depends on the latter for life-supporting air, water, and food. Under natural selection in nature, parasites and hosts tend to coevolve for co-existence; otherwise, if the parasite takes too much from its host, both will die if the parasite has only one host, as is the case with humans and the earth.

John Cairns (1997) is hopeful that, somehow, natural and techno-ecosystems will coevolve for mutualistic coexistence, preventing such a doomsday. But I'm afraid this won't happen until we overshoot carrying capacities and are forced to become proactive and get more effort and money to flow down the reward feedback loop of service, as shown in Fig. 1. This means "reconstructing" economics to include the life-supporting goods and services (natural capital), as suggested by Kenneth Boulding some 40 years ago, and as widely discussed now in the year 2001 by both economists and ecologists. (See the symposia in *Ecosystems* [Volume 3, Number 1] and *BioScience* [Volume 50, Number 4], and the book by Hawkins and the Lovins [1999].)

Literature cited

Boulding, K. 1962. The reconstruction of economics. Science Editions, New York, New York, USA.

Cairns, J. 1997. Global coevolution of natural systems and human society. Revista Sociedad Mexicana de Historia Natural **47**:217.

Hawkins, P., A. Lovins, and L. H. Lovins. 1999. Natural capitalism: creating the next industrial revolution. Little, Brown, Boston, Massachusetts, USA.

Neveh, Z. 1982. Landscape ecology as an emerging branch of human ecosystem science. Advances in Ecological Research **12**:189–237.

Wackernal, M., and W. Rees. 1996. Our ecological footprint: reducing human impact on the Earth. New Society Publishers, Cabriola Island, British Columbia, Canada.

Eugene P. Odum
Institute of Ecology
University of Georgia
Athens, GA

Best Practices for Preparing Ecological Data Sets to Share and Archive

Introduction

Historically, ecological data have been collected to support studies at small temporal and spatial scales by single or small numbers of investigators. These data have been published, but typically have not been made available for use by others. Over the past decade, ecologists have recognized that, collectively, these data are extremely useful and are needed for modeling, synthesis, and assessment of such interdisciplinary issues as global change, biodiversity, and sustainability (NRC 1991, Michener et al. 1997). The report of the ESA Ad Hoc Committee on the Future of Long-term Ecological Data (FLED) (Gross et al. 1996) <http://esa.sdsc.edu/FLED/FLED.html> describes the importance of long-term data sets and the need to develop mechanisms to promote their preservation, maintenance, and use. Recent advances in the Internet and electronic storage of data have improved our ability to use these data for broader scale ecological studies. As a result of the FLED report and advances in the Internet, ESA has established *Ecological Archives* (Peet 1998) <http://esa.sdsc.edu/Archive/> to publish data papers, digital appendices, and supplements for articles published in ESA journals. However, while the need for sharing data has become well recognized, training for ecologists often does not include how to produce and document data sets to ensure that other investigators can find, understand, and use the data.

The purpose of this paper is to provide guidance on data management practices that investigators should perform during the course of data collection to improve the usability of their data sets. This guidance is tailored for those who perform ecological and other ground-based measurements, although many of the practices may be useful for other data collection activities. These practices could be performed at any time during preparation of the data set, but we suggest that researchers consider them before measurements are taken.

Seven practices that researchers could implement to make their data sets ready to share with ecologists and global-change researchers are the following:

1. Assign descriptive file names
2. Use consistent and stable file formats
3. Define the parameters
4. Use consistent data organization
5. Perform basic quality assurance
6. Assign descriptive data set titles
7. Provide documentation

Additional information on preparing data sets may be found in Christensen et al. (2000), Kanciruk et al. (1986), Michener and Brunt (2000), ORNL DAAC (2000), Porter (1997), and USGS (2000).

1. Assign descriptive file names

File names should reflect the contents of the file and include enough information to uniquely identify the data file. They should contain information such as project acronym, study title, location, investigator, year(s) of study, and file type. The file name should be provided in the documentation (Section 7) and in the first line of the header rows in the file itself.

Clear, descriptive, and unique file names may be important later when your data file is combined in a directory or FTP site with your own data files or with the data files of other investigators. Avoid using file names such as mydata.dat or 1998.dat.

An example of a great file name is:

narsto_texas_PM2.5_study_1997-1998.csv

where NARSTO is the name of the project, Texas is the location, "PM2.5 Study" is the project name, 1997–1998 is the date of the study, and .csv is the file type (format).

When choosing a file name, check for any database management limitations on file name length and use of special characters. Also, in general, lowercase names are less software- and platform-dependent.

You may want to use similar logic when designing directory structures and names. Also, the data set title (see Section 6) should be similar to the data file name(s).

2. Use consistent and stable file formats

Using ASCII file formats is the best way to ensure that field data are readable into the future.

Use the same format throughout the file; don't have a different number of columns or rearrange the columns within the file. At the top of the file, include several header rows. The first row should contain the file name, data set title, author, date, and companion file names. Other header rows (column headings) should describe the content of each column, including one row for parameter names and one for parameter units.

Within the ASCII file, delimit the parameter fields using commas, pipes (|), tabs, or semicolons; these are listed in order of our preference. Avoid delimiters that also occur in the data fields. If this cannot be avoided, enclose the data fields that also contain a delimiter in single or double quotes. Don't include rows with summary statistics; it is best to put summary statistics, figures, and other comments in a separate file or in the documentation.

Some field researchers may generate raster data (image data or gridded GIS data). We don't offer any general recommendations about raster data, except that the format needs to be clearly documented. Binary file formats are used for most raster data, especially large-volume raster data. For small-volume raster data (coarse-resolution global data or fine-resolution data of a field site), ASCII format may be appropriate.

If you cannot use ASCII or binary files formats, another option is non-proprietary public domain data formats such as NET-CDF or HDF. Both formats have been used extensively to date and are reasonably well supported with open source versions of the software needed to read and write these formats.

Whatever file format you use, be sure to thoroughly document the format (see Section 7).

3. Define the parameters

In order for others to use your data, they must fully understand the parameters in the data set, including the parameter name, unit of measure, and format.

Parameters reported in the data set need to have names that describe the contents. The documentation should contain a full description of the parameter. Use commonly accepted parameter names, e.g., Temp for temperature, Precip for precipitation, Lat and Long for latitude and longitude. See *Literature cited* for additional examples. Be sure to use consistent capitalization (not temp, Temp, and TEMP in the same file), and use only letters and numerals in the parameter name. Because some software allows a limited number of characters, be sure the first eight characters are unique.

The units of reported parameters need to be explicitly stated in the data file and in the documentation. We recommend SI units, but recognize that each discipline has its own commonly used units of measure. The critical aspect here is that the units must be defined so that others understand what is reported.

Within each data set, choose a format for each parameter, explain the format in the documentation, and use that format throughout the file.

We recommend the following formats for common parameters:

Dates: Use yyyymmdd, e.g., January 2, 1997 is 19970102.

Time: Use 24-hour notation (13:30 hours instead of 1:30 p.m.). Report in both local time and Coordinated Universal Time (UTC). Include local time zone in a separate field. As appropriate, both the begin time and end time should be reported in both local and UTC time. Because UTC and local time may be on different days, we suggest that dates be given for each time reported.

Spatial coordinates: Spatial coordinates should be recorded in decimal degrees format to at least 4 (preferably 5 or 6) significant digits past the decimal point.

Provide latitude and longitude with south latitude and west longitude recorded as negative values, e.g., 80°30'00" W longitude is -80.500000.

Make sure all location information in a file uses the same coordinate system, including coordinate type, datum, and spheroid. Document all three of these characteristics, e.g., Lat/Long decimal degrees, NAD83 (North American Datum of 1983), WGRS80 (World Geographic Reference System of 1980). Mixing coordinate systems, e.g., NAD83 and NAD27 (North American Datum of 1927) will cause errors in any geographic analysis of the data.

Elevation: Provide elevation in meters. Include detailed information on the vertical datum used, e.g., North American Vertical Datum 1988 (NAVD 1988) or Australian Height Datum (AHD).

Missing values: Use a decimal point (.) or extreme value (-9999). Do not use character codes in a numeric field. Use the same notation for each missing value in the data set. Codes should not be parameter specific. Supply a flag or tag in a separate field to explain the reason for missing data.

4. Use consistent data organization

We recommend that you organize the data within a file in one of two ways. Whichever style you use, be sure to place each observation in a separate line (row). Most often, each row in a file represents a complete record and the columns represent all of the parameters that make up the record. This arrangement is similar to a spreadsheet or matrix (Table 1).

Table 1. Example of data organization.

Station	Date (YYYYMMDD)	Temp (°C)	Precip (mm)
HOGI	19961001	12	0
HOGI	19961002	14	3.3
HOGI	19961003	19	-9999

The final value of –9999 is a missing value code for this data set. If you use a coded value or abbreviation for a site or station (e.g., HOGI stands for Hog Island, Virginia), be sure to provide a definition, including spatial coordinates, in the documentation.

A second arrangement may be more efficient when most records do not have measurements for most parameters, i.e., a very sparse matrix of data, with many missing values. In this arrangement, one column is used to define the parameter and another column is used for the value of the parameter. Other columns may be used for site, date, treatment, and units of measure (see Table 2).

An important issue in data organization is the number of records in each file (file size). Many factors determine the optimal number of records in a file, and we don't have any hard and fast rules. In general, keep a set of similar measurements together (e.g., the same investigator, methods, and instruments) in one data set. Please do not break up your data into many small files, e.g., by month or by site if you are working with several months or sites. Instead, make month or site a parameter and have all the data in one large file. Researchers who later use your relatively large data file won't have to process many small files individually. There is an upper limit to file size, though. Large files (on the order of several tens of thousands of records, or several megabytes) do become unwieldy and may be too large for some applications. Such files must be broken into logical smaller files.

If you are collecting several different types of measurements at a site (e.g., leaf area index and above- and belowground biomass), put each type of measurement in a separate data set. For each data set, use similar data organization, parameter formats, and site names, so that users understand the interrelationships between data sets.

5. Perform basic quality assurance

In addition to scientific quality assurance (QA), we suggest that you perform basic data QA on the data file:

- Check file format by making sure the data are delimited/line up in the proper column.
- Check file organization and descriptors to ensure that there are no missing values for key parameters (such as sample identifier, station, time, date, geographic coordinates). Sort records by key data fields to highlight discrepancies.

Table 2. Example of layout with a column defining parameters, useful for records with missing data.

Station	Date	Parameter	Value	Unit
HOGI	19961001	Temp	12	°C
HOGI	19961002	Temp	14	°C
HOGI	19961001	Precip	0	mm
HOGI	19961002	Precip	3.3	mm

- Check the content of measured or derived values. Scan parameters for impossible values (e.g., pH of 74; negative values where negative values are impossible). Review printed copies of data file(s) and generate time series plots to detect anomalous values.
- Perform statistical summaries (frequency of parameter occurrence) and review results.
- If location is a parameter (latitude/longitude), use scatter plots or GIS software to map each location to see if there are any errors in coordinates.
- Verify data transfers (from field notebooks, data loggers, or instruments). For data transfers done by hand, consider double data entry (entering data twice, comparing the two data sets, and reconciling any differences). Where possible, compare summary statistics before and after transfers.

6. Assign descriptive data set titles

We recommend that data set titles be as descriptive as possible. When naming your data sets and associated documentation, please be aware that these data sets may be accessed many years in the future by people who will be unaware of details of the project.

Data set titles should contain the type of data and other information such as the date range, location, and instruments used. If your data set is part of a larger field project, you may want to add that name (e.g., LBA or SAFARI 2000). In addition, we recommend restricting the title length to 80 characters (spaces included) to be compatible with other global change data collections.

The data set title should be similar to the name(s) of data file(s) in the data set (see Section 1). Bad titles: "The Aerostar 100 Data Set;" "Respiration Data;" and "Amazonian Respiration Data." A great title: "LBA Respiration Data for Broadleaf Evergreen Trees in Rondonia, Brazil, 1999–2000."

7. Provide documentation

The documentation accompanying your data set should be written for a user 20 years into the future. What does that investigator need to know to

use your data? Write the document for a user who is unfamiliar with your project, methods, or observations.

To ensure that documentation can be read 20 years in the future, put it in a stable nonproprietary format. We recommend ASCII format for text. For figures, maps, equations, or pictures, use a nonproprietary document format such as html (hypertext markup language). Images, figures, and pictures may be included as individual gif (graphics interchange format) or jpg (Joint Photographic Experts Group) files. Stable proprietary formats such as rtf (rich text format) or pdf (portable document format) are a suitable last resort.

Documentation should be in a separate file, identified in the data file. The documentation file name should be similar to the data set file name.

The data set documentation should provide the following information:

- The data set name, which will be the documentation title (Section 6)
- The scientific reason why the data were collected
- What data were collected
- What instrument (including model and serial number) (e.g., rain gauge) and source (meteorological station) were used
- Who collected the data and who to contact with questions (include e-mail and Web address if appropriate)
- Who funded the investigation
- The name(s) of the data file(s) in the data set (see Section 1)
- How to cite the data set
- Where and with what spatial resolution the data were collected. If codes are used for location, be sure to define the codes in the documentation (e.g., HOGI in Section 4)
- When and how frequently the data were collected
- How each parameter was measured or produced (methods), its units of measure, the format used for the parameters in the data set, the precision and accuracy if known, and the relationship to other data in the data set if appropriate (see Section 3)
- Environmental conditions (e.g., cloud cover, atmospheric influences)
- The data processing that was performed, including screening
- Standards or calibrations used
- Software (including version number) used to prepare the data set
- Software (including version number) needed to read the data set
- Quality assurance and quality control that have been applied (see Section 5)
- Special codes, including those for missing values (see Section 3) or for stations (see Section 4)
- Date when the data set was last modified
- Summary statistics generated directly from the final file
- Example file record
- Pertinent field notes or companion files; file names should be similar to documentation and data file names
- Related or ancillary data sets
- Known problems that limit data use

Documentation can never be too complete.

Acknowledgments

Preparation of this document has benefitted from the thoughtful reviews of Larry Voorhees, Raymond McCord, Lisa Olsen, Betsy Horwedel, and Sig Christensen (ORNL), John Porter (University of Virginia), and Don Strebel (Versar Incorporated). This work was sponsored by the U.S. National Aeronautics and Space Administration, Earth Science Data and Information Systems Project. Oak Ridge National Laboratory is operated by UT-Battelle, LLC for the U.S. Department of Energy, under contract DE-AC05-00OR22725.

Literature cited

Christensen, S. W., T. A. Boden, L. A. Hook, and M.-D. Cheng. 2000. NARSTO data management handbook. ORNL/CDIAC 112/R2. Oak Ridge, Tennessee, USA. Available online: <http://cdiac.esd.ornl.gov/programs/NARSTO/narsto.html>

Gross, K. L., C. E. Pake, and FLED Committee members. 1995. Final report of the Ecological Society of America Committee on the Future of Long-term Ecological Data (FLED). Volume 1: Text of the Report. Ecological Society of America, Washington, D.C., USA.

Kanciruk, P., R. J. Olson, and R. A. McCord. 1986. Quality control in research databases: the U.S. Environmental Protection Agency National Surface Water Survey experience. Pages 193–207 *in* W. K. Michener, editor. Research data management in the ecological sciences. The Belle W. Baruch Library in Marine Science Number 16. Belle W. Baruch Institute, University of South Carolina, Columbia, South Carolina USA.

Michener, W. K., J. W. Brunt, J. Helly, T. B. Kirchner, and S. G. Stafford. 1997. Non-geospatial metadata for ecology. Ecological Applications **7**:330–342.

Michener, W. K., and J. W. Brunt, editors. 2000. Ecological data: design, management and processing. Methods in Ecology. Black-well Science, Oxford, UK.

National Research Council (NRC). 1991. Solving the global change puzzle: a U.S. strategy for managing data and information. Report by the Committee on Geophysical Data of the National Research Council Commission on Geosciences, Environment and Resources. National Academy Press, Washington, D.C., USA.

ORNL DAAC. 2000. Guidelines for producing ecological data sets for distribution and archive. Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA, *in review*.

Peet, R. K. 1998. ESA journals: evolution and revolution. ESA Bulletin **79**:177–181.

Porter, J. H. 1997. Data and information submission at the Virginia Coast LTER. Available online at: <http://www.vcrlter.virginia.edu/data/submission.html>

USGS (U.S. Geological Survey). 2000. Metadata in plain language. Available online at: <http://geology.usgs.gov/tools/metadata/tools/doc/ctc/>

Robert B. Cook, Richard J. Olson, Paul Kanciruk, and Leslie A. Hook
Environmental Sciences Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831-6038