DEVELOPMENT OF A DATA MANAGEMENT SYSTEM FOR LONG-TERM ECOLOGICAL RESEARCH

William K. Michener, Baruch Institute
Robert A. McLaughlin, Baruch Institute
Marvin F. Marozas, Coastal Carolina College

INTRODUCTION

The Belle W. Baruch Institute for Marine Biology and Coastal Research (one of eleven Long-Term Ecological Research sites in the nation funded by the National Science Foundation) has developed a flexible system for ecological data management and analysis. The Baruch Data Management System (BDMS) contains four years of data collected from ecological research on coastal and estuarine habitats. The system is designed to efficiently incorporate data from short-term to multi-decade studies. Specific design features include modular data base design for optimal flexibility in data set addition and modification. Both hierarchical and relational data structures are employed. Full screen data entry can be accomplished on-line or off-line using microcomputers. Quality assurance programs, cataloging and documentation procedures, and extensive analysis and report writing procedures are incorporated in the BDMS. Also, data are routinely protected using both tape and mass storage devices.

Topics addressed include current directions in ecological research and the mandate for effective data management, design and development phases of a data management system, and BDMS implementation using SAS software.

DATA MANAGEMENT IN ECOLOGICAL RESEARCH

Efficient systems for data management are becoming necessary as research directions in ecology change. Historically, most ecological studies have been short-term, limited in scope, and required one or few investigators. Characteristics of the investigations were determined by funding level and the time frames established by the granting agencies. Funding periods were often limited to two or three years and funding levels were usually insufficient to support more than a single or few investigators.

Through the Long-Term Ecological Research Program (established by the National Science Foundation), funding is now available to support ecosystem level projects which focus on long-term processes. These projects require the talents of many investigators from various disciplines (for further discussion see Callahan, 1984). A single database generated at a site supported by the Long-Term Ecological Research Program (LTER) may exhibit the following characteristics:

1. Variety. Ecosystem level research requires understanding environmental conditions (water quality, meteorological factors, etc.), natural system interrelationships (i. e. how the marsh and bordering forest interact), population parameters (i. e. species growth, reproduction, and mortality), and community dynamics (predator-prey interactions, etc.).

2. Large size. Open-ended collections can result in extremely large databases. Also, many additional data sets may be generated to address specific hypotheses.

3. Scale. Numerous spatial and temporal aspects are generally incorporated.

Complex ecological database analyses require a wide array of statistical and graphical procedures. Also, techniques for data element integration across data sets and between temporal and spatial scales are necessary.

As these long-term data sets grow, it can be assumed that their value increases. Understanding long-term trends may only be possible through time series analysis of data sets comprising many years. A long-term data set represents a considerable investment in time, effort, and money. It is apparent that 'personnel, equipment, and data are the three most important resources that a research institution possesses.'

DESIGN and DEVELOPMENT PHASES

Once data is recognized as being an important institutional resource, a decision must be made how to best manage the data for current utilization and as a resource for future investigations. BDMS development was divided into five phases:

I. Existing system evaluation
II. System criteria selection
III. System component selection
IV. Integration of components into an overall scheme
V. Implementation

BDMS design and development was subject to institutional constraints; primarily, budget and staffing.

I. EXISTING SYSTEM EVALUATION

Numerous approaches have been taken in data management system development. One emphasizes an all-encompassing central database using a database management software package. Many packages require initial declaration of variables and hierarchical/relational structures. Often, it is difficult to add new variables or change data structures. Systems developed using this approach may require maximal planning effort and programming. In addition, it may be difficult to access specific statistics, graphics, and modelling packages.

Database structures often change as the ecological research questions change. Flexibility, therefore, is a prime consideration during system development and implementation. Consequently, such systems should allow addition of new data types and variables. Since ecological research is often multi-disciplinary in nature (i. e. zooplankton catch data as related to phytoplankton abundance and nutrient availability), mechanisms which provide the ability to utilize multiple data sets and types should be inherent to the system.

In developing the BDMS we opted not to tie ourselves into a particular database management package. Instead, we chose to define required system criteria and then software and hardware which would satisfy these requirements.

II. SYSTEM CRITERIA SELECTION

After examining the capabilities of many existing systems and software packages we chose the following criteria:

1. Modular structure where new data sets can be easily added.
2. Ability to write quality assurance programs and provide suitable documentation.
3. Wide ranging statistical and graphical capabilities.
4. Ability to integrate data from different scales.
5. The final system must be 'Friendly' to both the user and the data manager requiring minimal maintenance and staffing.

In addition, software package(s) designed or chosen for the system must have the folowing capabilities:

1. Full screen data entry
2. Quality assurance
3. File manipulation and exploratory analyses
4. Statistics and graphics
5. Ecological modelling
6. Archival (mass storage, tape)
7. Documentation
8. Communication

III. SYSTEM COMPONENT SELECTION

In developing the BDMS, we had access to various hardware including mainframes, minis, and micros through the University of South Carolina's Computer Services Division. It was then necessary to specify software. SAS was chosen as the primary software package since it satisfied most desired system criteria. Other packages are used for communication and occasionally for ecological modelling.

We found SAS to be very flexible. It was easy to add new variables and data sets and make changes. SAS procedures and routines could be used to generate formatted output, merge files, and take advantage of relational and hierarchical data structures. In addition, SAS is a high level language (high programmer productivity), programs are easily documented, and numerous statistical and graphical analyses are possible.

IV. INTEGRATION OF COMPONENTS

The generalized data management scheme is outlined in Figure 1. Long-term data sets collected at our site result from group research efforts. Each research project is treated as an individual database. The database integrates the project's various data sets. Collectively, the databases are integrated into the BDMS. A Database Catalog provides detailed documentation about database contents as well as location of specimen jars, maps, etc. A project's database contains local level management modules which are networked into a central archival facility (Figure 2). To increase localized data management efficiency, the LTER database is broken down into two subsets. Generally, the data sets within each subset follow the same sampling regime (sampling frequency, location, etc.). One data manager who received primary training in biological sciences manages the biological data sets. The other data manager who received training in physical and chemical sciences manages the physical and chemical data sets. Requests for secondary usage are routed to the appropriate data manager. This scheme's value is most noticeable to the secondary user who deals with a data manager familiar with the data and the scientific background leading to that particular investigation.

Administering the BDMS requires careful planning and continued cooperation of the investigators and data managers. Initially, the researchers and data manager meet and discuss data collection within the framework of the project objectives. The researchers are responsible for addressing potential data set relationships. After an initial planning and design phase, the data manager develops efficient programs for data entry, quality assurance, manipulation, and retrieval which follow the scheme outlined in Figure 1.

Close contact with top administration is important in assuring that data management strategies evolve in parallel fashion with the Institute's long-term research goals.

V. IMPLEMENTATION

Data entry is accomplished with full screen entry programs which allow quick and easy conversion from raw data sheets to computerized files. Each program screen corresponds to a segment of the raw data sheet. Data entry programs are designed by the data manager in conjunction with the researcher and associated technician. Entry programs and raw data sheets are designed simultaneously to provide an efficient scheme. Data are entered directly by the investigator or technician. The data are then written out to a virtual system file which is accessed and archived by the data manager.

The full screen entry programs incorporate several quality assurance checks which alert the technician to data entry errors (i. e. entering alphanumeric data instead of numeric data or values which fall outside of established limits). A "comments" section is incorporated into most data entry programs so any problems or comments about a particular collection can be maintained. A report-generation program (written in SAS) is used to produce a hard copy for error checking against raw data sheets. In addition, several programs produce statistical and graphical data representations so outliers (potential entry errors) can be spotted.

Generally, raw data as recorded on raw data sheets (once certified to be error-free) are archived prior to further manipulation or summarization. Data manipulation and reduction can then be performed and stored as a subset to the original. This allows for future acquisition of the raw data as well as reduced data sets. Raw data from a single observation is stored in flat file format. One observation can be stored on one or more rows. None of the rows are redundant and each column within a row can be assigned a distinct variable name. Observations are stored sequentially by date and sampling time or according to a pre-defined format.

Data sets are generally grouped according to sample types, dates, and sampling times utilizing a partitioned data set structure. Other pertinent information which allows easy utilization by the manager and researchers are coded.

Multiple backups for raw data are provided including tape, mass storage, and hardcopy. Data routinely run through a conversion program for analysis are maintained on tapes and

mass storage in the processed form. If necessary, it is possible to convert processed data to its original form. Further data summaries and representations are usually maintained in at least two locations.

Analysis programs are written in SAS. In addition to report generation programs and data tabulation, SAS is used for graphical and statistical analyses. Also, analyses are used to monitor sampling program effectiveness. Since extreme temporal variability is inherent to ecological data sets, it is important to delineate periods when sampling must be intensified to document specific phenomena. Also, it may be possible to decrease sampling effort without hindering our understanding of certain ecological processes. Four SAS Procedures are used routinely to monitor sampling efficacy (MEANS, PLOT, GLM, and NESTED; SAS Institute Inc., 1982 a,b).

CONCLUSION

A data management system can make a valuable contribution to an ecological research program by providing:

1. Rapid analytical feedback to researchers
2. Integration of multiple data components with different temporal and spatial scales for analysis
3. Data protection for future investigations
4. Adequate documentation to enhance secondary usage.

Specific capabilities required for an efficient data management system include: (1) the need for personnel to derive a data management scheme and supervise operation of the database; (2) an integrated set of software tools for entry, quality assurance, analysis, and storage and retrieval; (3) hardware facilities which will support the system as the databases develop and requirements change.

Reliance on SAS software for development has resulted in increased programmer productivity, significant decreases in necessary program documentation, routine sampling program assessment, and rapid turnaround of analyses to investigators.

REFERENCES

Callahan, James T. 1984. "Long-Term Ecological Research." BioScience 34(6): 363-367.

SAS Institute Inc. 1982a. SAS User's Guide: Basics. 1982 Edition. Cary, N.C.: SAS Institute, Inc. 923 pp.

SAS Institute Inc. 1982b. SAS User's Guide: Statistics. 1982 Edition. Cary, N.C.: SAS Institute, Inc. 584 pp.

NOTES

1 SAS is a registered trademark of SAS Institute, Inc., Cary, NC, USA.

2 Design and development of the BDMS was supported by National Science Foundation Grant BSR 8012165.

3 Technical contribution No. 577 of Belle W. Baruch Institute for Marine Biology and Coastal Research.

4 For additional information, contact:

William K. Michener
Baruch Institute for Marine Biology
and Coastal Research (USC)
P. O. Box 1630
Georgetown, SC 29442
(803) 527-2067

GENERAL DATA MANAGEMENT SCHEME

EMPLOYED BY THE BARUCH MARINE LAB

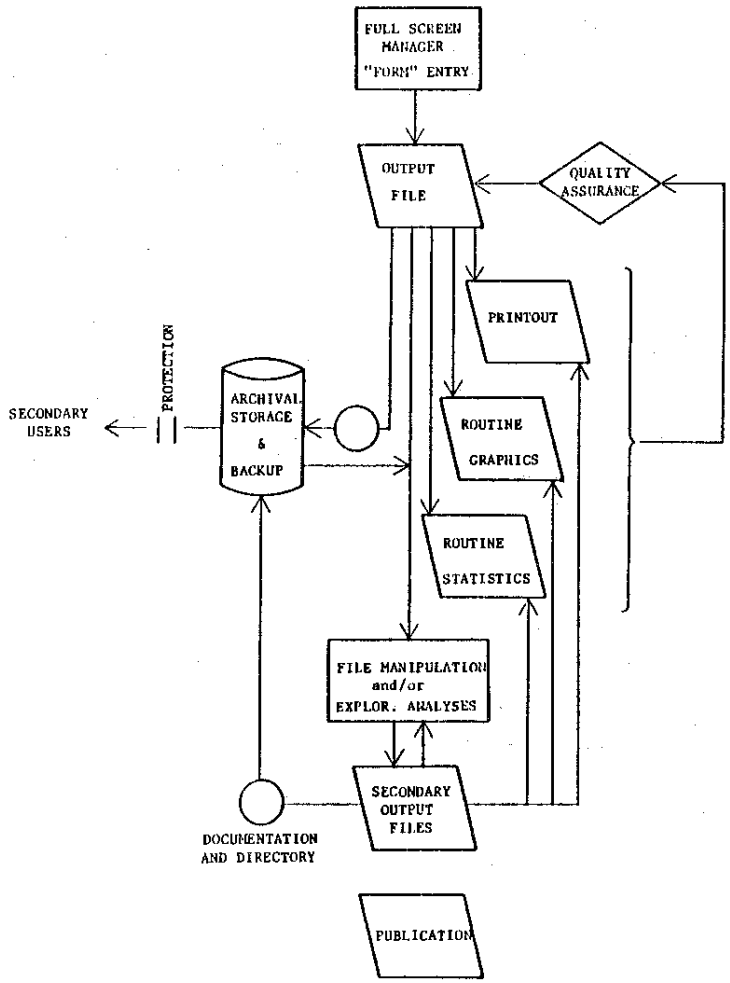

FIGURE 1

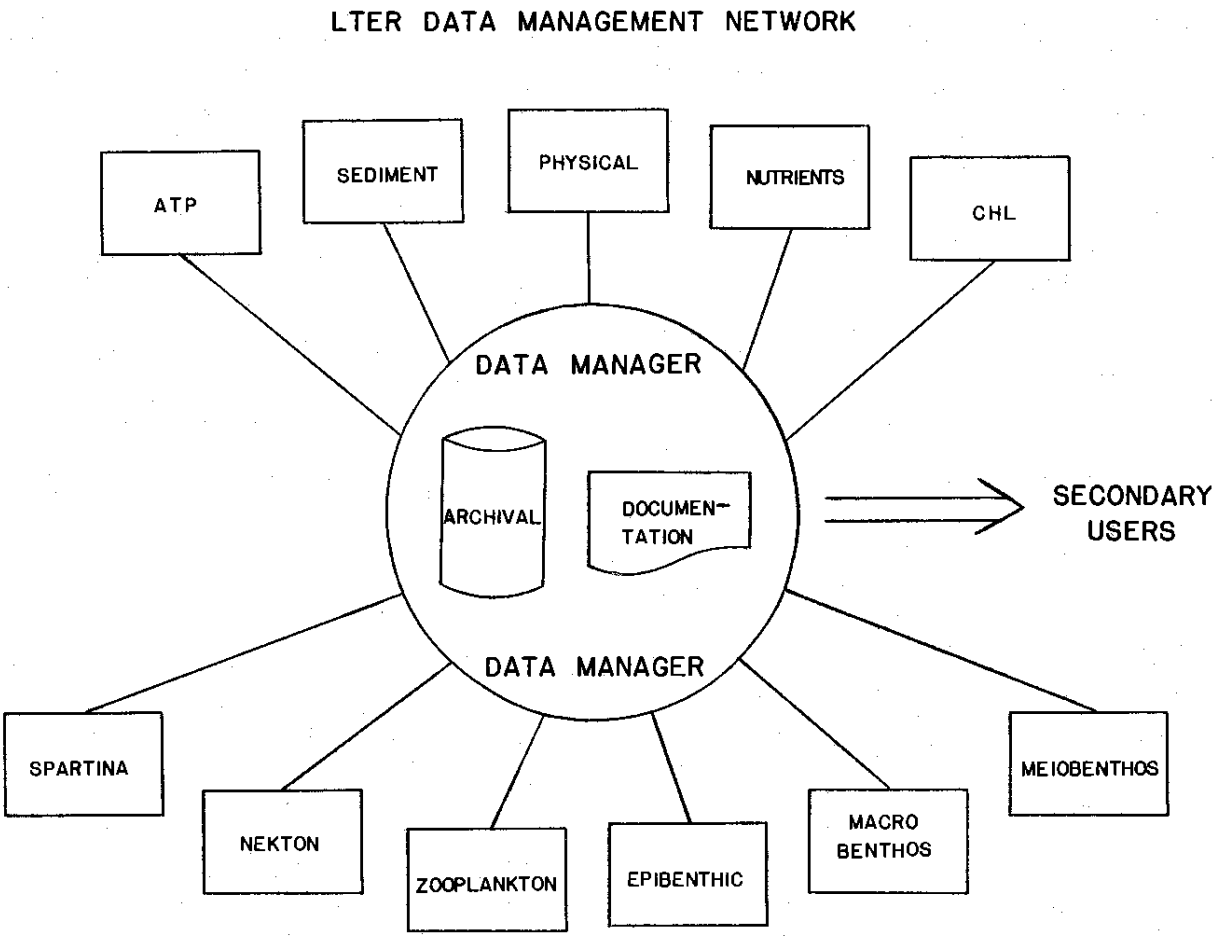
LTER DATA MANAGEMENT NETWORK
ATP
SEDIMENT
PHYSICAL
NUTRIENTS
CHL
DATA MANAGER
ARCHIVAL
DOCUMEN-
TATION
SECONDARY
USERS
DATA MANAGER
SPARTINA
NEKTON
ZOOPLANKTON
EPIBENTHIC
MACRO
BENTHOS
MEIOBENTHOS

FIGURE 2