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

INTRODUCTION 66

SUMMARY OF RECOMMENDATIONS 68

CHAPTER 1 VIEWS OF DATA MANAGEMENT 72

The Perspective of Biological Field Stations 72
A Research Perspective 72
The Perspective of Secondary Users 73

CHAPTER 2 DATABASES 74

Data Sets 74
Biological Inventories 77
Documentation Systems 79
Data Catalogs and Directories 82
Data Banks 83
Integrating Databases 84

CHAPTER 3 COMPUTER SOFTWARE SYSTEMS 88

Data Entry 88
Data Dictionaries 90
Data Management Systems 91
Integrating Software Systems 93

CHAPTER 4 DATA ADMINISTRATION 96

Relation of Data Manager to Site 96
Role of Site Administrators 96
Priorities 96
Computer System Selection 97
Data Inventories 99
Documentation Procedures 99
Security 100
Budgets 100

CHAPTER 5 EXCHANGE OF INFORMATION BETWEEN SITES 102

Data Exchange Network 102
Protocol for Exchange of Data 104
Mechanisms of Exchange 104
Sharing of Expertise on Information Management 106

BIBLIOGRAPHY 108

APPENDIX LIST OF PARTICIPANTS 110

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 developing 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: 1) 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 necessary 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 Title	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.

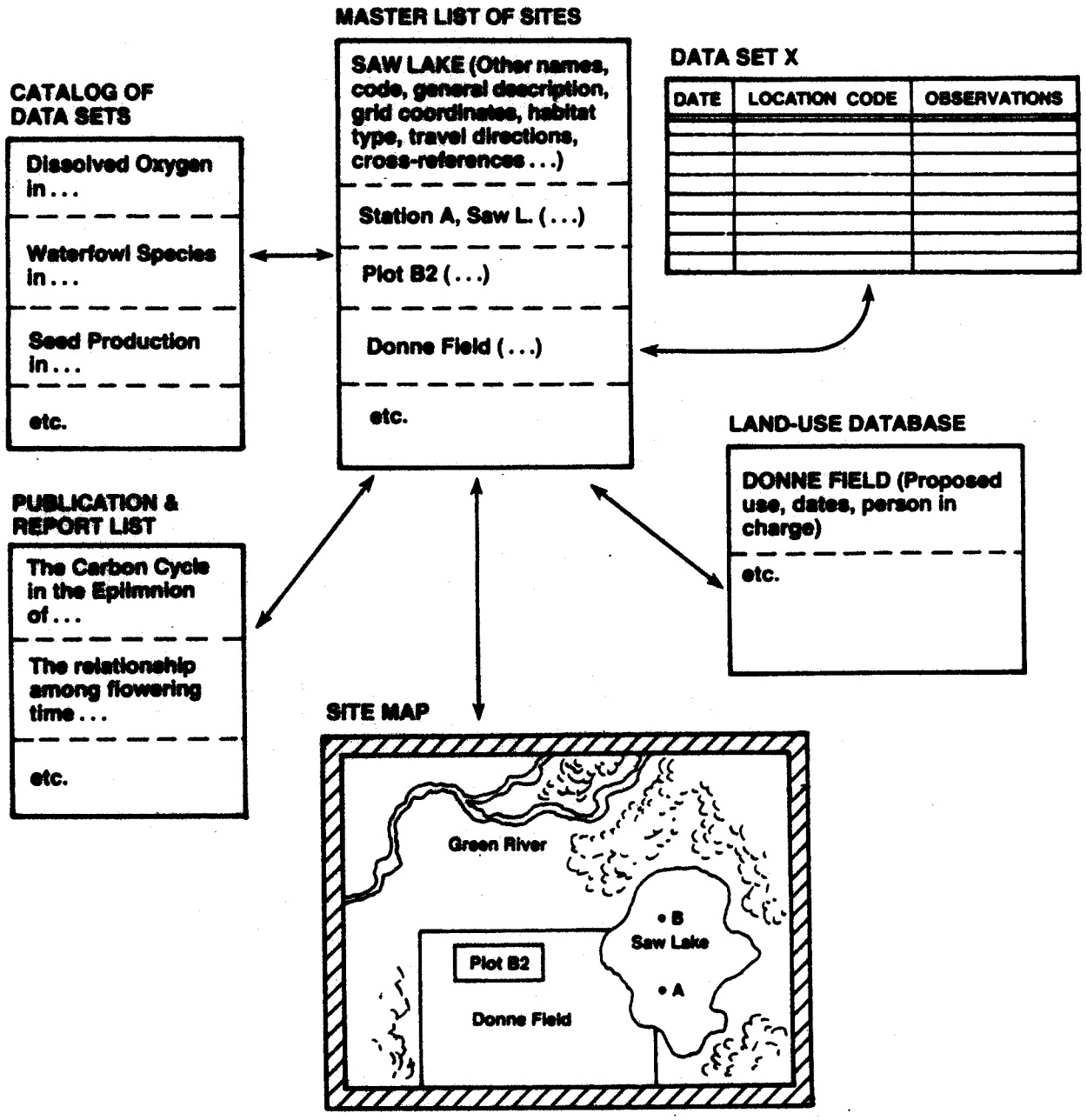
Taxonomic Data

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.

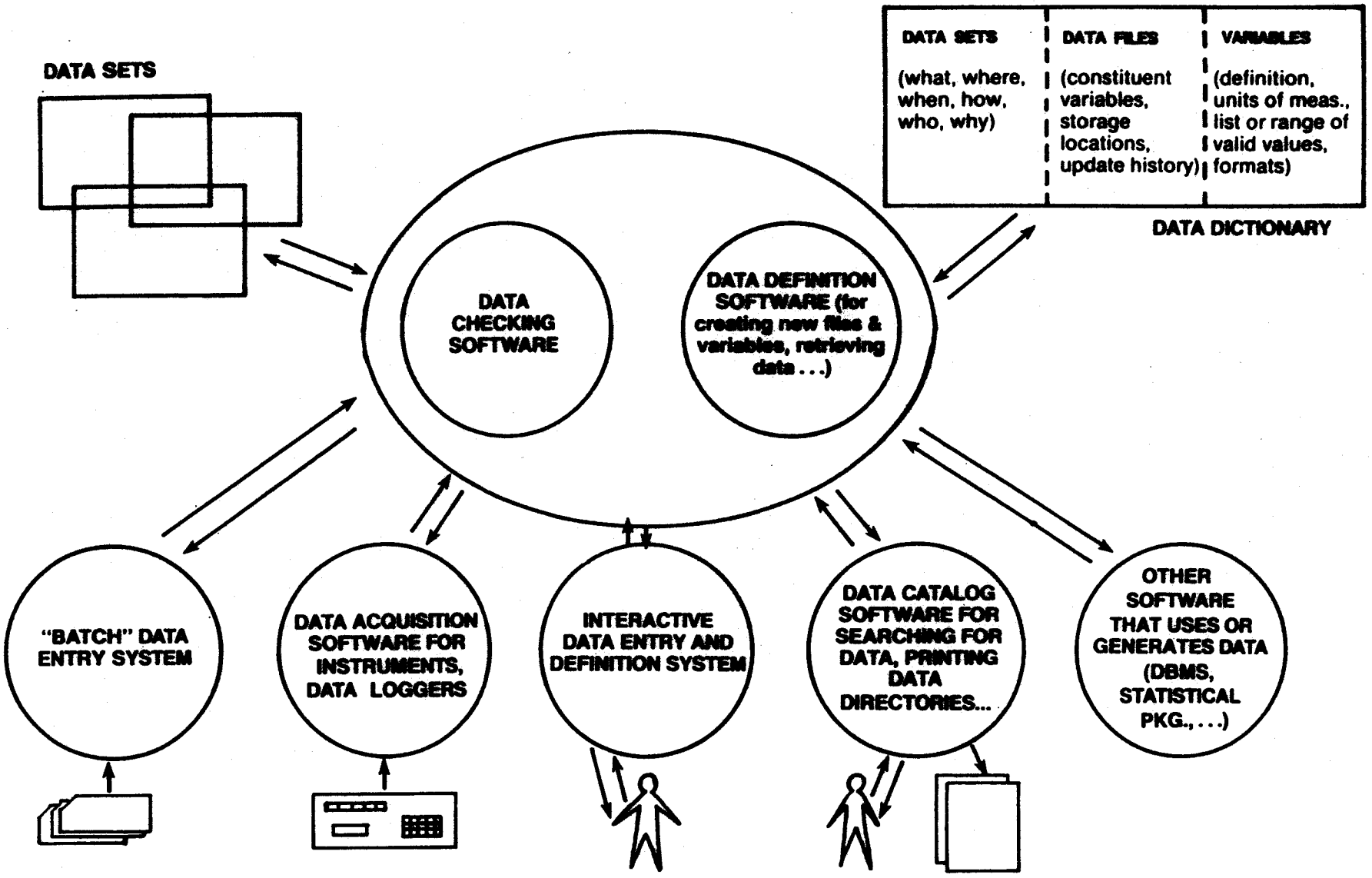


Figure 2. Role of a data entry system and data dictionary in Research Data Management.
A core of data entry/data dictionary software acts as a filter to ensure consistency and documentation of all data, and serves as a common means of access to information needed in order for humans and software to use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

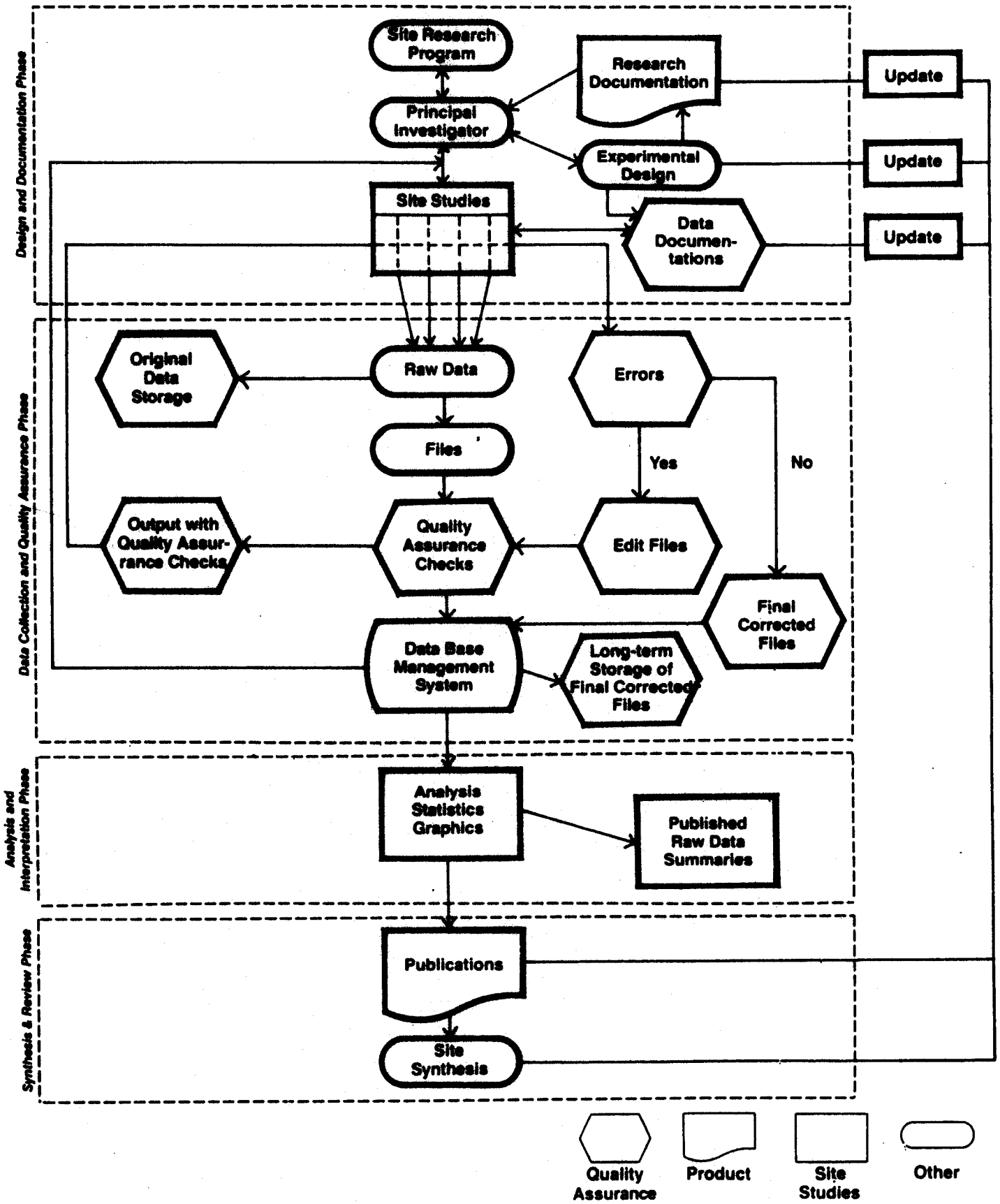
COMPUTER SYSTEM SELECTION

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. Non-computerized 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.

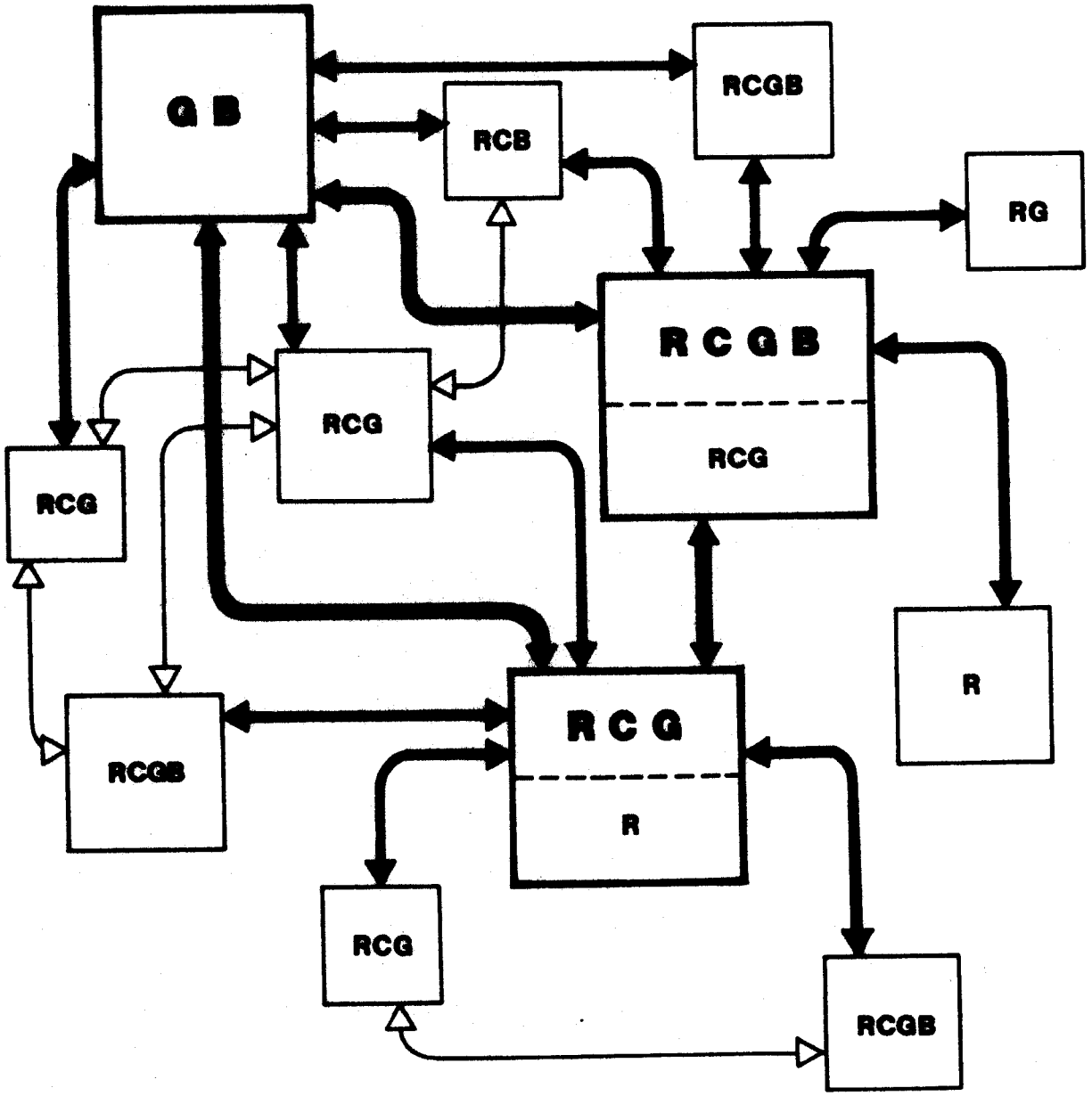
A few secondary institutions could serve as data banks, or repositories, for data that have no other means of long term care. There are important data sets, often the result of work by researchers now deceased, that cannot be cared for properly at field stations or on college campuses.

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 = Data banks



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. 27 pp.

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 Farrell
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)