Reproduced by permission of the University of South Carolina Press from Research Data Management in the Ecological Sciences edited by William K. Michener (Belle W. Baruch Library in Marine Science Number 16). 1986.

DATA MANAGEMENT PROCEDURES IN ECOLOGICAL RESEARCH*

S.G. Stafford, P.B. Alaback, K.L. Waddell, and R.L. Slagle

ABSTRACT

Ecological research requires a flexible, organized system for acquiring, documenting, and managing data. Careful documentation of data, best done through close collaboration of the researcher and data manager, is important if all users are to benefit from a centralized database management (DBM) system. A systematic approach to data management and analysis comprises four key steps: comprehension, planning, execution, and evaluation and interpretation. Statistical consulting at the beginning of a project helps scientists plan well-designed, efficient research strategies. Data collection forms should be designed to encourage recorders to enter all essential identifying information, thereby minimizing errors. Consistent, high-quality data verification, through either the visual or double-entry method, allows most data collection errors to be caught and corrected before analysis. Standardized forms help maintain uniformity in data documentation. Documentation and corresponding data files should be linked in a carefully organized relational DBM system in which all information may be easily stored and retrieved. Using commercial software for data management and analysis is usually more cost effective than developing and maintaining customized software. Scarce DBM resources should be invested in data validation, equipment

*FRL 1944, Forest Research Laboratory, Oregon State University, Corvallis. The mention of trade names or commercial products does not constitute endorsement or recommendation for use.

acquisition, system design, and data documentation, not programming. Database maintenance is costly, but critical, to keeping credibility with users. The key to future advances in ecological data management lies in the ability of data managers and scientists to move into a more cooperative, integrative mode through which comprehensive databases are established and more fully used, increasing overall research efficiency.

INTRODUCTION

Most ecological studies produce a wide variety of data over a period of years. These data are expensive to collect. The better they are documented, the more likely they are to be useful to other scientists presently and in the future. Research data management is a serious attempt to anticipate future needs.

Different types of data are normally stored in separate data files, all of which constitute the data set for a given study. Each data file is documented in a systematic process that identifies and records its structure and origin; this information is entered into the data catalog. The result is the organized storage of data files and corresponding documentation in a logical, easily retrievable format. All data and documentation can then be permanently archived into a central data bank so that a scientist, even one unfamiliar with the research, can readily obtain data for comparison with current, related studies, support of a new hypothesis, or synthesis with other studies. The methods used to organize, store, and retrieve data in this fashion make up the local database management (DBM) system.

The Forest Science Data Bank (FSDB), which serves the Department of Forest Science at Oregon State University, has evolved over more than a decade. Faculty and students have collected sizable amounts of data from ecological studies conducted throughout the Northwest. In addition, research at the H.J. Andrews Experimental Forest, a National Science Foundation (NSF) Long-Term Ecological Research (LTER) site, has produced vast quantities of data now stored with documentation in the FSDB (Stafford et al., 1984). The prime reason for NSF's emphasis on sound, reliable data management in its LTER program has been to help overcome the difficulty experienced in the past of finding appropriate data sets and coordinating research from multiple institutions or geographic regions. Success or failure of an LTER DBM system

may well rest on its ability to provide a uniform standard of documentation and facilitate accurate transfer of experimental data among all 11 LTER sites. Thorough study documentation enhances the understanding and accessibility of valuable research results and allows for greater research efficiency and opportunities.

In this paper, we give an overview of the systematic approach we have developed for retrieving past data and managing current data. Specific DBM procedures, as they relate to the Andrews LTER site, are described. Procedures at other LTER sites are referred to when they are known to differ from those at the Andrews LTER site.

OUR SYSTEMATIC APPROACH: AN OVERVIEW

The FSDB personnel include a consulting statistician, data managers, data analysts, and a microcomputer specialist, all with a combination of statistical, ecological, and biological training. Through time, we have seen how some research strategies succeed and others fail. From this experience we have evolved our own philosophy, a systematic approach to data analysis and management, comprising four key steps: comprehension, planning, execution, and evaluation and interpretation (Fig. 1). The approach itself is a modification of a more general schema currently used in some teaching programs (Chervany et al., 1980).

The first step involves recognizing a specific problem as one instance of a general category of problems and organizing the approach to find a solution. The experimental design needed to test the problem must be understood so that statistically valid data will be available for final analysis. Problems that might have been avoided had a statistician been consulted before the experiment was initiated often develop during analysis of a poorly designed experiment.

During the second step, the strategy (plan) for analysis should be formulated (Fig. 2) and then critically reviewed. The data bank's involvement at this point encourages careful thinking and precise planning relative to experiments that are long-term and expensive. A complete set of documentation should be maintained as a permanent record for future reference and to ensure consistency over time. As research progresses, changes in the design should be reviewed and documented; some changes are inevitable because

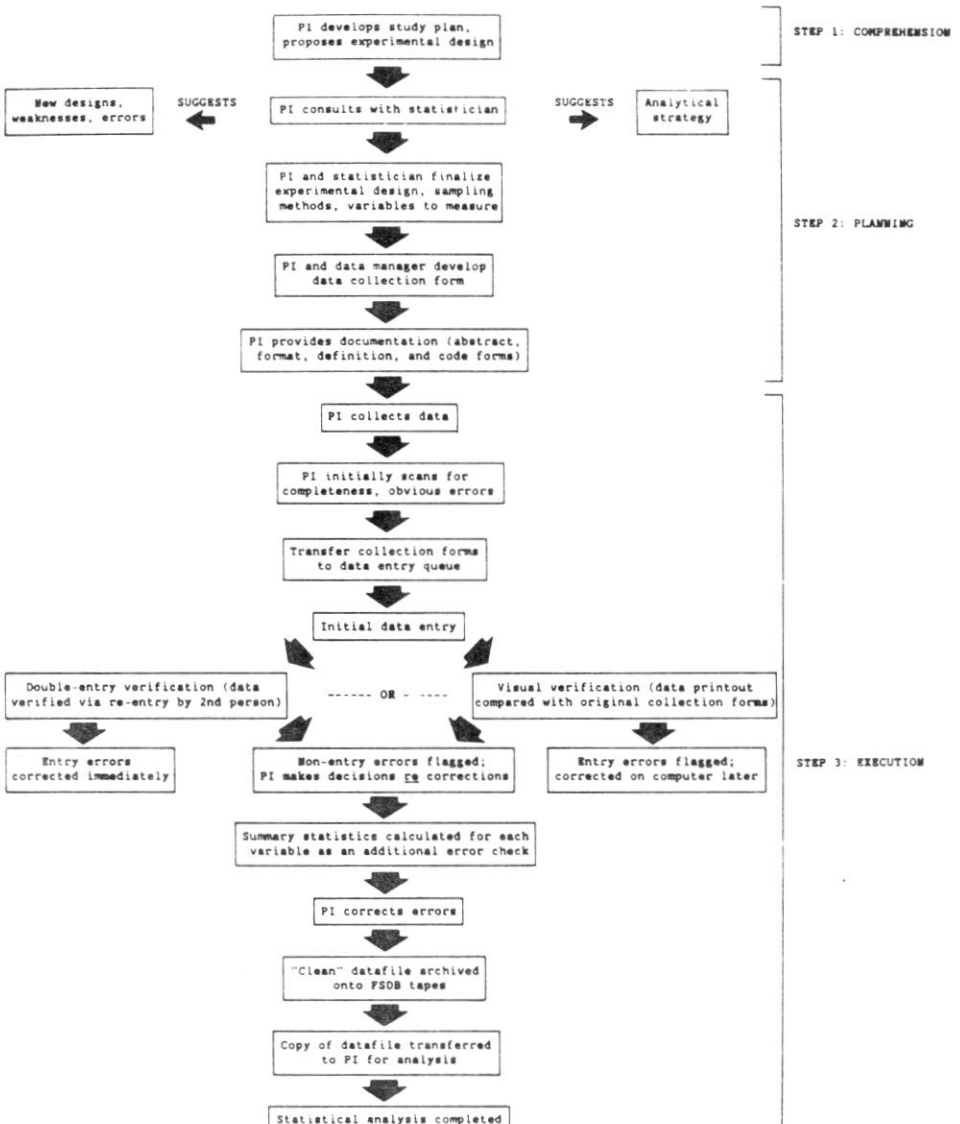

Fig. 1. Flow chart for efficient data collection, documentation, and analysis at the Andrews LTER site. The procedures are blocked according to the four steps specified by the FSDB's systematic approach to data analysis and management.

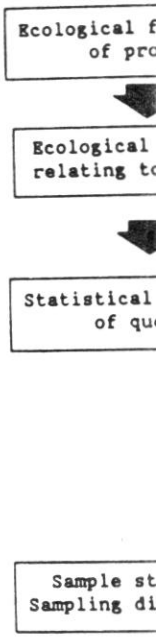

Fig. 2. A s
res
et

many facets
foresee or ar

Once da
should be va
structure wi
problems can
plex, and co
is then anal
procedures) t
this process

Often,
procedures a
age used; fo
al., 1975) r
SAS (SAS Ins
may be so co
into a serie

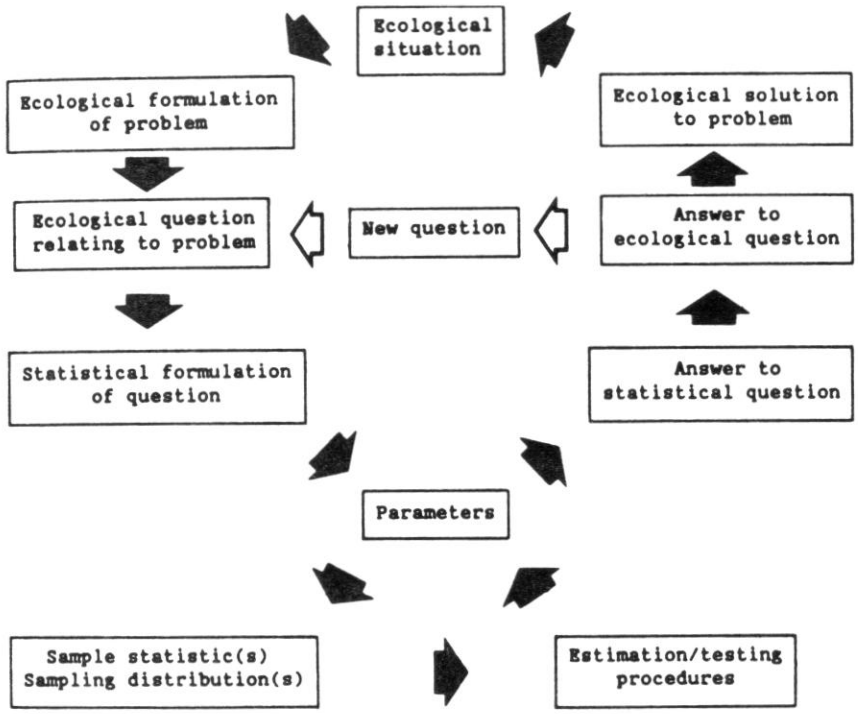

Fig. 2. A strategy for sound, computer-aided analysis to resolve ecological problems (adapted from Chervany et al., 1980).

many facets of ecological experimentation are impossible to foresee or are beyond direct human control.

Once data have been collected and entered, data files should be validated and summarized to cross-check data file structure with the documented experimental design. Many problems can be caught at this stage, before extensive, complex, and costly analyses have begun. The clean data file is then analyzed with a statistical procedure (or group of procedures) to test the experimental hypotheses of interest; this process is the heart of the third step.

Often, the data file structure and sequence of analysis procedures are influenced by the statistical analysis package used; for example, the MANOVA procedure in SPSS (Nie et al., 1975) requires the data to be more structured than does SAS (SAS Institute Inc., 1982). The experimental analysis may be so complex and lengthy that it should be broken apart into a series of smaller analyses, completed individually.

We recommend starting simply, working with one component at a time and reviewing the initial results, before undertaking the entire analysis.

In the fourth and final step, each set of results should be scrutinized to ensure that variables were read correctly and that procedures were appropriate for the intended statistical tests, and to determine if new procedures are suggested by the results. Once the final output is obtained, a statistician can help the researcher interpret analytical results in an ecological context. This interpretation should be the basis for discussion in the final presentation of the data.

Long-term ecological research poses different types of statistical problems than traditional agronomic or agricultural experiments. Most ecologists recognize the difficulties of specifying appropriate statistical criteria for replication and determining appropriate error rates in a world full of unique sites and situations (Hinds, 1984; Hurlbert, 1984). Statistical consultation through all four steps of the systematic approach will help ensure that statistically sound, credible inferences be drawn.

THE STATISTICIAN'S ROLE

Data produced from any research project must be collected according to a sound experimental design. Whether the design is simple or complex, thorough analysis can be completed only with statistically valid data.

We recommend consultation with a statistician to confirm that the field or laboratory procedure is consistent with the statistical design and study objectives (Fig. 1). Suggestions for analysis, study weaknesses, or potential sources of error can be pointed out and dealt with early on. Consultation usually benefits both parties: the researcher learns how all phases of experimental research and attendant assumptions of statistical inference are interrelated and how they will most likely affect the project's outcome, and the statistician becomes more familiar with the research project as well as unavoidable constraints on design. Data bank personnel and scientists should work closely together to chose variables, design data collection and documentation forms, and anticipate future analyses so that data can be properly documented as permanent, logically arranged files in the data bank.

DESIGNING DATA

Careful de
rate, efficient
field or labor
sampling inten
sured should b
forms should b
courage the re
Many items are
promise quality
tical analyses.

Media use
include analog
paper forms.
forms. Althoug
ized, data ba
develop a pro
extended to ev
the following:

- The most
  variable
- All per
  ment le
  sample
- Simplic
  observa
  contain
  value e
- Additio

The real
capture the ne
collecting pha
tions, and fo
form's success
and data manag
crucial. Term
a forest mensu
to 8.5 inches
mentation mus
diameter clas
question of p
able allocate
really demand
greater preci

DESIGNING DATA COLLECTION FORMS

Careful design of data collection forms fosters accurate, efficient data collection and a smooth transition from field or laboratory to the computer. Experimental design, sampling intensity, periodicity, and variables to be measured should be determined before forms are designed, and forms should be developed before data are collected to encourage the recorder to enter all essential information. Many items are neglected when no form exists, which can compromise quality control and later prevent complete statistical analyses.

Media used to record data in the field or laboratory include analog, digital, or audio electronic devices and paper forms. The most common are project-specific paper forms. Although these forms cannot be completely standardized, data bank staff and scientists at each site should develop a prototype form whose general structure can be extended to every study and which should take into account the following:

- The most efficient order for collecting data for each variable.
- All pertinent experimental information (e.g., treatment level, location, site, plot or subplot, block, sample date).
- Simplicity and legibility; areas for recording observations should be clearly labeled and should contain adequate space to accommodate the maximum value expected.
- Additional space for recording pertinent comments.

The real test of how well a data collection form can capture the needed information comes during the active data-collecting phase, which tests all the predictions, assumptions, and formats imposed on the data by that form. The form's success will partly depend on how well the researcher and data manager communicate; an "open line" between both is crucial. Terminology often is misunderstood. For example, a forest mensurationist traditionally classifies a tree 7.6 to 8.5 inches in diameter as an 8-inch tree; the study documentation must distinguish this value as within an 8-inch diameter class and not a true decimal measurement. The question of precision must be posed. For instance, a variable allocated three spaces as an integer value, but which really demands two or three additional decimal places for greater precision, may be squeezed into too small an area,

compromising legibility, or, even worse, rounded off to fit the space provided. Poorly designed forms increase the likelihood of errors in the data and unnecessarily complicate the data entry procedure. Moreover, transcribing values onto a more appropriate form introduces additional potential for error. If possible, briefly test a new form before deciding on its final design.

A data collection form is only a tool--a means, not an end. Occasions will arise when experimental conditions change or when assumptions about the data are proven incorrect. For example, mortality of a tree may entirely eliminate a species, or the number of places allocated to a particular variable may need to be increased due to unexpected extreme values. Both of these situations require only a simple modification of the form, but severe changes during data collection may require that an entirely new form be created.

Direct electronic input from analog or digital devices is becoming increasingly common for meteorological and laboratory data. These instruments can automatically record large quantities of data at set time intervals on digital media such as cassettes or EPROM (erasable programmable read-only memory) microchips. The hardware-dependent, limited storage capacity of most of these devices produces data files in a condensed format that often cannot be read by the data collector. Documentation for these types of data should include a textual header, entered at the beginning of each data file, that specifies, for example, the type of experiment, date, time, people involved, measurement interval, sensors or instruments used, and output data format. If calibration data are required to convert raw data to useful measurement units, the source of this calibration data and the calibration function should also be included.

Special problems occur when re-measurement data (e.g., annual tree height and diameter growth measurements) are being collected during a long-term research project. If a new or updated form is designed for the re-measurement sample, the existing data and documentation should be reviewed to assure that the subsequent (new) data file will be compatible with the older one. It is often useful to have the original data on the collection sheet in the field to serve as a baseline for verifying the new data. Commercial or custom software can efficiently produce a collection sheet containing the previous measurement, with adjacent spaces to record re-measurements.

ENTERING AND VE

Before da
searcher revie
ness and other
forms are unacc
researcher may
Occasionally, i
ized, transcrip
it is and ref
costly, this
errors. Easy-t
for transcribin
or older data s
ized, and refo
commercially av
fically for da
cess by allowin
lection form on

The resea
corded and comp
or code check i
data set. For
collected data
checked for err
of data entry.
should not be r
even when it ma
a numeric field

Two syste
entered: doub
tion. Double
completely by
a second perso
person may be
This is the s
data and is fr
tions. If a
match that of
match and the
data-entry per
to know how to

Visual ve
cially when in
data file prin

ENTERING AND VERIFYING DATA

Before data are entered into the computer, the researcher reviews every data collection sheet for completeness and other obvious errors (Fig. 1). If the collection forms are unacceptable for data entry, the data collector or researcher may have to transcribe the data onto new forms. Occasionally, if the collection forms are not too disorganized, transcription can be avoided by entering the data as it is and reformatting later. Though time consuming and costly, this route reduces the number of transcription errors. Easy-to-use DBM software tends to minimize the need for transcribing poorly formatted data because data from new or older data sets can be efficiently restructured, reorganized, and reformatted after it is entered. A variety of commercially available software packages are designed specifically for data entry. Many simplify the data entry process by allowing the user to reproduce the actual paper collection form on the computer's video display.

The researcher is responsible for the accuracy of recorded and computerized data. For many LTER sites, a range or code check is built into the data entry software for each data set. For others (including much of the Andrews LTER), collected data are rapidly entered into the computer and checked for errors later, in the second (verification) stage of data entry. In any case, the person entering the data should not be responsible for correcting faulty information, even when it may be obviously wrong (e.g., an alpha value in a numeric field).

Two systems are commonly used to verify the data being entered: double-entry verification and visual verification. Double entry requires that the data set be entered completely by one person, then reentered (i.e., verified) by a second person. Two different people are used because one person may be more likely to make the same mistake twice. This is the system used by the FSDB for virtually all new data and is frequently used by commercial data-entry operations. If a value typed during the second entry does not match that of the first entry, the computer signals a mismatch and the error can be corrected. However, the second data-entry person will have to consult with the researcher to know how to make the correction.

Visual verification is widely used on LTER sites, especially when initial costs and time prevent double entry. A data file printout is compared with the original collection

sheets; any errors or changes are flagged and edited later. In our experience, and that of many large businesses, double-entry verification is the most straightforward, accurate system and is preferred for critical data sets.

DEVELOPING DATA DOCUMENTATION FORMS

Standardized documentation containing all pertinent information describing data collection is an essential component of a DBM system. A standard set of documentation forms, which may be created by DBM software or hand drawn on paper forms, should be developed for each research organization. Most LTER sites currently use the paper forms.

Documentation initially includes variable definitions, formats, measurement units, codes, and research methods and should be permanently recorded long before data collection begins. All information contained on the finalized data-collection form should be documented as soon as possible after actual data collection to avoid the problems incurred when the researcher cannot recall details as time passes. Summary handouts describing how to fill out each form make it easier for newcomers to the system to provide complete documentation and understand its rationale. Again, the most desirable forms usually are produced when the researcher and data manager collaborate.

Every study should be assigned a code or brief title (often called a data code) to uniquely identify its documentation and data files. If multiple files are produced because data types are diverse, each file should be documented with a unique name easily related to the data code. The Andrews LTER site uses a format-type identifier after the data code number to name each file in the study. To develop a successful relational database (a coding system linking all data files and supporting documentation with data location referenced within a catalog or directory), all documentation pertaining to each file should be labeled with the data code and format-type identifiers. Ideally, before the data manager agrees to assume responsibility for permanent data storage, documentation should be complete.

Every format type within a data code should be documented with a complete set of the following: (1) forms recording variable names and formats, with units of measurement; (2) forms explaining variable definitions and their acronyms; (3) forms specifying the meaning of each code for a given variable; (4) forms identifying the names of all

stored data files under one data code and their description; and (5) a form abstracting the study purpose, goals, locations and availability of supporting documentation, and methodology for additional background.

For example, a long-term study investigating the effects of fertilization, crown pruning, and stand density on the population dynamics of a young forest stand at the Andrews LTER site was begun in 1981 (Perry, 1982). Various measurements are already stored within several different data files, each with a unique format type. Complete documentation, filed with the FSDB, is described here to demonstrate the database documentation system. We anticipate that the resulting database will be continually updated. The data will be analyzed periodically and will serve as a baseline for other studies and comparative analysis. For example, a researcher deciding to study the population dynamics of a forest 10 years from now could retrieve this data set and its documentation to aid in the planning phase of that new study. The study plan underwent critical statistical review, and the data collection forms, with supporting documentation, were developed before data collection was begun.

The data code assigned to this study is TP88, a unique identifier recorded on every documentation form associated with this project. Relevant details are summarized on the abstract form (Fig. 3). The permanent file name and a brief description of the data stored in each file are listed on the data file description form (Fig. 4). Each file contains a different set of measured variables; thus, variable formats (Fig. 5), definitions (Fig. 6), and code specifications (Fig. 7) are documented on separate sets of forms. To save space here, documentation for only one format type is provided. Although data collectors and analysts will come and go during the course of this long-term study, the fully documented files should pose no major problems to the successors.

STORING DATA AND DOCUMENTATION

Data files and study documentation can be permanently stored on a variety of media: paper files, microfiche, floppy diskettes, punched cards, hard disks, laser disks, EPROM chips, mainframe disk packs, and magnetic tapes or cards. For example, the FSDB currently stores all data on magnetic tapes on the University mainframe computer and has

RESEARCH STUDY ABSTRACT

DATACODE TP88 PRINCIPAL INVESTIGATOR: PERRY, D.A.

STUDYID — PROJECT FUNDING: LTER GRANT

STUDY TITLE: POPULATION DYNAMICS OF YOUNG FOREST STANDS AS AFFECTED BY DENSITY AND NUTRIENT REGIME

OTHER RESEARCHERS INVOLVED: SCHROEDER, P.; CHOQUETTE, C.
CONTACT PERSON: PERRY, D.A.

DATA COLLECTION PERIOD: Begin 81/08/01 End __/__/ONGOING

KEYWORDS: POPULATION DYNAMICS, FERTILIZATION, PRUNING, FEEDBACK MECHANISMS, LOGISTIC GROWTH, STAND DENSITY, LTER

PARAMETERS/MEASURED VARIABLES: SPECIES, DBH, HEIGHT, SAPWOOD AREA, WOOD INCREMENT

VEGETATION ZONE: TSHE

PLANT COMMUNITIES IN STUDY AREA: TSHE: RHMA: GASH, TSHE: BENE: GASH, TSHE: ACCI: BENE

TAXA (List Scientific Abbreviations of Plant Species studied):
PSME, TSHE, THPL

SOIL TYPE: NA

DETAILS OF SITE CHARACTER: 2000-3050 FT ELEVATION RANGE, MESIC SITES, 20-50% SLOPE RANGE, BROADCAST BURNED IN 1959 FOR SITE PREP., 2 SITES ON N-FACING AND 2 SITES ON S-FACING SLOPES.

RESEARCH AREA/REGION: OR-WESTERN CASCADES, H.J. ANDREWS EXPERIMENTAL FOREST, BLUE RIVER RANGER DISTRICT, WILLAMETTE NATIONAL FOREST.

PERMANENT PLOT NAMES: HJA STAND UNIT NUMBERS L107, L405, L701, L111

DATACODE TP88

STUDY PURPOSE AND GOALS: TO TEST HYPOTHESES CONCERNING FEEDBACK MECHANISMS (COMPETITION AND MORTALITY) WHICH AFFECT FOREST STAND DEVELOPMENT. SITE QUALITY WILL BE MANIPULATED VIA FERTILIZATION, AND LEAF AREA WILL BE ALTERED VIA PRUNING. GROWTH WILL BE MODELED AS STANDS APPROACH SELF-THINNING.

EXPERIMENTAL/SAMPLING DESIGN: A SPLIT-SPLIT PLOT DESIGN WITH 4 REPS., 3 DENSITY LEVELS (WHOLE PLOT), 2 FERTILIZATION (SPLIT PLOT) AND 2 PRUNING TREATMENTS (SPLIT-SPLIT PLOT). ONE PERMANENT PLOT WILL BE ESTABLISHED WHERE HEIGHT, DIAM., AND SAPWOOD AREA WILL BE SAMPLED.

EXPERIMENTAL METHODS: DENSITY TREATMENTS THINNED TO: (1) 15% OF MAX. STAND DENSITY; (2) 25% MAX. STAND DENSITY; (3) NO THINNING. FERTILIZATION TREATMENTS: (1) 50 kg/ha/yr OF UREA ADDED UNTIL CURRENT FOLIAGE REACHES 2% N; (2) NO UREA. PRUNING: TREE BRANCHES REMOVED TO ATTAIN CERTAIN CROWN WIDTHS.

REFERENCE CITATIONS--(Only those that cite data, relate to methods, etc.):

Num.	Year	Authors	Title	Journal and Vol.
1	1981	PERRY, D.A.	LTER STUDY PLAN-POPULATION DYNAMICS OF YOUNG FOREST STANDS	AVAILABLE FROM AUTHOR

Fig. 3. The research study abstract form documenting ecological studies may request items such as study location, habitat type, site and soil characteristics, experimental design, analytical methodology, and keywords. The contents of this form should provide a reasonable overview of the experimental procedures and study objectives.

DATACODE
STUDYID
STUDY TIT
PRINCIPAL
OTHER PER
FORMAT TY
FORMATS
DEFINITIO
CODES
FORMAT
TYPE

1	TP
2	TP
3	TP
4	TP
5	TP
6	TP
7	TP

Comments

Fig. 4. The
ize
co
bri
ea

recently con
floppy and
location and
accessibilit
well as numb
mation. In
secure backu
burned, or o
our FSDB tap
the Universi
also is an
disks deteri
and magnetic
periodic use

Fig. 3. The research study abstract form documenting ecological studies may request items such as study location, habitat type, site and soil characteristics, experimental design, analytical methodology, and keywords. The contents of this form should provide a reasonable overview of the experimental procedures and study objectives.

DATAFILE DESCRIPTION FORM
(FSDB)

841010

DATACODE TP88 DATE 83/10/11 (yr/mo/da) ONE-TIME COLLECTION
STUDYID – RECORDER KW,CC LONG TERM STUDY ✓
DATA HEADERS EXIST —

STUDY TITLE POPULATION DYNAMICS OF YOUNG FOREST STANDS AS AFFECTED BY DENSITY AND NUTRIENT REGIME
LTER ✓ CRAFTS — FIR —

PRINCIPAL INVESTIGATOR PERRY, D.A.

OTHER PERSONS INVOLVED SCHROEDER, P.; CHOQUETTE, C.

DOCUMENTATION STATUS (✓)

FORMAT TYPES →	1	2	3	4	5	6	7	8	9	10
FORMATS	✓	✓	✓	✓	✓	✓	✓	—	—	—
DEFINITIONS	✓	✓	✓	✓	✓	✓	✓	—	—	—
CODES	✓	✓	✓	✓	✓	✓	✓	—	—	—

FORMAT TYPE	DATAFILE NAME	BRIEF DESCRIPTION OF DATAFILE CONTENTS
1	TP8801	PRE-AND POST-THINNING DENSITIES ON PLOTS
2	TP8802	INITIAL TREE DIMENSIONS
3	TP8803	DELAYED HEIGHT MEASUREMENTS AND CHANGES
4	TP8804	TREE GROWTH MEASUREMENTS
5	TP8805	EXTRA TREE DATA OUTSIDE EXPERIMENTAL AREA
6	TP8806	COMPETITION DATA FOR SAMPLE TREES
7	TP8807	MINERALIZABLE NITROGEN SAMPLES

Comments: YEARLY MEASUREMENTS ARE EXPECTED

Fig. 4. The data file description form contains an organized list of all data files pertaining to one data code (study). Individual data file names and a brief description of the type of data stored in each file are recorded.

recently converted storage of paper file documentation to floppy and hard disks on a microcomputer. The storage location and media chosen depend primarily on computer accessibility, available software, ability to store text as well as numbers, and cost of storing and retrieving information. In any case, it is important to have adequate, secure backups in the event that originals are lost, stolen, burned, or otherwise destroyed. We maintain four copies of our FSDB tapes, one of which is stored at a location outside the University computer center. The durability of media also is an important consideration. Magnetic tapes and disks deteriorate after extensive use. Cards, paper tapes, and magnetic tapes become brittle with time and require periodic use, cleaning, or replacement.

Initially, we store most study documentation in paper files, where completed forms, sample field collection sheets, the study plan, and other related information are maintained. Some DBM systems store much of their documentation as a header at the beginning of each data file. The storage device for this type of file must be able to handle text efficiently and search for and retrieve specific information. Many systems have specialized software for relational databases, but in our experience, mainframe software is too expensive to run, and disk storage costs are too high to leave documentation continuously online. If files must

841010

VARIABLE FORMAT FORM (FSDB) Page 1 of 1

DATACODE TP88 DATE 10/11/83 (mo/da/yr) RECORDER KW, CC.

FORMAT TYPE 4

STUDYID —

DATA TITLE: TREE GROWTH MEASUREMENTS

VAR. NUM.	VARIABLE NAME	COLUMNS OCCUPIED	FORTRAN FORMAT	CODED (✓)	UNITS	MISSING VALUE CODE
1	DATACODE	1-4	A4			ALL "BLANK"
2	FORMTYPE	5-6	I2			
3	SITE	7-11	1X,A4			
4	DENSITY	12-13	1X,A1	✓		
5	TMT	14-16	1X,A2	✓		
6	YRMODA	17-23	1X,I6			
7	TREENUM	24-27	1X,I3			
8	DBH	28-31	1X,F3.1		cm	
9	HT	32-35	1X,F3.1		m	
10	SAPWI	36-39	1X,I3		mm	
11	GRINCI	40-42	1X,I2		mm	
12	PRERING1	43-46	1X,I3		mm	
13	SAPW2	47-50	1X,I3		mm	
14	GRINC2	51-53	1X,I2		mm	
15	PRERING2	54-57	1X,I3		mm	
16	TREECOND	58-59	1X,I1	✓		
17	BARK1	60-62	1X,I2		mm	
18	BARK2	63-65	1X,I2		mm	

CONTINUED, Reverse side —

Fig. 5. The variable format form delineates the exact structure of a data file by identifying the columns occupied by each variable, the precise format, and related information. Each variable is assigned an acronym (variable name no greater than eight characters) that remains constant on every documentation form.

first be retr
inevitable and
ware is used.
computers for
mentation and
files themselve
tion files on t
file name, and
library on the

CATALOGING AND

Paper file
mentation form
invaluable for
ple, with a co
abstracts of

841010

DATACODE
FORMAT TYPE
STUDYID

VAR. NUM.(1)	
1	F
2	F
3	H
4	S
5	T
6	\
7	T
8	1
9	T
10	S
11	G
12	P
13	S
14	G
15	I
16	T
17	
18	

(1) Use

Fig. 6. The
eac
men

first be retrieved from magnetic tape, time delays are inevitable and additional costs accumulate before the software is used. We have relied on publicly accessible microcomputers for organizing, storing, and retrieving data documentation and the University mainframe for storing the data files themselves. Interested users can scan our documentation files on the microcomputers, note the data code or data file name, and retrieve the file(s) from our magnetic tape library on the mainframe.

CATALOGING AND RETRIEVING DATA AND DOCUMENTATION

Paper files may be suitable for storing completed documentation forms, but a computerized relational database is invaluable for quick search and retrieval tasks. For example, with a computerized database, a file containing study abstracts of all projects currently being managed may be

841010 **VARIABLE DEFINITION FORM** Page 1 of 1

(Complete one form for each format type)

DATACODE TP88 DATE 10/11/83 (mo/da/yr) RECORDER KW, CC

FORMAT TYPE 4

STUDYID -

VAR. NUM.(1)	BRIEF DESCRIPTION OF EACH VARIABLE	PRECISION
1	FSDB DATASET CODE	
2	FSDB FORMAT TYPE	
3	HJ ANDREWS SITE IDENTIFICATION CODE	
4	STOCKING-DENSITY LEVEL FOR WHOLE PLOT TRTMT.	
5	TREATMENT-FERTILIZED OR PRUNED	
6	YEAR-MONTH-DAY OF SAMPLING	
7	TREE TAG NUMBER	
8	DIAMETER AT BREAST HEIGHT	±0.1cm
9	TOTAL TREE HEIGHT	±0.1m
10	SAPWOOD WIDTH - DIRECTION #1	± 1mm
11	GROWTH INCREMENT-LAST TWO YEARS, #1	± 1mm
12	PRE-THINNING RING WIDTHS-RINGS 4-8 COMBINED, #1	± 1mm
13	SAPWOOD WIDTH - DIRECTION #2	± 1mm
14	GROWTH INCREMENT-LAST 2 YEARS, #2	± 1mm
15	PRE-THINNING RING WIDTHS-RINGS 4-8 COMBINED, #2	± 1mm
16	TREE CONDITION-DAMAGES	
17	BARK THICKNESS-DIRECTION #1	± 1mm
18	BARK THICKNESS-DIRECTION #2	± 1mm

(1) Use exact numbers from the corresponding Variable Format Form.

Fig. 6. The variable definition form briefly describes each variable and specifies the desired measurement precision.

841010 VARIABLE CODE FORM Page 1 of 1

(Complete one form for ALL format types)

DATACODE TP88 DATE 10/11/83 (mo/da/yr) RECORDER KW,CC

STUDYID —

VARIABLE NAME	CODE VALUE	BRIEF DEFINITION OF EACH CODE VALUE
DENSITY	C	CONTROL
	L	REMOVAL OF 15% MAX. STAND DENSITY
	M	REMOVAL OF 25% MAX. STAND DENSITY
TMT	C	CONTROL - NO UREA OR PRUNING
	F	FERTILIZED WITH UREA
	P	PRUNED TO SPECIFIC CROWN WIDTHS
	FP	FERTILIZED AND PRUNED
TREE COND	0	LIVE
	1	DEAD
	2	BROKEN TOP AND DEAD
	3	STEM DEFORMITY DUE TO SNOW
	4	ROOT ROT (PHELLINUS) INFECTION
	5	BROKEN TOP - WILL LIVE
	6	LEANING TREE
	7	ANIMAL DAMAGE - BASAL SCARS
	8	ANIMAL DAMAGE - IN CROWN

CONTINUED, Reverse side ___

Fig. 7. The variable code form describes the codes allocated to constant nonmeasurable variables or categorical variables (used where each observation is classified into a predefined category). Specific codes should be identified before active data collection begins.

searched for a particular topic. Once the appropriate abstract is found, the data code or name is extracted and used as a key variable in the database. The most efficient method of referencing paper files, if used, would again be by data code or name. If documentation can be centralized on a computer, commercial DBM software packages can be used. Because these packages are more user friendly, DBM personnel need be only minimally involved in routine searches and retrievals. Once the desired information has been located, the computer can produce a printed report fully describing the data sets.

Ease of retrieval depends on the existence of a relational system for cataloging information for a given data

code and on the access mode of the storage medium: random access (disks) or sequential access (magnetic tapes). Software that simplifies data retrieval is available for data files permanently stored on magnetic tape. Currently, researchers at the Andrews LTER site use a program that requests the name of the data file to be retrieved, automatically creates a batch file that mounts the proper tape, retrieves the file, and stores a copy on a requested computer account, all without DBM staff assistance. If data are stored on disk or a similar random-access device, the optimal strategy is to tie data and documentation together so that retrieving both is one basic process. Researchers using the SIR system (Robinson et al., 1979), for example, can retrieve data-structure documentation easily with a query language. Data files can also be stored with a documentation header, which allows documentation and data to be retrieved as one unit, but this approach may require more custom software development or be incompatible with some software.

SOURCES OF ERROR

A popular misconception is that errors first occur when the data are collected. Actually, problems potentially begin much earlier, with the design and documentation of data collection forms. Detectable errors may still exist even after data files are in final form. A preliminary data screening analysis, designed to further validate the data, will identify these errors. Validation analysis serves two purposes: first, to validate the correspondence between data and associated documentation; second, to identify data collection and entry errors where possible. We have found that data screening always pays. Locating and correcting as many errors as possible before a full statistical analysis is undertaken prevents costly reruns and wasted time.

During validation analysis, the documented data file format is compared with the actual data file format. For discrete variables, a list of frequencies for each code is generated. For continuous variables, the following statistics are produced: minimum, maximum, mean, number of valid observations for each variable, and total number of observations in the data file. The analysis reveals where incorrect codes were used or where data were out of the expected range. These errors, which may have occurred during field collection, data entry, or recording of the documentation, are then identified and located on the collection forms so

that the researcher can decide how to correct them. Validation checks have been incorporated into DBM systems at several LTER sites to catch errors when the data are entered. At other sites, graphical displays identify possible errors. The most elaborate systems produce printouts of raw data with annotated comments, which point to possible errors. Once again, responsibility for changing faulty values lies with the researcher.

After data entry, data files frequently need to be reformatted and then transferred to another medium or storage device. These procedures should be done by data bank personnel to avoid the possibility of introducing "unprotected errors" (errors created by a program or by another logical error that could produce a major, but virtually undetectable, error in the data file). For example, merging two data sets for which there is no one-to-one matching variable (which exactly identifies the matched cases) can create unprotected errors. A successful merge requires that the sorted order of each data file be identical. If cases are out of order or missing, the completed merge may appear to be normal, but actually may have produced undetectable errors in the data file. Where reformatting and manipulating data are unavoidable, exercise extreme caution.

Though documentation and data entry errors are usually easily corrected in the data file, field or laboratory errors must be handled quite differently. One approach, though not widely recommended, is to leave the data exactly as recorded on the collection forms. In many cases, however, the data file cannot be processed with these errors; for example, an alpha value in a numeric field could cause a program failure. Another approach is to try to update the data where the correction is obvious (and we do mean obvious), for instance, a misspelling. Such corrections, once approved by the researcher, can be made efficiently without running the risk of introducing extraneous or incorrect information. We have found that the best approach concerning faulty information is to set the observation equal to a "missing" code to avoid guessing the correct value; these observations are then excluded from all analyses.

Data collection errors usually are not rectified easily. Continuous variables that are out of the expected range often require field or laboratory validation. If no other validation method is available, the value should be set to "missing." Losing information is always unfortunate, but a value that is out of range can be far more damaging to analysis results than missing information. Similarly, data

that are ille
observation n
tion method is
In a few
a "best appr
mended, but ca
tree diameter
ally for many
was found to
could be appro
the adjacent y
very small. T
consequences of
stance. Proble
data are not a
have been ident
were obtained
do occur when
modifications
changes range
tion forms to e
Another t
the failure to
lead to atypic
way to plan fo
except to prov
are not genera
researcher int
the archiving
microfiche) to
studies in whi
ments are enc
this must be
easily and exc
Sites in
data validati
sophisticated
tated output,
Inc., 1982).
at least once
formation on
sions on what
searcher. Pr
not adequately
must sign or

that are illegible should be entered as "missing" and that observation noted for later evaluation if no other validation method is available.

In a few circumstances, data may be modified purely on a "best approximation" basis. This is not often recommended, but can sometimes have merit. For example, suppose tree diameter on a particular plot had been measured annually for many years, but the value for an intermediate year was found to be completely out of range. That diameter could be approximated according to the values recorded for the adjacent years, and the probability of error would be very small. The alternative is entering "missing." The consequences of each approach should be judged for each instance. Problems like this arise when long-term ecological data are not analyzed yearly (in which case the error would have been identified and corrected at the time) or when data were obtained from other sources. In any case, situations do occur when recorded data must be modified, and those modifications must be documented. Systems for documenting changes range from merely jotting them down on the collection forms to elaborate notation in the actual data files.

Another type of error common to ecological research is the failure to record unusual sampling situations, which may lead to atypical data values. Unfortunately, there is no way to plan for these when designing a data collection form except to provide space for comments. Although the comments are not generally archived with the data file, they help the researcher interpret the outliers in the data. We encourage the archiving of raw data collection forms (usually in microfiche) to preserve this information. In some long-term studies in which re-measurements are frequent, certain comments are encoded in the data files themselves; however, this must be done so that the comments can be separated easily and excluded during analysis.

Sites in the LTER network have functionally similar data validation (quality assurance) systems. The most sophisticated systems produce graphical analysis and annotated output, usually via the SAS system (SAS Institute Inc., 1982). Data bank personnel verify the entry of data at least once at all sites and provide researchers with information on illegal codes and possible outliers. Decisions on what to do about outliers always rest with the researcher. Problems sometimes arise because scientists do not adequately proof and validate data; at some sites, they must sign or initial approval forms after reviewing entered

data. Providing a uniform standard of quality assurance for all archived data remains the most critical test of the usefulness of an ecological data bank.

The issue of data integrity must be emphasized to assure that the data analyzed are a valid, accurate representation of the facts. In this process we often take advantage of the "Law of Large Numbers," which basically states that the larger the sample size, the closer the sample mean will be to the true population mean. To require every bit of data to be an exact representation of reality is unrealistic. A more reasonable aim is to record measurements as accurately as possible, given the experimental constraints and overall research perspective.

One additional source of error remains after documentation, data entry, and validation are completed: errors in the execution of the statistical analysis. These can range from relatively minor mistakes to completely misleading results that may be invalid if the data are not analyzed with the statistical design initially chosen.

CONCLUSIONS

Evolution of hardware and software for data management in recent years has made possible great advances in the flexibility and utility of ecological data management. In our experience, developments in the computer industry have far outstripped ecologists' training to use them. The most substantial barriers to improving data management are our own failures of imagination. New ways of thinking and cooperating in multidisciplinary, long-term research lay the groundwork for future, more fundamental advances in the field. Increased availability of computers, more powerful software that is easier to use, and support from DBM personnel should augment scientists' abilities to improve database quality and utilization. Once credible, reliable, comprehensive databases are established, more integrative data-intensive research projects will be a practical reality for answering the key questions of systems ecology.

LITERATURE CITED

Chervany, N.L., P.G. Benson, and R.K. Iyer. 1980. The planning stage in statistical reasoning. Amer. Statistician 34: 222-226.

Hinds, W.T.
in terre
11-18.

Hurlbert, S.H.
ecological
187-211.

Nie, N.H., C.H
D.H. Brent
Social Sci
pp.

Perry, D.A. 19
stands: Es
Research at
D. In: Pr
Term Ecolo
Forest. Un

Robinson, B.N.,
1979. SIR
edition.
pp.

SAS Institute I
edition. S

Stafford, S.G.
Klopsch.
bank. J. F

Hinds, W.T. 1984. Towards monitoring of long-term trends in terrestrial ecosystems. Environ. Conserv. 11(1): 11-18.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 54(2): 187-211.

Nie, N.H., C.H. Hull, J.G. Jenkins, K. Steinbrenner, and D.H. Brent. 1975. SPSS: Statistical Package for the Social Sciences. 2nd edition. McGraw-Hill, NY. 675 pp.

Perry, D.A. 1982. Population and dynamics of young forest stands: Establishment report for Long-term Ecological Research at H.J. Andrews Experimental Forest. Appendix D. In: Proposal to National Science Foundation Long-Term Ecological Research on the Andrews Experimental Forest. Unpublished. 10 pp.

Robinson, B.N., G.D. Anderson, E. Cohen, and W.F. Gazdzik. 1979. SIR: Scientific Information Retrieval. 2nd edition. Northwestern University, Evanston, IL. 321 pp.

SAS Institute Inc. 1982. SAS User's Guide: Basics. 1982 edition. SAS Institute Inc., Cary, NC. 923 pp.

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