercised to ensure that databases will be error free, consistent, and well documented. A good data entry system can make researchers more efficient at a troublesome task, and at some sites may even be the dominant element of the data management system.

In Chapter 2 we discussed the advantages of entering data as soon as they are collected, by the persons most familiar with the data. This sort of timely and personal data entry is only practical in an environment where personal computers or terminals are easily accessible, and only with a system that is easy to learn and to use.

A simple approach is to use an interactive text editor for data entry. The main advantage is that text editors are often used by researchers for other purposes, so no additional learning is required. However, specialized data entry systems that can be tailored to a data set offer many features that text editors lack. They can control and guide the entry process by, for example, displaying forms on a video screen with blanks to be filled in, and by doing some initial error checking.

Data entry systems can also capture documentation about data. Typically, in order to get started, a person must first define the data to be entered. Names must be given to variables, and ranges or lists of valid values must be specified. This is the basic information needed by the software to check for valid values, but additional information about each variable and file (such as is listed in Tables 2 and 3) can also be collected. The most convenient time for the researcher to record such documentation is probably at this step.

The information must be stored somewhere, and the logical place to put it is in a "data dictionary" type of database. Data dictionaries are discussed in more detail in another section, but for now it will suffice to think of them as central repositories for data about data. Each researcher might have a data dictionary, or there could be one central data dictionary for an entire field station, or some combination of the two.

The link between the data entry software and the data dictionary is very important. The data definition task can be made easier, and at the same time some consistency can be enforced, if a common pool of variable definitions is available to the researcher. For example, if a data file needs to contain a variable that identifies treatments (via a treatment code), and this is a variable that has already been defined for another file, it would be good if the researcher did not need to redefine it. Instead, he could specify that he wants to use a prestored definition. This capability is especially important for variables that identify sites, species, or dates, because these are often the basis for linking data sets together, and for indexing data sets. A central database of definitions of these variables can help make data sets compatible with each other.

Although **interactive** data entry is efficient, a complete data entry system should also handle data that are entered in batches (e.g., on keypunch cards), or from real-time data acquisition systems, data loggers, and instruments. The data entry systems should have a component that serves as a filter (Figure 2) to ensure that all data are defined in a consistent fashion, and that error checking is done on all data, whatever their source. To serve in such a flexible fashion, it is necessary that the software modules that do data definition, interactive data entry, and error checking be independent, so they can be incorporated in all types of software that do data entry.

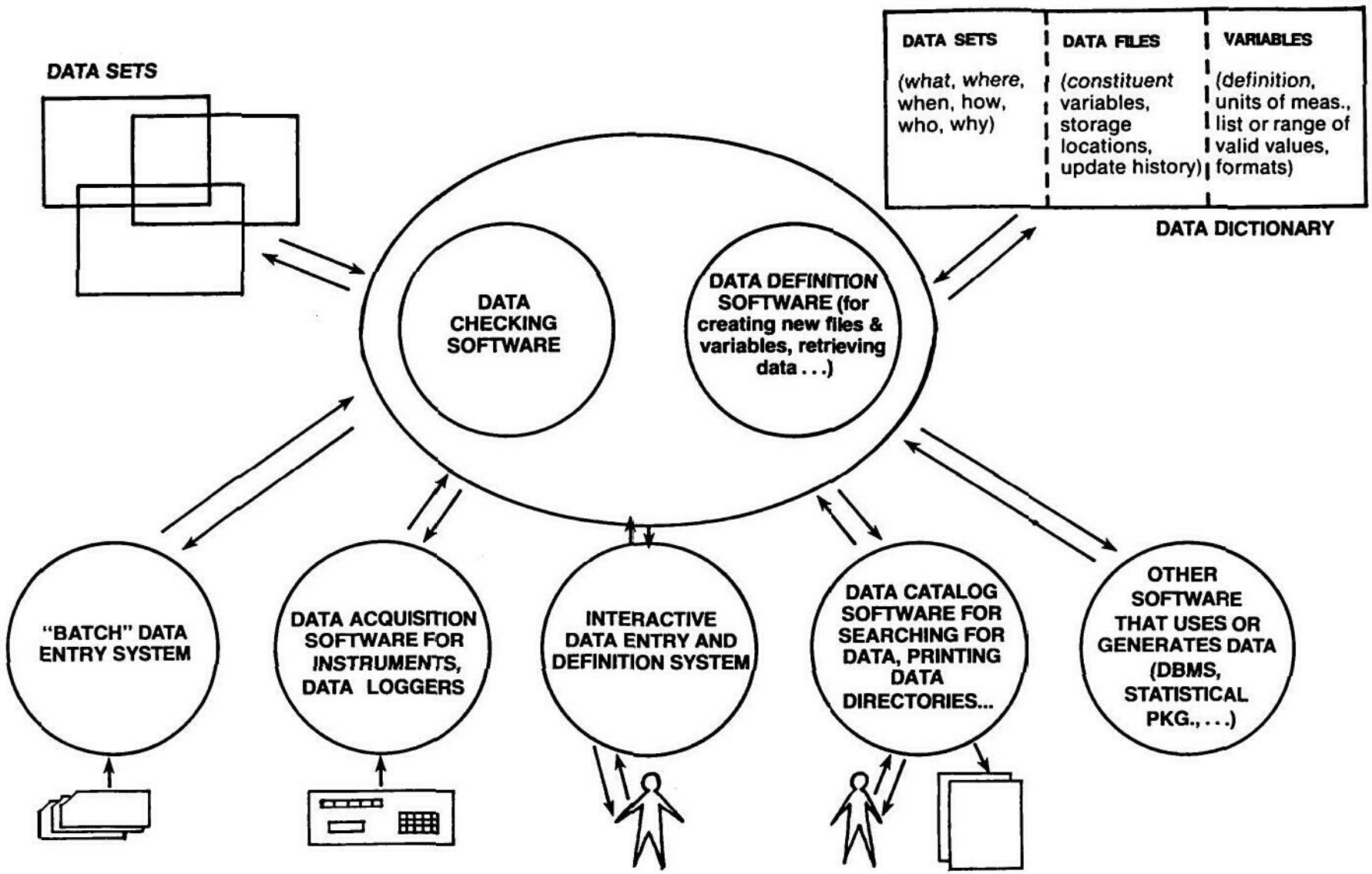


Figure 2. Role of a data entry system and data dictionary in Research Data Management.
A core of data entry/data dictionary software acts as a filter to ensure consistency and documentation of all data, and serves as a common means of access to information needed in order for humans and software to use data.

The ideal data entry system should be able to compare sets of data and report differences, so that the double entry type of error checking (discussed in Chapter 2) can be done. It also needs a good link to a "report writer" so that printed copies of data can be produced for proofreading and safekeeping. (When data are entered directly, rather than being transcribed from field forms, these will take the place of the field forms.) A final helpful feature is a means of maintaining a revision history (as discussed in Chapter 2).

A data entry system can be a good beginning, especially at many of the university field stations that have visiting researchers, but do not have resident research programs. Visiting researchers typically collect data during the field season, and analyze them elsewhere during the remainder of the year. The technique of offering these researchers the use of a computer system in exchange for copies of their data and documentation (and their cooperation) will not be effective if they perceive the computer's only value to be as a data analysis tool. They do not care to spend time on data analysis when time spent in the field is at a premium. An appropriate service to offer those researchers is in the area of data entry. If they can use a computer as a convenient data entry (and documentation) device, they can enter their data as soon as they are collected, and later transfer them elsewhere for analysis. The benefit to the researchers is that they can enter data in a more timely and reliable fashion. (If in addition, they have some basic data processing capabilities with which they can easily produce simple data summaries, they can use this information to make timely adjustments to their data collection procedures.) The field station, in turn, benefits because it has a better opportunity to capture data and documentation at the source.

DATA DICTIONARIES

A data dictionary is a specialized database that contains data about data (sometimes called "metadata"). It contains definitions of variables and files, as discussed in the preceding section. However, it could also be much more general, containing a data directory or catalog, or even complete documentation of all data, computerized or not.

The main purposes for computerizing documentation (including that in data directories and catalogs) are:

1. To impose an organization on the documentation and enforce consistency and completeness.
2. To keep data and documentation together so that, given a data set, its documentation can be easily located, and vice versa. (This does not necessarily imply a physical togetherness, e.g., on the same magnetic tape.)
3. To enable researchers to locate existing data more easily, on the basis of indexed documentation.
4. To cross-reference related documentation in ways that are appropriate to the nature of the documentation, but impractical without computers.
5. To be used as information for controlling the maintenance of databases (especially at the data entry step) and software (although the latter is beyond the scope of this report).

In order to computerize documentation, software with features not found in most general purpose data management software is usually needed. Although the software products known as data dictionaries, information storage and retrieval systems, and card file systems (for microcomputers) all have some of the necessary features, data dictionaries, as described by Ross (1981), are conceptually the most comprehensive. Some desirable features include:

User definition of entities: The software should allow the data manager or other users to specify the types of entities that they want to document, and to specify a list of categories, or fields, of information to be kept for each entity type. For example, it should permit a data manager to set up databases of documentation for data sets, maps, methods, variables, or any other entity types, and should allow him or her to specify the categories of information, such as "investigators" and "time period," to be included. This contrasts with the inflexibility of the built-in data dictionaries sold with many database management systems (DBMSs) which typically manage only certain information about those files and variables managed by that DBMS. They do not permit one to maintain information about other entities that a site may wish to document, such as data sets, maps, and publicatons. Also, because they are inseparable from a particular DBMS, they are of no help for data that are not managed by that DBMS. This is a major disadvantage, since field stations will likely also wish to document data that are not on computers (which may often be the bulk of the data).

Cross-referencing: The software should allow the user to establish cross-references (two way references) between classes of entities. For example, if data sets and publications are cross-referenced, it means that the documentation on publications includes a list of all related data sets, and vice versa. This is one feature usually missing from information storage and retrieval systems (such as those com-

monly used for bibliographic databases), which otherwise might serve some data dictionary functions.

Cross-referencing is sometimes confused with indexing. An index is much like the index to a book. It enables one to find all the data sets or publications on a topic. By contrast, cross-referencing is the means by which one data set can refer to another particular data set or publication, and vice versa.

The software should support automatic cross-referencing. This means that if someone enters a reference in data set A to publication B, the corresponding reference will automatically be placed in publication B. Without such a capability, cross-referencing is tedious and error prone, and could just as well be done manually.

Indexing: The software should enable the user to set up indexes. This is often done by allowing the database creator to define a particular field (or category) as being an indexed field. For example, if a data catalog has "keyword" and "taxa" fields that are indexed, it means that researchers interested in data about insect pollinators of goldenrods can specify search terms such as pollinators, insects, and Sohdago, and receive a list of all data sets that have been indexed accordingly. Many software systems that support indexing also enable users to search a database on the basis of information in any field, not just indexed fields, although such searches are less efficient for the computer.

An additional de *facto* indexing technique can be provided by cross-referencing. Suppose that a field station documents both data sets and research sites. If the data sets are cross-referenced with research sites, then the list of research sites serves as an additional index to data sets, i.e., given a particular research site, all related data sets can be located. Also, if-the list of research sites is itself indexed, say according to habitat type, the habitat index also serves as a de *facto* index to data sets.

Thesauruses: If indexing is to be consistent, a list of valid indexing terms (also called a "controlled vocabulary") must be available for indexers to use. These lists can be maintained in the form of thesauruses as described by Lancaster (1979, Chapter 12). They can contain simple lists of terms, or can link together related terms, such as narrower and broader terms or synonyms. It is best if the software can maintain multiple thesauruses for each database, and if the thesauruses are independent of particular databases (or entity types). For example, it should be possible to maintain at least two thesauruses for a database of data sets, one containing general subject terms, and another containing taxonomic names. It should be possible to use these same two thesauruses elsewhere, say to index a database of publications.

Textual data types: A data dictionary must be able to handle textual data. Many general purpose database management systems can handle character data types, but very few handle textual data types (where each datum is an arbitrarily long chunk of text). However some systems that allow long character strings may allow a procrustean solution. The lack of this feature makes many general purpose database management systems unsuitable for documentation.

DATABASE MANAGEMENT SYSTEMS

Although database management systems (DBMSs) are the most general and basic of the software we consider, different persons will have different ideas of what they are and what they are used for. This is in part because the meaning of "database management system" often depends on whether it is used in the context of large "mainframe" computers, minicomputers, or microcomputers. Persons who work with business databases on large computers would not consider the DBMSs available for microcomputers to be worthy of the name, while to a person operating in a microcomputer environment, the DBMSs used on large computers might seem unnecessarily complex and more of a hindrance than a help to accomplishing useful work.

Rather than concentrating on DBMS features commonly found in any one of these computing environments, this discussion will cover certain features that are especially appropriate to biological field stations. We need to be aware that DBMS priorities for research tend to differ from those for business. Researchers often do ad hoc analyses, while much business data processing is (or at least used to be) devoted to regularly scheduled, repeated processing of databases with a relatively static structure. Research data processing involves a multitude of data sets, whose structure may often need to be modified (added to) and upon which a multitude of analyses are performed. The databases themselves and the analyses that are performed are ad hoc. Researchers are constantly collecting new types of data and looking at their data in new ways. Data management at a field station is typically done by the primary users of the data (or by someone who works very closely with them), while in the world of business a separate department is typically responsible for data management. However, now that business users are doing more personal ad hoc computing, researchers are likely to benefit from the products developed to meet business needs.

A DBMS, if comprehensive enough, can tie all other software and data together by serving as a general purpose storage and retrieval system for all types of data. A common data structure can make possible a consistent treatment for all data. Tools for error checking, documentation, and security are easy to develop and to use if the data are in a common form. A DBMS can also include a language for retrieving and manipulating data to prepare it for use by data analysis or data reporting software. These two features, a generic structure for data, and a set of generic operations to manipulate data, can free the researcher from many of the details involved in performing the same functions in general purpose programming languages such as FORTRAN or Pascal.

The DBMS can be a stand-alone system for entering, manipulating, retrieving, and analyzing data, but it can also be a component, or building block, of other software. For example a special purpose program for meteorological data could be made to use a DBMS internally for storing and retrieving data. It is also conceivable that data dictionaries, statistical analysis software, and even word processing systems could use a DBMS internally to maintain their data.

Data management can, of course, be done without the software that goes under the name "database management system." Sometimes other software products, alone or in combination, provide some of the functions that we might otherwise obtain from a DBMS. We consider here some important features.

Generic data **structure:** A uniform data storage structure can do much to integrate data management. It is far too confusing and wasteful to have to store data in one way for one analysis and in another way for others. A good DBMS will make it possible to store all data in a uniform way, yet retrieve them easily in the form required by any other software.

To be completely universal, so as to serve as a foundation for all types of software, a DBMS should support numeric, character, and textual types of data. (At present, very few DBMSs can handle textual data types, but several software vendors are working toward it.) A more specialized data type that is very useful is a "date" data type. Built-in means for handling missing values are also important.

DBMS data structures are generally categorized according to three models—hierarchical, network, or relational. Relational DBMSs are based on a normalized structure (as described in Chapter 2). Both network and relational DBMSs can handle data of any complexity. The network structure is not quite as simple, and network DBMSs require one to define in advance the relationships between entities. They dominate in business environments where very large databases are maintained. Relational DBMSs that are currently available are mostly too slow and inefficient (in their use of computer processors) to be used on large databases that are in constant use. However, relational DBMSs are often well suited to ad hoc development of data bases and ad hoc data processing, so are often well suited to the research environment. Since most ecological data contain hierarchies, it might appear that hierarchical DBMS should be used. They sometimes are appropriate, but it should be noted that although most data sets contain hierarchies, hierarchies are often not sufficient to represent an entire data set.

Data manipulation language: A DBMS should provide the user with a language to do three basic types of data manipulation: subsetting, merging, and aggregating. High level languages that perform these operations are a tool that can greatly increase the productivity of researchers doing data analyses, and can free them from dependency on programmers.

Data independence: A DBMS can make the data storage structures independent from the programs that use the data. This makes it possible to change a database without disrupting programs that use it. A common type of change is that which results when a researcher decides in mid experiment to collect a new type of data. For example, he may have been collecting data about individual plants of a population when he decides to start collecting data about insect damage and insect populations in his study plots. The insect information must be tied to the information about plots, and so must be integrated into the database. Without a DBMS, the structure of a database will probably be referenced explicitly in programs written in a language such as FORTRAN or Basic, i.e., in its READ or WRITE statements. If the programs deal with complex sequences of records, changes will be difficult, especially if several programs need to be changed. A good DBMS, however, will make many types of changes possible without necessitating changes in the programs that read or write the data.

Redundancy control: Redundancy can be eliminated or controlled with a DBMS. Redundancy often occurs during data analysis when the data need to be merged and aggregated in a certain way for one analysis and merged with a subset of other data and aggregated differently for another analysis. If each results in a different copy of the data, and if the original data from which these copies were derived is changed due to an update or error correction, then the several derived copies must be changed also. If a data management system allows, or tempts, researchers to make error corrections on derived data rather than on the raw data, great confusion can result. There are two common ways that a DBMS can

help. One is to store with each file a copy of the commands that were used to create the file. The commands can then be executed again at a later time, if necessary. The other is to make it possible to "view" the same data in different ways. The definition of the data processing steps is stored, but not the resulting data. It appears to the researcher that there is another copy of the data, derived from the raw data, but no actual copy exists. Instead, each time the researcher uses the stored view, the data records are created anew from the up to date raw data.

Data integrity control: Sophisticated DBMSs should assist in controlling the integrity of a database by allowing one to specify constraints on the values of variables and on relationships between variables and entities. For example, it can enable one to specify for a "temperature" variable that its values must lie between -20" and 35°C. Or, for a "species" variable, it should enable one to specify that only values that also exist in column X of table Y (which might be a species list) can be put in the database. This capability is especially valuable at the data entry step.

Security control: A DBMS can control access to data by allowing the manager of a database to make specified portions of it available to specified persons, for specified purposes (e.g., updating, reading), and **at** specific times and places.

Auxiliary functions: There are some auxiliary components that are often packaged with a DBMS. They can include a data entry system, report writer, and statistical and graphical functions. The DBMS that has such functions should also permit easy interfacing with other such software components that are not part of the same package.

Multiple interfaces: Ideally, it should be possible to execute DBMS commands both via a special data manipulation language and through higher level languages such as FORTRAN. If the latter is possible, the DBMS can then serve as a building block for further customization.

INTEGRATING SOFTWARE SYSTEMS

Data management software should be integrated with other software into a coherent whole. The ideal data management system will be comprehensive, have a high degree of data compatibility, and operate consistently in all parts. Thus far, this chapter has discussed several types of software: data entry systems, data dictionaries, and database management systems. Statistical and graphics packages, report writers, and word processing system have also been mentioned. There are several ways that all of these software modules need to work together.

Researchers often need to use different types of software to analyze a set of data. In an integrated system, the analysis should proceed smoothly without problems caused by converting between different, incompatible data formats. Output from one type of software should be usable as input to another, as when one wants to use graphics to portray the results of a statistical analysis. Even if a field station has the necessary tools to do all sorts of data processing, they might not be used effectively if they do not operate consistently. Keystrokes and commands should be as similar as possible in all components of the system. For example, it is confusing to have the command "quit" mean in one place that you are finished, and in another that you want to undo what you just did. It is confusing (and even dangerous) for the command "purge" to mean, in one place, "remove old, outdated versions of a file," and in another, "delete the one and only copy of a file." It is hard for the casual user to learn different keystrokes to do the same editing function in different places. Editing, especially, should be consistent, because it is done in many places. Documents get edited in word processing, data get edited, commands get edited, and documentation gets edited.

Commands for entering and editing documentation should be similar to those used for data. (The distinction between data and documentation is often fuzzy, anyway. One person's documentation is another's data.) Consistent operation is more likely attained if there is a comprehensive, consistent data structure underlying the system.

This sort of integration is not easy to achieve, but is worth working toward. There are several approaches to the task. Some provide only a partial degree of integration, but can be done with products that are currently available. Others provide more complete integration, but require more work.

Buying a Single Software System

One approach is to use a single software system for all purposes. Given the dominant role of statistical analysis in research computing, this most likely means that the system will be a statistical package that has some data management capabilities. For this purpose, a statistical package will need, in addition to its statistical capabilities, a generic set of data manipulation operations that allow the user to do the complex combinations of subsetting, aggregating, and merging that may be needed to prepare data for statistical analysis. For some research these capabilities are more useful than the statistical tests per se, and the lack of them is often the major bottleneck for researchers doing analysis of complex data sets.

Most statistical packages provide some sort of storage structure for data and maintain some

rudimentary forms of documentation (such as labels for files, variables, and data values), and some provide for storage of user defined procedures for manipulating and analyzing the data. The use of these packages thus makes data more self-documenting. Some also have useful graphics and report writing capabilities.

The foremost disadvantage of this method of integration is that it is not likely to be a complete solution. For example, a fully integrated system should be able to handle not only data, but also data about data, including that in textual form. At present the same software systems that handle numeric and character data well do not handle textual data well, and vice versa. And no single package is likely to be able to do everything that a researcher might want to do with his or her data. The primary advantage is simplicity. There is only one system for researchers to learn, and only one system for a support staff to maintain. Documentation is also made simpler because it can all be done in terms of a single system.

Developing a Comprehensive Systems from Scratch

At the other extreme is the strategy of developing a complete system in-house. While dissatisfaction with existing products may tempt some persons to try this approach, it is not recommended. It would of course be possible to make a system as comprehensive and as consistent as one wants, but it would not be likely to find its way off the drawing board. Designing such a system, much less implementing it, would tax the resources of even the largest biological field station. Even if the resources were available, it would not be cost effective unless it were developed for sale. It would certainly include much "reinventing of wheels."

In any event, "custom programming" is often so dependent on specific personnel that when they leave, the software is no longer useful. It should be kept to a minimum.

Exchanging Data Between Software Modules

No matter how comprehensive a particular software package may be, researchers will sometimes need other software. An obstacle is that each software system tends to have its own data input format and internal data format. To deal with this need, links can be developed between pairs of software packages, so that any one package can read and write data in the other's internal format. Some statistical packages already have such capabilities. In some cases where those links do not already exist, a field station could develop its own. This is a reasonable task if the software modules to be linked have interfaces to general purpose programming languages for reading and writing data, so that the programmer does not have to become involved with internal storage formats. There are some disadvantages to this approach. Difficulties will arise where not all types of data used in one system are supported by the other. Consistency and ease of use will not likely be obtained, since each module will probably have a different command syntax. And developing a link between each pair of packages can result in a great number of links, and therefore a cumbersome system.

Exchanging Data via a Common Data Structure

Rather than converting data between each pair of formats, it may be better to adopt a single data format to be used to store all data, and to develop utilities to convert data between the common format and those required by each of the software modules. Not only can this reduce the number of conversion utilities needed, but it also makes both data analysis and data documentation simpler. The documentation system can be based on the common format. This format could be one that a station develops in-house, or one that it adopts as part of a DBMS. (A good candidate for a common data structure is one that is normalized.)

There is a disadvantage to the use of a data format that is independent of a site's most commonly used software. The step of converting from the common format to that needed for data analysis is potentially a clumsy extra step that is wasteful of computer resources, and makes it difficult to take advantage of the machine efficiencies afforded by a software system's internal format. However, the concept of a common data format is a necessary component of all schemes to completely integrate data and software systems.

Developing a Single System from Software Modules

The techniques discussed so far can provide some integration, but it is not comprehensive. It would be good if statistics, graphics, word processing, record keeping, and modeling software could be mixed and matched into a unified system in which data could be freely passed from one function to another, and which operated in a consistent, uniform fashion.

What is needed are flexible modules that we can buy and easily incorporate into a system that has an underlying data structure and user interface of our choice. The main obstacle is that the available software usually has its program control, input and out-

put functions all intertwined with its main function. Input and output should be designed so that output from one module can be used as input to another. Specialized, printable outputs are fine, but each software module should also be able to produce output in a raw form readable by other programs. It is good for software to have a user interface in the form of a command language or menu system, but it should also be "callable" from general purpose programming languages.

It is possible that, in the future, software developers will make their software more modular so that it can be interfaced easily with other software. An analogy is in the computer hardware industry. At one time manufacturers did not design their equipment so that others' peripheral devices could be easily attached, but now many of them do. If the same type of developments take place in the software industry, we will be able to, with reasonable effort, develop software systems that are truly comprehensive and coherent.

CHAPTER 4
DATA ADMINISTRATION

RELATIONSHIP OF DATA MANAGER TO SITE

The role of research data management (RDM) is to facilitate and integrate research at the site and thus serve to sharpen the focus of the research program. The effectiveness of a research data management program depends upon the support of the site administration as well as individual researchers. To function most effectively, a research data management group should be established which has its own identity and a sufficient base level of institutional support to insure a sustained program. Establishment of the RDM group requires full and continued financial support from the administration. However, as the RDM evolves, financial support may diversify due to increased levels of external support. At some sites it may not be necessary to have a full-time data manager, provided that its goals and level of activity are modest, but a successful program is not likely to be a natural outgrowth of other activity with computers.

The qualifications of the research data manager should stress primary training and expertise in ecology or other appropriate scientific disciplines combined with knowledge of information management, data processing, and statistics. An RDM group with these qualities lends credence to reports and publications, and increased credibility to the administration's overall planning and organization. More narrowly specialized data managers may lack the perspective needed to assist researchers with data analysis, review data documentation, and integrate data management with other research activities.

The major responsibilities of the RDM unit include:

1. Advising researchers on the development of research plans, including format of data forms, experimental design, sampling design, etc.
2. Developing a research data management computer system (including documentation, data input, management of data files, etc.) appropriate to the level of activity and resources of the site.
3. Providing quality assurance of data through appropriate procedures such as checking for missing elements, valid codes, and outliers.
4. Performing analyses of needs in relation to data accessibility, hardware and/or software.
5. Continuing evaluation of the research data management system (RDMS) with modification as necessary.
6. Participating in related professional activities, including workshops, conducting training or orientation sessions for users of the RDMS, preparing reports or papers on data management or other research interests.
7. Increasing the awareness of researchers and administration to RDM'S ongoing activities and capabilities through close interpersonal communication, development of newsletters, data catalogs, annual reports, public presentations, and other means.

ROLE OF SITE ADMINISTRATORS

The site administrator is responsible for defining the RDM program to be developed at that particular site for current and future demands and for determining the particular mix of duties of the RDM personnel. Priorities will obviously differ between sites and projects within sites, but a clear understanding of what these priorities will be is important to insure an effective data management system.

It is important that, having defined the RDM program for the site, the administrator vigorously support and promote it by all possible means. This should include a commitment to maintain an RDM group as a continuing component of the site program regardless of variability in outside funding, and to enhance the visibility of the program to encourage active cooperation of investigators using the site. It is essential that the administration foster integration of the RDM unit into the total research organization by including it in planning activities and budgetary considerations, and by continuing to enhance the concept of RDM as a vital component in the institution's organizational scheme. The means of accomplishing these goals will obviously vary with the site and even within the site depending upon the relationship of the various levels of research to the RDM program.

PRIORITIES

The role of administration is central to effective RDM insofar as policy and its implementation defines the framework for developing a RDM program and

setting activity priorities for the field station. The goals must be clearly defined by the administration. Once these goals are established, based on historical, current, and anticipated needs of the station, activities can be prioritized.

Recommended steps to be followed in determining priorities are as follows: (1) inventory, (2) define task. (3) determine priorities of needs, (4) determine availability of resources, (5) reassess, (6) select methods.

1. Inventory—The administration must first conduct an inventory of the data base(s) and RDM resources. These resources include past, present, and future research programs; types, amounts, and forms of the data; and staff, money and facilities.
2. Define task—After the inventory, decisions should be made regarding objectives for each data set. These decisions should consider the condition of the data set and needs for future implementation in terms of site goals, research programs, schedules, and/or user needs.
3. Determine priorities of needs—Tasks should be ranked using a synthesis of field station goals and the data. A diversity of priorities exists among field stations. These site-specific priorities reflect the different goals and resources of the facilities. For example, RDM at some sites focuses on current research activities whereas other sites emphasize existing databases. Most sites manage data from both ongoing and past research.
4. Determine availability of resources—Once the data management tasks have been identified and ranked according to priority, available resources (number and training of the data management personnel, availability of software and hardware, estimated staff time for project completion, project duration, project lead times, and projected budgets) should be examined to determine the extent to which they are adequate for accomplishing the tasks. For certain tasks, in-house capabilities may not exist. It is also quite likely that the desired set of data management tasks demands more than the available data management resources. Thus, further decisions must be made.
5. Reassessment—Based upon the overall goals of the station and the analysis of resources, the data management tasks should be reprioritized in terms of feasibility. If certain important tasks cannot be accomplished in-house, then financial resources must be allocated to have them completed externally. Other less important tasks may be deferred for an indefinite time period.
6. Selection of methods—The next step is to determine detailed methods for completing the desired data management tasks. One of the most basic decisions is the determination of whether the task should be manual or computerized. Irrespective of the method, data must be organized and documented so that the data are available for secondary users and amenable to future computerization.

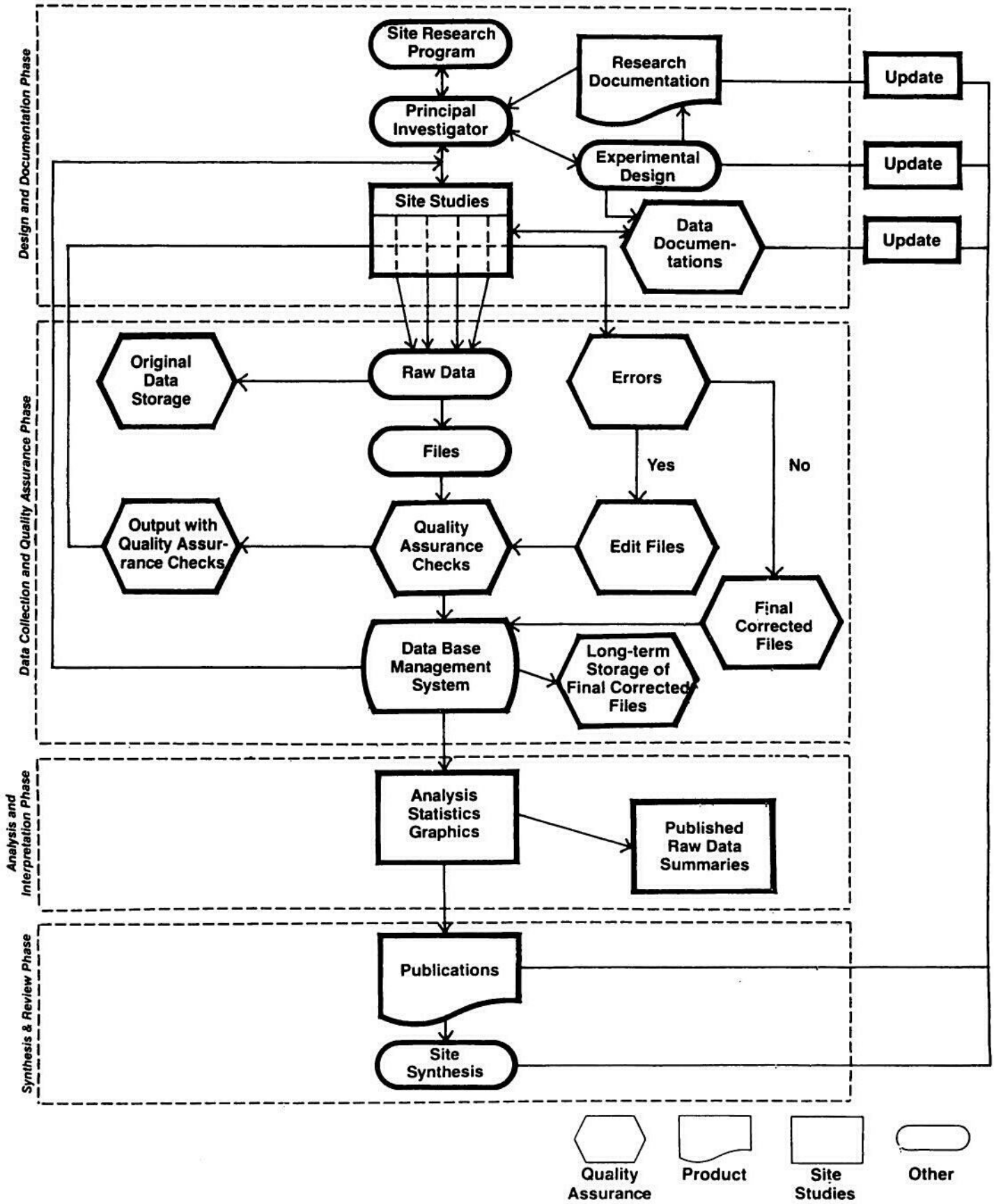
COMPUTER SYSTEM SELECTION

If the decision is made to computerize the database, a series of system selection criteria should be formulated outlining software requirements and subsequent hardware configurations. The selection or development of appropriate software is of primary importance for accomplishing RDM tasks. To augment this selection process, it must be noted that computer software is universally constrained by available computer systems and that in-house development of application programs for data handling and analysis is usually not cost effective. When examining available systems to meet anticipated research needs, the major system selection criteria from an administrative viewpoint are:

1. Vendor support of the system's software, including help in troubleshooting user applications.
2. Research data management capabilities that are easily programmed (user oriented), flexible, possess simple instructions for sorting, merging, and updating, and accept user programmed instructions for input, output, and quality control procedures.
3. A basic complement of statistical analysis routines, graphic and cartographic capabilities, report generation routines, and more advanced statistical analysis capabilities.
4. Ease of interfacing with other software packages and/or application programs.
5. A common syntax for batch and interactive operation.
6. Cost effectiveness not only in terms of computer costs but also in the personnel time needed for implementation and maintenance.

From the administrative viewpoint, all research data management activities must be planned. What is not clear perhaps is the amount and direction of planning necessary after a software package has been selected. The amount of planning for integrating research databases appears to be inversely proportional to the degree to which the selected software package meets the system selection criteria. If the criteria are adhered to closely, then planning the integration of the RDMS

Figure 3. Research Data Management Sequence



can be minimal. On the other hand, if the system selection criteria are not followed closely, planning time may be increased and the emphasis shifted more to the mechanics of documentation, data entry, and file manipulations. Therefore, careful selection of the software system permits the research data manager to be more involved with research end products, such as exploratory graphical displays, publication quality graphical output, computer generated tables, and quality assurance controls. In turn, the scientist benefits by becoming more involved with interpreting results of the study than with initial data management tasks. Such an approach to RDM places an emphasis on the needs of the scientists. Additionally, efficiency is gained in the field operations, where the majority of the cost is usually involved, without additional cost to the data management program.

DATA INVENTORIES

A data inventory is the process (and result) of compiling an exhaustive list of data of potential usefulness to the data management objectives. Before a field station can develop a data management system, it should have a good idea of what data it has to manage. Thus, a data inventory should be a first step, and should be the basis for decisions regarding the development of data management systems and databases. However, compiling this list of data is not a one-time project. It should be an ongoing list, reflecting a station's current awareness of extant data, and therefore part of an iterative sequence of evaluation and development of a data management system.

A data inventory is useful not only for planning purposes, but also to provide continuity in data management. In addition to containing a list of data sets, the inventory should also include a record of decisions (and rationale) regarding field station support and responsibility for these data sets.

The inventory process consists of two parts. One is to inventory historical data sets, and the other is to maintain an awareness of data sets as they are created. These two parts can be treated somewhat differently.

The first may require a bit of detective work. A list may be compiled by examining existing data management schemes, perusing publications, and soliciting information from researchers about data collections from their own past research or that of their colleagues.

The process of maintaining an awareness of current data sets can be more systematic. Some stations simply require that all researchers using the site's facilities leave copies of their data at the station, although the politics of the station aren't always amenable to that approach. Some stations use computer resources as a "carrot," requiring all persons who use those resources to cooperate with data managers in making their data and documentation available. Other stations do not insist on such cooperation, but rely on the usefulness of computer resources to bring researchers into contact with data managers, thus making their research and data known. Tools for data analysis are especially attractive to resident researchers, but some researchers use a field station for field work during the summer or during short term visits, and do their data analysis elsewhere. Good quality data entry systems can foster communication with such researchers, if the capability exists for a smooth transfer of their data to other systems. If a data entry system or data analysis system is powerful and easy to use, it can even attract researchers who would ordinarily think of their data sets as too small to bother putting on a computer, and might even be useful to those whose data are of an anecdotal nature.

DOCUMENTATION PROCEDURES

A great challenge to data administration is the comprehensiveness and quality of data documentation. Data managers must give high priority to developing a system of incentives to encourage researchers to document their data thoroughly.

Among the most effective incentives for ensuring the cooperation of researchers is the provision of a system that will produce tangible improvements in the efficiency and effectiveness with which their data can be analyzed. Other incentives can be given by providing help in designing efficient field sheets, thorough quality assurance procedures, and efficient interfacing to powerful and flexible graphical and statistical analytical tools. Reduced file storage costs and an automated data retrieval/security system are additional incentives for sites in which these services are not normally available to researchers.

Policing (enforcement) policies can, in combination with voluntary incentives, provide a high degree of documentation and researcher participation in an RDMS. At several sites, documentation standards are mandatory at the time of data entry if the data are to be input to the RDMS. At other sites, funding sources are tied to the researchers' fulfillment of data documentation requirements. A combination of incentives and policing often makes for the most effective administrative system.

Once a successful system of data documentation procedures has been established, the potential value of archived data to the biological field station is dramatically enhanced, making the cataloging and organization of the data a logical and essential follow-up step to reach the ultimate goal of increased data accessibility and use.

One aspect of documentation that is often overlooked is that of RDMS documentation—all of the policies and procedures governing the operation of the RDMS. An RDM newsletter can often provide a useful way to begin this process. A user's manual or operations manual including details of data entry procedures, archiving and cataloging, and general policies is not available from most sites—yet could be a useful tool for increasing continuity of procedures in the case of personnel turnover and for evaluating RDMS effectiveness.

SECURITY

RDMS vary in their attention to data security. Noncomputerized data files may be stored in filing cabinets or other appropriate facilities; computerized data may be stored in card image files or in various database management systems. In all cases, several copies of the final data should be archived for long term storage at several different locations. For computerized data these copies should include both magnetic and hardcopy forms.

During analysis, synthesis and publication, updates of research and data documentation may be necessary, On rare occasions, even experimental design may have to be updated and additional data collected. A very important step is the publication of final raw data summaries; hard data copy deposited in a number of libraries is the only truly permanent data record, and for the forseeable future, the most accessible.

A data manager who is attempting to encourage researchers to use a research data management system must be prepared to offer assurance of security from unauthorized use or manipulation. This assurance can take several forms. For example, when using computer systems, the file of interest can be protected by requiring passwords for access. Another form of security is to have the researcher maintain all copies of the raw data prior to publication. In this case, only the documentation is made available to other researchers with a potential interest in the data.

BUDGETS

Budgets for research data management systems are difficult to separate from other objectives at biological field stations. The wide spectrum of RDMS capabilities presently existing at biological field stations further complicates comparisons of RDM budgets. Some systems feature full implementations of each of the major types of computer capabilities. For other institutions, data management primarily consists of archiving and organizing manual files of data and associated documentation. The budgets for RDM usually reflect these different levels of system capabilities and uses.

Total operating budgets for biological field stations vary from less than $100,000 to over $20 million per annum. The proportion of operational budgets devoted to RDM varies from 2 percent for sites at the initial stages of organizing a RDMS to almost 10 percent. Most sites are supporting RDMS with 5-6 percent of the field station's operational budget. Although the suggested proportions of RDM budgets can be used as a rough guideline to the overall level of financial commitment to RDM, the size and diversity of data being managed can significantly influence the amount of resources that will be needed. Sites that manage a few large data sets often require a smaller percentage of station operating funds than sites that deal with many smaller data sets. Similarly, the initial cost of the conversion of data and operational procedures from a dispersed manual RDMS to a centralized and computerized RDMS will be more than the maintenance of a centralized system for ongoing projects. If all research data are to be fully organized and documented for secondary analyses, more financial commitment and administrative skills are required. If a site is not committed to the treatment of data as a long term resource, then less immediate financial commitment is necessary. However, short term financial savings will often be overshadowed by long term scientific loss.

CHAPTER 5
EXCHANGE OF INFORMATION BETWEEN SITES

DATA EXCHANGE NETWORK

Each field station or other agency that manages ecological data should view itself, not as an isolated entity, but as a node in a data management network. Many of the components of this network already exist, but if data management plans are made using a network perspective, many types of data exchange can be made more efficient. (Some aspects of this network are depicted in Figure 4.)

There are many obstacles to data exchange. Oftentimes useful data exist, but there is no convenient means for researchers to find out about them, at least not in sufficient detail. The network can include information centers that make this sort of information available.

Another obstacle is caused by incompatibilities between data sets. A data network should foster common exchange formats for data and documentation. It would also be possible for some of the institutions in the network to develop and distribute (for example) taxonomic thesauruses that can be used at field stations to standardize the handling of taxonomic data. The necessary funding and cooperation for such efforts is more likely to be obtained in a network context.

A third obstacle is the lack of documentation. Data often have insufficient documentation to be of use. Efforts can be made to develop standardized, complete systems of documentation throughout the network.

The data network consists of two types of institution. The first is the typical biological field station where research is conducted. The second type deals with tasks beyond the scope and capacity of a single field station. The latter can be called "secondary agencies," since they focus on secondary use of data.

There are several possible roles for secondary agencies. One is to serve as information centers to help researchers locate and obtain data kept elsewhere. They can maintain data catalogs similar to those kept at field stations, except that since they are centralized, they are more easily accessible. These agencies will not be able to work as closely with contributing researchers as data managers at field stations do, so they will need to rely heavily on data cataloging efforts taking place within field stations. One example (represented at the workshop) is the National Environmental Data Referral Service operated by the National Oceanic and Atmospheric Administration.

Another role is as compilers and custodians of large databases, which can be thought of as national data resources. These databases are often developed where an agency has been charged with studying a large scale environmental problem, such as acid rain. They represent data gathering efforts that exceed the capability of a single field station, but they are compiled from data that originate at field stations. The database maintained by the National Atmospheric Deposition Program (NADP), and the Geoecology database at Oak Ridge National Laboratory (ORNL) are examples that serve some rather urgent research needs.

Ecological and taxonomic thesauruses represent another type of database that should be maintained by secondary agencies. The development of taxonomic databases by the Association of Systematics Collections is an example. One of their uses is in developing standardized indexing and coding systems for data and documentation.

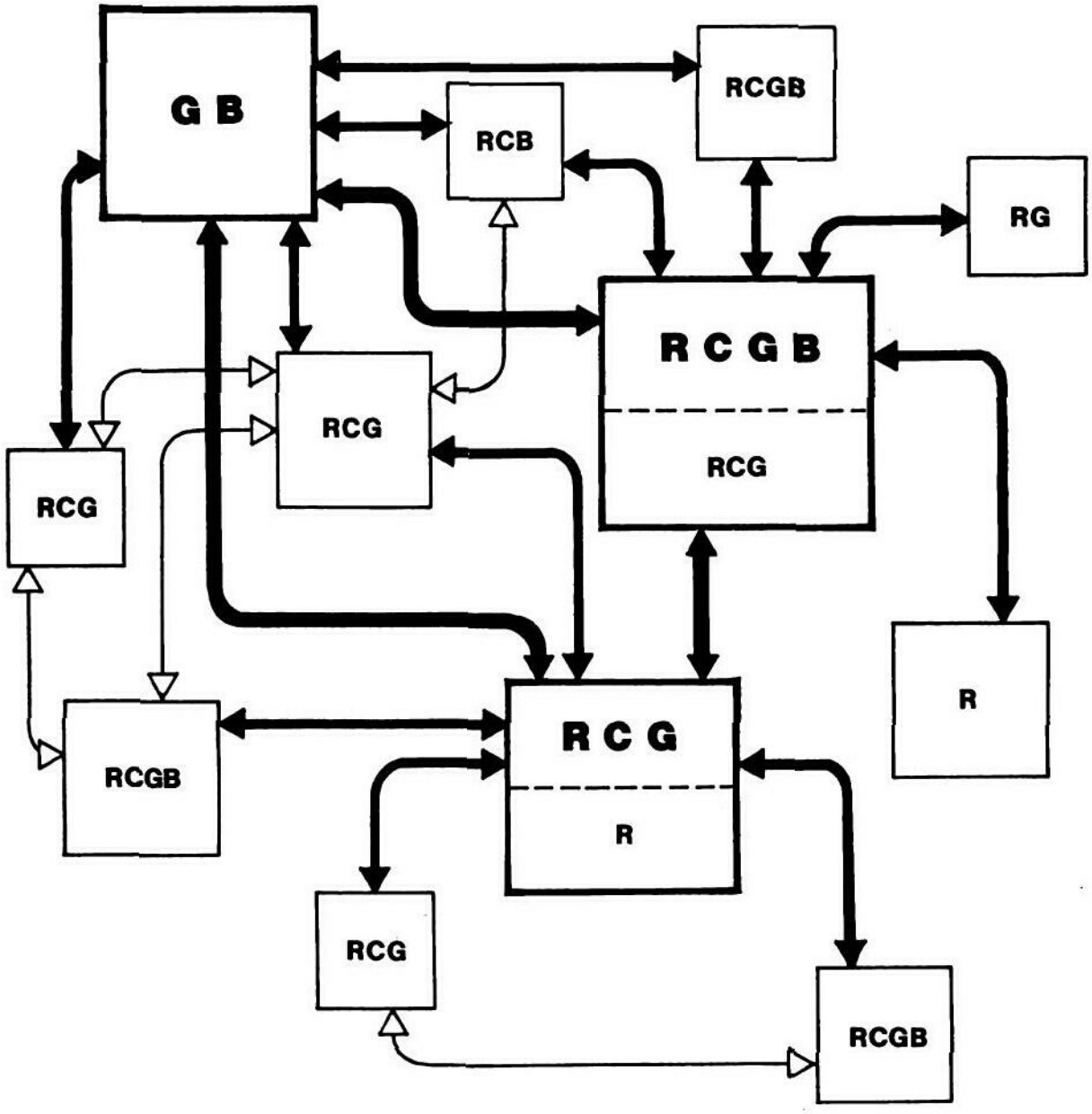
A few secondary institutions could serve as data banks, or repositories, for data that have no other means of long term care. There are important data sets, often the result of work by researchers now deceased, that cannot be cared for properly at field stations or on college campuses.

Although these functions need to be centralized, decentralization is desirable where possible. The network should serve as a "distributed database." That is, data should be accessible from anywhere in the network, but they should be stored and managed locally. This takes advantage of the interest and motivation of the originators of the data, and avoids error prone redundancy. (If a redundant copy of a data set is stored in a repository, it is in danger of becoming outdated due to changes or additions in the original.) Rather than store data in central repositories, it is better to just keep a central directory (although it too must be kept up to date).

Figure 4. A hypothetical data network, consisting of biological field stations plus a few secondary agencies. Each institution is represented by a box. The letters represent four types of data management activity, which can involve local data (small letters), or data across many sites (large letters). The different combinations of letters in each box represent the diversity of activities among institutions. Secondary agencies deal with data across many sites, and serve to tie all the field stations together, reducing the number of links necessary. The arrows represent data exchange paths, with the heavier lines representing especially efficient, heavily traveled paths. All are two-way paths, allowing not only exchange of data and information about data, but feedback on their use.

Legend:

R = Research data analysis
C = Information centers and data catalogs
G = Compilation of databases for general use
B = Databanks



Although information centers will expedite information transfer, it should not be inferred that all data transfer must go through secondary agencies. Researchers at field stations will continue to maintain direct ties with other field stations, and can obtain data directly, without having to go through intermediaries. However direct transfer between sites will also benefit from work done to make transfers via secondary information centers more efficient.

Although more than one data management role may be performed at a given secondary agency, the roles should not be combined or confused. An agency that puts together a national database (for example) might be well situated to maintain a national data directory. However, it should not be assumed that because it has large computers or great expertise in one area, that it will be able to perform all other data management roles. Each task needs separate and sufficient funding, administration, and expertise.

To be mutually beneficial, all data transfer pathways should involve feedback mechanisms. All secondary use of data should be acknowledged, and researchers should be informed of the utility and use of their data for secondary purposes. This is especially important for research on environmental problems of a large geographic scale. It is often far too expensive for these programs to generate all the necessary data themselves; they must rely on data generated locally. However, even though researchers at field stations do not view themselves as data generators for large projects, they might be persuaded to make alterations or additions to their research programs to produce data that are also of use to others, especially if it could increase their own visibility in the eyes of funding agencies.

As a final point, it should be noted that this network approach, while it can meet some pressing needs, is a low risk approach. It takes advantage of existing resources and expertise. It does not involve grandiose plans that will not work until every piece is in place. It can develop gradually, with every stage being useful in its own right, because it meets primary as well as secondary data management needs.

PROTOCOL FOR EXCHANGE OF DATA

Relationships between primary and secondary researchers deserve careful attention in any data exchange. It is not uncommon for researchers to hesitate to make their data available to others. One reason is that researchers are (naturally enough) jealous of the time and expense that went into collecting the data. Another is that data can be misused in ways that might reflect badly on the contributing researcher. A set of data that is quite adequate for one purpose may be inappropriate for another. The contributing researcher will not want to expose himself to criticisms resulting from misuse of the data.

Whenever a researcher does learn of data at another site that he or she would like to obtain, the following steps should be taken as a matter of courtesy, and to protect reasonable proprietary rights.

1. Where the original investigator has so specified, permission for use of the data should be obtained. The original investigator should also be invited to provide relevant information concerning the collection of the data, and to collaborate in the new research in an appropriate role.
2. If the original investigator gives approval (or is deceased), the use of the data by the new investigator should proceed.
3. Any use of the data should be given prominent acknowledgment. The original investigator should be informed of its utility and use.

In some circumstances a data set may include information that should not be made available to the public in general. For example, it would seem inadvisable to reveal the locations of specimens of some threatened and/or endangered species.

MECHANISMS OF EXCHANGE

The actual exchange of data between sites involves four important considerations: 1) the medium on which the data will be transferred (paper, cards, magnetic tape, telephones lines, etc.), 2) the structure of the data to be transferred, 3) documentation describing the data and how it was prepared and formatted for the transfer, and 4) verification that the transfer was completed without error.

A very simple way of exchanging data is via a printed listing. For small volumes of data, it is a quick and efficient method. It can be easily documented and does not require verification. For readability, listings with tightly packed data fields and obscure codes should be avoided. For example, a date may be stored as the number 053082, but is more readable if printed as "30 May 1982." Sample identification codes that pack several items of information into a single code should be avoided. Headings and labels (with units) should be used, and the data should be arranged for maximum readability. Printed output should be labeled with the date of printing, the source of the data, and other identifying information (such as file names). If additional documentation is available (perhaps from a data catalog), it should also be provided.

If the data set is large, or if the secondary user plans to do computerized data analysis, then transfer media such as magnetic tape, cards, or floppy disks are more

appropriate. It is sometimes an easy matter for the sender and the recipient to find a mutually compatible transfer medium. For example, if both sites are using the same model of computer, the problem may be greatly simplified. For transfer between different kinds of systems, 9-track magnetic tape is the most common "standard" medium for large computers, and the 8-inch "CP/M format" floppy disk is one of the few standards for microcomputers.

The proliferation of microcomputers with nonstandard floppy disk formats will make media compatibility an increasingly difficult problem. Fortunately, microcomputer users often develop telephone links to transfer data to and from larger computers, such as those at campus computer centers. These links can then be used to access magnetic tape drives. Unfortunately, data communication over telephone lines can be very slow and error prone (depending on available equipment and software), expensive over long distances, and troublesome to set up for various combinations of computers.

The second important data exchange consideration is the structure of the data. For ease of use and documentation, it is best that the data be sent in a "normalized" form. This means that the data should be as organized as a set of files containing two-dimensional arrays (i.e. tables with rows and columns). Note that this is the required input form for most statistical packages. Records should not contain repeating groups, and there should be only one type of record in each file. One example of normalization is the separation of sample identification or description records from sample measurement records (when there are multiple measurement records for each sample), placing each record type in a separate file. In general, the simpler the file structure, the easier it is for the receiving site to process the data.

• The third consideration, documentation, is frequently given insufficient attention. There are two distinct kinds needed: 1) documentation of the data itself, which has already been emphasized in this report as being of critical importance for secondary use, and 2) documentation of the precise form in which the data exists on the transfer medium. The emphasis here will be placed on the second kind of documentation.

The information that is needed for a trouble free transfer depends on the medium used, and the complexity of the data set. For example, when a data set is transferred on paper, no technical documentation is needed, but if that data set consists of 50 assorted listings, obviously some explanation or index would be helpful. When the medium is magnetic tape or disk, technical details are essential. They may include: 1) the physical data recording format, 2) identification of any special software needed, 3) the number of files, 4) an index to the contents of each file, 5) how much storage space is required, and 6) what means of verification and error recovery is provided. These ideas are illustrated in the sample guidelines for preparing magnetic tapes that appear at the end of this section.

The final consideration is how to ensure an accurate transfer. Sending sites should always verify that magnetic transfer media (especially tapes) were written correctly and are readable, by using software to read them back and compare them to original copies. They should also provide some sort of redundant information that the receiving site can use for verification. A simple and reliable method is to send two complete copies of all data files. The receiving site then reads them both and uses comparison software to verify that they are identical.

Another approach is to send some sort of summary information along with the data. It could be as simple as the number of records in each file, or the range of values of each variable. A much more reliable method is to compute summary parameters such as the mean and variance of each variable, which the receiving site can then recompute for comparison. More technical methods (such as software generated checksums or cyclic redundancy codes) are not recommended unless both sites have appropriate compatible software.

Verification is especially important when data are transferred over telephone lines. Some sophisticated data communication protocols are designed to detect and correct transmission errors automatically, but the commonly used asynchronous dial-up link does not provide such luxuries. Reliability can be very poor, especially in rural areas. There is software available for many computers (including microcomputers) that will handle transmission errors, but it must be run on both the sending and the receiving computers. Lacking such software, reliable transmissions can be ensured using the methods outlined above for magnetic media. That is, multiple copies or summary information can be sent and compared.

The following is an example of a guideline that could be developed into a standard for writing magnetic tapes for data exchange. It illustrates some of the documentation and verification ideas discussed above. Magnetic tape is widely regarded as the exchange medium of choice because of its low cost, high capacity, common usage, and (most of all) its standard physical recording methods. Unfortunately, using tapes generated at other sites is often quite a struggle, unless procedures such as those suggested below are adhered to.

1. Tapes should be written on a "9-track" tape drive. (7-track drives are obsolete and becoming quite rare.)
2. They should be written at a density of 800 or 1600 BPI (bytes per inch), preferably 1600 for better reliability. (800 BPI is becoming obsolete, and 6250 BPI drives are less common than 1600.)
3. They should be written in "card image" format, using the ASCII character set; they should never be written in binary form, or in any "internal" form such as that used by statistical packages. (Other character sets such as EBCDIC and "half-ASCII" may be required at some sites.)
4. The tape should not be "labeled." That is, no special heading information should be recorded at the beginning of the tape (as is common when tapes are used only within a site). Such information is typically formatted differently between sites, and thus not usable.
5. All files written on one tape should use the same "block size," and all records (lines) within these files should be of some fixed length. (Variable length records must be truncated or padded with blanks to achieve a fixed length.) Block size must be at least as large as the record length, and if there is more than one record per block, no record should span across blocks.
6. It is a good idea to record two copies of all files (especially if there is extra room on the tape) in case a file cannot be read due to dirt or defects on the tape. Also, the redundant copies can be used to verify that the tape was read accurately.
7. The following information should be written on the tape reel (e.g., on one or more adhesive labels):

   a. character set that was used
   b. recording density in bytes per inch
   c. record length in characters
   d. number of records per physical tape block
   e. block length in characters (or bytes)
   f. some indication of the tape's contents
   g. name, address, and telephone number of the tape's owner
   h. name and telephone number of the tape's preparer
   i. a note of any documentation files contained on the tape

8. The documentation describing how the data are organized and formatted on the tape should be provided in printed form, and should also be recorded as the first file on the tape. Then, even if the printed information is misplaced, the tape is still fully documented. (The information on the reel itself is sufficient to allow reading of the documentation file, and it then provides the information needed to read the data files.) If any of the documentation on the data itself is available in machine readable form, it should also be included on the tape.
9. Sample printouts of the data on the tape should be provided as additional documentation, and to help the receiving site verify that they have successfully unloaded the tape.

Note that these guidelines are based on the assumption that the sending and receiving sites do not have computers with the same "operating system" software, which is the most common situation. When both sites do have the same operating system, there are typically better ways to format tapes and ensure reliability.

SHARING OF EXPERTISE ON INFORMATION MANAGEMENT

There is a need to share not only data, but also expertise on information management itself. For field stations to make their data management methods compatible with those of other sites in a network, there must be an awareness of what is being done elsewhere. Most field stations are at a very early stage in developing data management systems. It would be better for them to learn from the experience of others rather than to repeat each other's mistakes. The limited funds and personnel of most stations make it particularly important to avoid expensive mistakes.

There are several possible ways to share expertise. Some of them require special funding, while others can be done on the initiative of individual field stations.

One possibility is to have courses, consulting services, and internships that take advantage of the experience and expertise of leaders in scientific information management. Several cooperating institutions would need to be involved to ensure a sufficiently flexible approach to different needs.

A second type of exchange is a cooperative effort or pooling of resources, undertaken by a group of field stations. This would be most appropriate for small field stations within a single region, or where similar research is being conducted. Such an approach might use resources efficiently, and promote compatible systems and collaborative research syntheses. It might also help keep research personnel from getting bogged down in information management responsibilities.

Conferences or workshops are of great value. Expenses could be reduced if they could be held in conjunction with meetings of professional societies. They are of greater value if other, more frequent, exchange can take place between meetings. A national newsletter would be an ideal medium.

A simple way for stations to share their expertise is to communicate (i.e. advertise) their current data management activities to each other. For example, one field station currently produces an in-house newsletter which it also mails to other sites. Stations could share in-house announcements or other printed materials. (The appendix lists workshop participants who can be contacted for specific information about data management at their sites.) Such exchange, while simple, can easily lead to valuable personal exchange of information between data management personnel. It would also provide a higher visibility for the field station at a relatively low cost, and could be the precursor to a more formal newsletter.

BIBLIOGRAPHY

Altman, P. L. and K. D. Fisher. 1981. Guidelines for development of biology data banks. Federation of American Societies for Experimental Biology. 71 pp.

Lancaster, F. W. 1979. Information retrieval systems. John Wiley and Sons. 381 pp.

Lee, W. L., B. M. Bell. and J. F. Sutton. 1982. Guidelines for acquisition and management of biological specimens. Association of Systematics Collections. 42 pp.

Martin, J. 1977. Computer data-base organization. 2nd edition. Prentice Hall. 713 pp.

National Science Foundation. 1977. Long-term ecological measurements: Report of a conference. 26 pp.

National Science Foundation. 1978. A pilot program for long-term observation and study of ecosystems in the United States: Report of a second conference on long-term ecological measurements. 44 pp.

National Science Foundation. 1979. Long-term ecological research: Concept statement and measurement needs. 27pp.

Olson, R. J., C. J. Emerson, and M. K. Nungesser. 1980. Geoecology: A county-level environmental data base for the conterminous United States. ORNL/TM-7351. Oak Ridge National Laboratory. 312 pp.

Oppenheimer, C. H., D. Oppenheimer, and W. C. Brogden. 1976. Environmental data management. Plenum Press. 244 pp.

Ross, R. G. 1981. Data dictionaries and data administration. AMACOM. 454 pp.

Sarasen, L. 1981. Why museum computer projects fail. Museum News 59(4):40-49.

APPENDIX
Workshop Participants

Paul Alaback
Forest Research Laboratory
Oregon State University
Corvallis, OR 97331 (WL)

John Balling
Chesapeake Bay Center for Environmental Studies
Smithsonian Institution
P.O. Box 28
Edgewater, MD 21037

Carl Bowser
Department of Geology and Geophysics
University of Wisconsin
Madison, WI 53706

Craig C. Brandt
Science Applications, Inc.
Jakcson Plaza Tower, Suite 1000
800 Oak Ridge Turnpike
Oak Ridge, TN 37830

Warren Brigham
Illinois Natural History Survey
607 East Peabody Drive
Champaign, IL 61820

Richard Coles
The Washington University Tyson Research Center
P.O. Box 258
Eureka, MO 63025

Melvin I. Dyer
Environmental Sciences Division
Oak Ridge National Laboratory
Oak Ridge, TN 37830 (WL)

John S. Eaton
Section of Ecology and Systematics
Biological Science Building
Cornell University
Ithaca, NY 14855

Stephen R. Edwards
Association of Systematics Collections
Museum of Natural History
University of Kansas
Lawrence, KS 66045

Michael FarreU
Environmental Sciences Division
Oak Ridge National Laboratory
Oak Ridge. TN 37830 (WL)

Robert R. Freeman
Environmental Science Information Center
National Oceanic and Atmospheric Administration
11400 Rockville Pike
Rockville, MD 20852 (SR)
Current address:
National Environmental Data Referral Service Program
Office, 3300 Whitehaven St. N.W.,
Washington, D.C. 20235

Charles Gish
Office of Biological Services
U.S. Fish and Wildlife Service
Department of the Interior
Washington, D.C. 20240

John Gorentz,
W.K. Kellogg Biological Station
Michigan State University
Hickory Corners, MI 49060 **(WL,SR)**

Frank Harris
Division of Biotic Systems and Resources
National Science Foundation
Washington, D.C. 20550 (OB)

Robert Jenkins
The Nature Conservancy
1800 North Kent Street
Arlington, Virginia 22209

Claudia L. Jolls
University **of** Michigan **Biological Station**
Pellston, MI 49769 (OB)

Greg Koerper
Forest Research Laboratory
Oregon State University
Corvallis, OR 97331 (WL)

Vera Komarkova
Institute of Arctic and Alpine Research
Box 450
University of Colorado
Boulder, CO 80309

George H. Lauff
W.K. Kellogg Biological Station
Michigan State University
Hickory Corners, MI 49060

James Layne
Archbold Biological Station
Rt. 2, Box 180
Lake Placid, FL 33852 (SR)

Orie L. Loucks
The Institute of Ecology
Holcomb Research Institute
Butler University
Indianapolis, IN 46208

Ken Lubinski
Illinois Natural History Survey
Box 221
Grafton, IL 62037 (SR)

Marvin Marozas
P.O. Box 1630
Baruch Institute
University of South Carolina
Georgetown, SC 29440 **(WL,SR)**

G. Richard Marzolf
Division of Biology
Kansas State University
Manhattan, KS 66502 (WL)

William Michener
P.O. Box 1630
Baruch Institute
University of South Carolina
Georgetown, SC 29440

Paul Risser
Illinois Natural History Survey
607 Peabody Drive
Champaign, IL 61820

I. Robert Stottlemyer
Great Lakes Area Research Studies Unit
Department of Biological Sciences
Michigan Technological University
Houghton, MI 49931

John Tester
Department of Ecology and Behavioral Biology
University of Minnesota
Minneapolis, MN 55455 (OB)

Stephen Threlkeld
University of Oklahoma Biological Station
Star Route 1
Kingston, Oklahoma 73439

Robert Vande Kopple
University of Michigan Biological Station
Pellston, MI 49769

Vicki Watson
213 Bacteriology
University of Wisconsin
Madison, WI 53706

Steven H. Weiss
W.K. Kellogg Biological Station
Michigan State University
Hickory Corners, MI 49060 (WL)

Robert G. Wetzel
W.K. Kellogg Biological Station
Michigan State University
Hickory Corners, MI 49060

Michael Wooten
Savannah River Ecology Laboratory
Drawer E
Aiken, SC 29801

Kathleen Zinnell
Department of Ecology & Behavioral Biology
University of Minnesota
Minneapolis, MN 55455

OB - Observer
SR - A person who can provide a site report or other printed material on request
WL - Working group co-leader