DATA MANAGEMENT FOR ENVIRONMENTAL RESEARCH

Rodney H. Strand

Environmental Sciences Division
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37830

1Research supported by the Eastern Deciduous Forest Biome, US-IBP, funded by the National Science Foundation under Interagency Agreement AG-199, BMS69-01147 A09 with the Energy Research and Development Administration, Oak Ridge National Laboratory.

2Contribution No. ____ from the Eastern Deciduous Forest Biome, US-IBP. Publication No. ____, Environmental Sciences Division.

3Operated by Union Carbide Corporation for the U. S. Energy Research and Development Administration.
INTRODUCTION

The underlying principle of environmental data management is that data are a resource. Data resources play an essential role in facilitating and integrating environmental research programs. "Data resources" is a phrase used to describe data which, when properly managed, constitute a valuable and reusable resource for environmental research. Data resource management, or more briefly, data management, includes the planning for or the formulation of data needs, the standardization of data collections, the documentation of data sets, the efficient management of data files, and the coordination of data processing needs and activities. An active data management group consists of personnel skilled in data management methods and experienced in data management activities. They develop data resources and capabilities and also promote communication among researchers through discussion of common problems or goals and serve to integrate research results by publicizing available information.

This paper summarizes the data management requirements for environmental research and the techniques and automated procedures which are currently used by the Environmental Sciences Division at Oak Ridge National Laboratory. In this paper I deviate from the normal emphasis on SAS and the little known techniques we have used. Instead, I direct my discussion to ideas and concepts which I have used to develop data management within integrated research. This involves handling large amounts and varieties of data, and full consideration of the implementation of a total effort encompassing all aspects of data management.
Data Management Requirements

The basic assumption made in environmental data management is that accurate and reliable data are available. Without these, data management is an exercise which produces misinformation. However, with careful data collection and appropriate data management procedures, reliable data sets are formed which can be used for problem solving and decision making.

Recent recognition of data management requirements provides a means to address a range of hypotheses that has long been ignored because large-scale data compilation for ecosystem analysis was too formidable. These requirements form the basis for advances in data management in ecosystem analysis.

The basic needs include adequate computer hardware, applicable computer software, and personnel with data management skills and research experience in environmental science. Each of these needs can be examined in order to emphasize the components of environmental data management.

Computer Hardware

Specific computer hardware items are, in most cases, predetermined by the computer system available. Generally, the required hardware items are (a) late generation computers (IBM 370/168, 360/195) with (b) large core storage (512-1024K bytes), (c) online storage space in vast quantities (e.g., 358-4728 bytes), (d) high-speed graphical or tabular display hardware (CRT's, electrostatic or laser printer/plotters), (e) interactive processing with immediate response, and (f) batch processing
with acceptable response times for queries as well as data bank additions, updating, and maintenance. Included with these hardware items are peripherals such as channel processors, Remote Job Entry (RJE) interfaces, minicomputers for automated data logging, collection, and manipulation, and computer systems-support software. Floor space and maintenance and operations personnel are also needed. All of these may require millions of dollars and extensive administrative support.

Computer Software

Computer software is perhaps the largest combined dollar, time, and manpower consideration in environmental data management. What data do I want to manage? How extensive is the management required? These are the initial questions which must be answered. Other questions too numerous to mention cloud the decision process. The computer hardware can and usually does constrain the software choices. A survey of available software is the initial step in developing or implementing a system. With software development a number of different requirements are imposed on the software and its capabilities in order to adequately accomplish the data management objectives. These requirements are summarized in Table 1.

Personnel Requirements

Personnel requirements for data management are determined by the extent of the data management activities. The information in Table 2 notes the kinds of activities that may be included in the total data management approach. Skills in more than one of the activities are required of all data management personnel. Hopefully, familiarity
Table 1. Some software requirements for data management

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ease of use</td>
<td>Allows more user identification with system and increases system use</td>
</tr>
<tr>
<td>Portability</td>
<td>Gives flexibility when used at other installations and decreases duplication</td>
</tr>
<tr>
<td>Access control</td>
<td>Marshals data, some of which may not be available to every user, and maintains integrity of data sets</td>
</tr>
<tr>
<td>Destruct safeguards</td>
<td>Maintains integrity of data bank files and increases system availability by decreasing down times</td>
</tr>
<tr>
<td>Extendability</td>
<td>Provides for using or creating other programs for data analysis and system improvements</td>
</tr>
<tr>
<td>Documentation</td>
<td>Provides understandable and complete descriptions which are essential to system use and maintenance</td>
</tr>
</tbody>
</table>
Table 2. Data management activities for environmental research

<table>
<thead>
<tr>
<th>Activity</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>File creation and maintenance</td>
<td>Subsetting, merging, sorting, concatenation—batch and interactive</td>
</tr>
<tr>
<td>Data transformation and manipulation</td>
<td>Standardization of units, quality control</td>
</tr>
<tr>
<td>Data display</td>
<td>x/y plotting, outlier tests, pattern analysis</td>
</tr>
<tr>
<td>Statistical analysis</td>
<td>Univariate, multivariate, nonparametric, and modeling</td>
</tr>
<tr>
<td>Data set documentation</td>
<td>Variable lists, format, experimental design, authors, title, keywords, abstract</td>
</tr>
<tr>
<td>Computer usage instruction</td>
<td>Program executions, program options, system changes and capabilities, file inputs and outputs</td>
</tr>
<tr>
<td>Directory of data sets</td>
<td>All authors, titles, tape numbers, data set use counts, availability of data</td>
</tr>
<tr>
<td>Dictionary of data sets</td>
<td>Permuted title keyword lists, authors, and cross referenced data sets</td>
</tr>
<tr>
<td>Publish availability of data sets</td>
<td>Quarterly journal of available data</td>
</tr>
<tr>
<td>Computer hardware purchase and development</td>
<td>RJE facilities, time-sharing options, input-output computers</td>
</tr>
</tbody>
</table>
with all of the activities is common to all personnel. Bear in mind that familiarity with environmental research and applications of research skills to these activities enhances the data management operations performed.

Once the computer hardware, software and personnel requirements have been mapped out, a systematic approach to data management must be developed. A systematic scheme for data collection, formatting, collation, storage, documentation, and analyses is the most efficient and realistic means by which great volumes and varieties of data are synthesized. A data management scheme is represented as in Fig. 1. The initial step is the definition of objectives and the types of hypotheses to be tested. This step is the statistical planning or design phase. Also defined in this step are the number and kind of environmental or biological variables, frequency and location of sampling, degree of precision required, and a description of the analytical tools which will be used. These all serve valuable functions of defining the research objectives and procedures explicitly and highlighting potential problems in data collection, data set formulation, computerization, and analysis. Additional variables which may be required analytically and/or data sets which may be merged with these data to provide more comprehensive analysis and evaluation may also be identified.

After the content of the data set is defined and a computer record layout is planned, the data are collected and computerized. Manual data collection requires some form of encoding of the values into a computer-compatible record. If automated, the data collection procedure generates a computer-compatible record on a medium such as punched paper tape or
Fig. 1. Numeric data management scheme.
magnetic tape. Telemetry systems for transmitting data directly into a computer, where it is stored on magnetic tape or disk, are also used for collecting ecological data. However, these systems are expensive to purchase and maintain and they require environmentally controlled enclosures for acceptable performance. But when operable, they provide an excellent means of reducing human errors in encoding and decrease time lags between collation and analysis.

Quality control for determining inaccuracies in the data is one of the most critical, time-consuming tasks in the data management scheme. In analysis and evaluation, ecological data are assumed to be "mistake-free". As scientists who have thoroughly looked at and evaluated data or who have applied extant data in synthesis and assessment efforts well know, errors in a data set are often found even after extensive editing and display of the data set have been performed.

Following quality control, a data set is documented and prepared for file storage. Documentation requires the description of a number of data set items such as: (1) data record layout of variables and units as well as any missing data; (2) experimental data collection methods; (3) an experimental design and sampling scheme; (4) detection limits and error-bounds of the data values; (5) a background description, e.g., geography, locale, frequency of sampling; (6) the name of the investigator most acquainted with the data; (7) the subject areas of the data (for keyterm search); and (8) documents which further describe or evaluate these data. This documentation is filed in the computer as a textual file describing and cross-referencing each unique data set in the data bank files. These textual files are
then available for on-line searches for specific subject matter and for display of the description of a particular data set.

Temporary data files are formed following retrieval from the permanent storage area of the numeric data bank. Data in these files can be then input into modeling programs, data analysis routines, and/or data display devices. Following analysis, the data are used by the researcher to evaluate his scientific problems or hypotheses. The decisions based on the evaluations of the data through the use of these data management tasks feed back into the research planning effort and provide insights into the questions and problems posed for later research.

Within this scheme, the contributions of the data management personnel are critical to the successful completion of the research. Experimental design analysts, data analysts, computer programmers, data preparation specialists, modelers, and data managers are some of the people who contribute their expertise and experience to performing the data management tasks for each research problem. Collaboration between the researcher and data management personnel is strongly emphasized as a means of efficiently solving the research problems and reporting the results.

Data Handling Systems

Automation of the various tasks described above increases the efficiency of the data management operation, which for environmental science at Oak Ridge National Laboratory has been developed as three systems. These systems serve to store, retrieve, and catalog numeric or textual data sets. Data analysis and manipulation are then performed
on the data sets with available computer programs or statistical analysis packages. The constraints of large textual file searches and the unavailability of some of the new SAS capabilities have led to the 3-system approach (Fig. 2).

Each of the following systems satisfies a specific data management need and is a widely used system for managing numeric and/or textual data:

**ORCHIS**

The Oak Ridge Computerized Hierarchial Information System [ORCHIS] (Brooks 1974) is used for storing and searching data set documentations in an interactive mode and producing textual reports, e.g., bibliographies, abstracts. Keyterms are formed for the subject matter of interest and used in searching the textual descriptions for data sets relevant to that subject matter. Each data set has a unique data set name cross-referenced in the data set descriptions. When the relevant data set(s) is found, the data set name is used to retrieve the data set from the numeric file.

**NUTIS**

The storage and retrieval of data sets and cataloging of the data set names and certain descriptive text are maintained by another system. The Numerical and Textual Information System [NUTIS] (Strand and Taylor 1974) maintains a catalog of data sets, the locations of the data sets in the numeric files for retrieval purposes and a textual directory of the contents of the numeric data files. This system eliminates manual logging of data sets and corresponding tape numbers,
Fig. 2: Systems for Environmental Data Management.
organizes the numeric data files in structure and content, and provides computer-generated tables describing each data set. These tables are used by individuals who prefer to scan the numeric data files manually for data sets of interest to them.

SAS

The Statistical Analysis System [SAS] (Barr and Goodnight 1972) is the most frequently used manipulation and analysis tool. With its sound statistical analyses, and its capabilities for subsetting, merging, sorting, and hierarchically processing data files, the user is provided with a general data management tool for production processing of data. This system is doubly desirable because of the short time required for learning how to use the system and the small amount of time required for maintaining a working level of competence among infrequent as well as frequent users of the computer.

SAS Applications. The initial uses of SAS in the Environmental Sciences Division were for statistical analysis. Since then, we have expanded the uses of SAS with the aid of its programming language and data set manipulation routines. In turn, this has aided in the optimization of our data management activities. One example of our use of SAS is in Appendix A.

The data set "NOO11" is one of many sets which are being edited, standardized, and archived within a project initiated under the International Biological Program at Oak Ridge. Each data set within this project was archived following massage with SAS to edit the data for inconsistencies. In addition to editing, variables were standardized
(e.g., year, month, day) to Julian date and metric units of measurement, where appropriate. Other checks (not shown here) were programmed into the data editing procedures to consider impossible ranges of data values. Blanks were inserted for variables in cases where a non-blank value had been used to indicate a missing value.

When completely updated, each data set was output onto a magnetic tape or disk for input into the archiving routines. A procedure "DPTPCH" (see documentation, Appendix B) was written and incorporated into SAS. DPTPCH allowed us to place blanks in our output fields wherever a missing value occurred for a data set variable in SAS. This routine followed closely the PRTPCH in the Supplemental Procedure Library (Perkins, 1973). Numerous other applications of SAS are used in our data management effort, but for brevity, I include only the previous example.

Data Management Benefits

Some of the benefits of data management in environmental research are indicated below. The archived, standardized data serve as a valuable resource and constitute another output from research projects. The synthesis of extant data at later dates enhances summaries of current knowledge of particular subject areas. Field data can be integrated by applications of extant data from similar sites within different projects, increasing the exchange of information among projects and investigators. Data handling conducted in an organized and structured manner tends to be more efficient not only for the project investigators but also for the project's supportive agency.
Investigators can use the data handling tools and contribute to the data resources development by documenting and submitting their data for archiving.

I have presented some of the concepts and principles used in environmental data management at Oak Ridge National Laboratory. Some may not be new; others may be academic. I believe we can appreciate the need for a data management tool, such as the 1976 version of SAS, which would complement the on-line bibliographic needs and data set access control. I have emphasized an evolution of our systematic data management approach within the constraints of a national laboratory operated by a private corporation for a Federal agency. Just as SAS has evolved to the enormous system it is now, so have the data management concepts and principles changed to meet the dollar support, the needs, and the tools available. I would hope that programs involved in data management of any magnitude would be able to apply some of the principles and guidelines I have set forth. Data management is not an exact science, but lends itself to the scientific mode of question, alternatives, hypotheses, testing, and solutions.

In the management of environmental data we hope to become better versed in the new capabilities of SAS, and, in this manner, be able to do more and more of the data management activities with this single system. Eventually, with on-line searching capabilities and access control of data sets and documentation, we may be able to use this system exclusively for our data management effort.
APPENDIX A

// EXEC SAS
//SAS.FTU7DD DD UNIT=2400,DISP=(NEW,KEEP),VOL=SER=1,LABEL=(NL),
// DCP=(RECIM=FB,LRECL=80,PLKSIZE=4000)
//SAS.SYSIN DD *

COMMENT MACRO TO CONVERT MONTH-DAY-YEAR (MN DY YR) TO JULIAN DATE (JDATE);

MACRO DATES;
   JDATE = 0;
   LEAP = 0;
   IF MN = 72 OR YR = 68 THEN LEAP = 1;
   IF MN = 7 THEN JDATE = DY + 31;
   IF MN = 8 THEN JDATE = DY + 59 + LEAP;
   IF MN = 9 THEN JDATE = DY + 120 + LEAP;
   IF MN = 10 THEN JDATE = DY + 151 + LEAP;
   IF MN = 11 THEN JDATE = DY + 212 + LEAP;
   IF MN = 12 THEN JDATE = DY + 334 + LEAP;
   COMMENT CHECK FOR REASONABLE DATE AND YEAR;
   IF JDATE < 1 THEN JDATE = 0;
   IF JDATE > 366 THEN JDATE = 0;
   IF YR > 99 THEN YR = YR - 1900;
   JDATE = JDATE + YR*1000;
   COMMENT END OF JULIAN DATE MACRO;
%

COMMENT CHANGE WEATHER DATA TO METRIC UNITS BY SETTING BASE AND FACTOR

BASE - VALUE SUBTRACTED FROM OLD VALUE PRIOR TO MULTIPLICATION

FACT - VALUE TO MULTIPLY OLD VALUE;

MACRO CONVERT;
   C = MISS( (C - BASE) * FACT, C);
   D = MISS( (D - BASE) * FACT, D);
   E = MISS( (E - BASE) * FACT, E);
   F = MISS( (F - BASE) * FACT, F);
   G = MISS( (G - BASE) * FACT, G);
   H = MISS( (H - BASE) * FACT, H);
   I = MISS( (I - BASE) * FACT, I);
   J = MISS( (J - BASE) * FACT, J);
   K = MISS( (K - BASE) * FACT, K);
   L = MISS( (L - BASE) * FACT, L);
   M = MISS( (M - BASE) * FACT, M);
   COMMENT END OF MACRO CONVERT;
%

DATA NN111;

INPUT
   A $ 4-7 YR 5-9 MN 10-11 DY 12 C 13-18 D 19-24 E 25-30
   F 31-36 G 37-42 H 43-48 I 49-54 J 55-60 K 61-66 L 67-72 M 73-78;
   COMMENT CONVERT MPH TO KM PH;
   FACT = 1.6 / 0.6213;
   BASE = 0.0;
CONVERT

\( \text{DY} = (B - 1) \times 11 + 1 \)

DATES CARDS:

\[
\begin{array}{ccccccccccccccc}
496170 & 11 & 3.0 & 2.5 & 8.6 & 6.5 & 8.0 & 5.5 & 11.4 & 16.4 & 9.6 & 5.3 & 6.8 \\
496170 & 12 & 8.2 & 3.3 & 9.6 & 9.1 & 9.9 & 12.2 & 8.9 & 6.3 & 13.4 & 8.6 & 12.4 \\
496170 & 13 & 6.8 & 11.2 & 5.9 & 6.8 & 9.8 & 8.8 & 13.7 & 9.4 & 10.4 \\
496170 & 21 & 13.7 & 17.1 & 11.8 & 17.5 & 9.2 & 5.2 & 6.2 & 12.2 & 11.1 & 9.5 & 8.6 \\
496170 & 22 & 7.9 & 4.1 & 3.9 & 6.2 & 9.9 & 8.8 & 14.1 & 12.7 & 12.5 & 9.9 & 9.1 \\
496170 & 23 & 8.9 & 12.2 & 15.4 & 15.1 & 9.6 & 5.0 \\
496172121 & 6.9 & 9.1 & 12.7 & 12.1 & 8.6 & 15.7 & 4.8 & 7.3 & 8.1 & 13.1 & 7.5 \\
496172122 & 7.5 & 8.6 & 9.9 & 10.2 & 15.8 & 12.8 & 9.5 & 7.3 & 6.0 & 8.3 & 11.8 \\
496172123 & 9.2 & 8.8 & 10.8 & 13.1 & 10.9 & 8.6 & 14.4 & 11.5 & 18.3 \\
\end{array}
\]

139 OBSERVATIONS IN DATA SET N0011

16 VARIABLES

PROC DPTPCH PUNCH OUT_UNIT=7;
VAR A JDATE B C D E F G H I J K L M Z;
PARM CARDS
(A4,'WIND',F6.0,T14,F2.0,T15,F20.0)
/**
/**

-------------------------------

LISTING OF FILE /

\[
\begin{array}{ccccccccccccccc}
4961WIND700011 & 4.8 & 4.8 & 14.2 & 10.5 & 9.7 & 8.9 & 18.3 & 26.4 & 15.5 & 8.5 & 10.9 \\
4961WIND700012 & 13.2 & 5.3 & 15.5 & 14.6 & 15.9 & 19.6 & 14.3 & 10.1 & 21.6 & 13.3 & 20.0 \\
4961WIND700023 & 10.9 & 18.0 & 14.3 & 11.1 & 15.8 & 14.2 & 22.1 & 15.1 & 16.7 \\
4961WIND700321 & 22.1 & 27.5 & 19.0 & 28.2 & 14.8 & 8.4 & 10.0 & 19.6 & 17.9 & 15.3 & 13.8 \\
4961WIND700432 & 12.7 & 7.4 & 6.3 & 10.0 & 15.5 & 14.2 & 22.7 & 20.4 & 20.1 & 15.9 & 14.6 \\
4961WIND700543 & 14.3 & 19.6 & 24.8 & 24.3 & 14.5 & 8.0 \\
4961WIND723361 & 11.1 & 14.6 & 20.4 & 19.5 & 13.8 & 22.1 & 7.7 & 11.7 & 13.0 & 21.1 & 12.1 \\
4961WIND723472 & 12.1 & 13.8 & 1.4 & 16.4 & 25.4 & 20.6 & 15.3 & 11.7 & 9.7 & 13.4 & 19.0 \\
4961WIND723583 & 14.8 & 14.2 & 17.4 & 21.1 & 17.5 & 13.8 & 23.2 & 18.5 & 29.5 \\
\end{array}
\]
Appendix B

A Modified SAS Procedure for Outputting Formatted Records: DPTPCH

R. J. Olson, November 1975

The DPTPCH procedure is a modification of the PRTPCH procedure as described in the Guide to Library by C. G. Perkins, August, 1971. It allows missing numeric values to be punched or printed as blank fields and allows specifying the output unit which can be defined by a JCL DD card. Fields which may contain missing values are indicated by adding 200 to the width indicator of the FORTRAN format statement following the PARMCARD, e.g. F208.3. Both PRTPCH and DPTPCH do not permit the use of integer ("1") specifications in the format statement as all SAS variables are stored as double precision floating point variables. Decimal points can be deleted by overlaying fields using the "T" option in the format. An example of printing an integer is given below as part of the sample SAS statements. DPTPCH was written by D. G. Taylor and is available with the standard SAS library on the X-10 IBM-360 system.

PROCEDURE DPTPCH [PUNCH] [DATA=data_set_name] [OUT_UNIT=nn];

PUNCH - causes the output to go to FT02F001 (card punch) or to the file specified by OUT UNIT; default is the printer if PUNCH is not specified.

DATA= - name of data set to be processed; default is data set most recently referenced.

OUT_UNIT= - specifies unit number to relate output to a file defined by a JCL DD card; options include tape, disc, print or punch file.

Example:

JCL to define unit 7 as a mag tape.

//SAS,FT07F001 DD UNIT=2400,DISP=(NEW,KEEP),VOL=SER=1,LABEL=(,NL), // DCB=(RECFM=FB,LRECL=80,BLKSIZE=4000)

SAS instructions to write data set to above tape assuming A and B are character variables, I and J are integers and X, Y and Z are floating point variables.

PROC DPTPCH PUNCH OUT_UNIT=7;
VAR A B I J X Y Z;
PARMCARDS;
(A3,1x,A1,F205.0,T10,",F210.0,T20,F205.1,2F210.3)