Managing a Database Pond Research Data - the CRSP Experience[1]

KEVIN D. HOPKINS
JAMES E. LANNAN
JAMES R. BOWMAN
Collaborative Research Support Program
Pond Dynamics / Aquaculture[2]
Office of International Research and Development
Oregon State University
Corvallis, Oregon 97331
USA

Introduction

The Collaborative Research Support Program in Pond Dynamics/Aquaculture (CRSP) was established in 1982 to describe the physical, chemical and biological principles of pond culture systems. Seven U.S. universities cooperate with researchers at seven stations in six countries: Honduras, Indonesia, Panama, Philippines, Rwanda, and Thailand.[3]

Experimentation started in late 1983 and is scheduled to continue until 1990. Two five-month experiments are conducted each year during which 96 variables are monitored. Data from each field station are reviewed in the field and at the USA universities before they are submitted to the CRSP Program Management Office for verification, consolidation and storage.

CRSP's First Database System

A computerized database system was developed to manage the large quantities of data being produced by the CRSP. Desirable characteristics for the system were:

1. Reliable data entry under adverse conditions in locations far from technical support;
2. Ready access to each station's data on site and at each USA university so that routine reports and statistical analyses could be prepared;
3. Permanent storage of all data submitted by the stations; and
4. The ability to sort and combine data into formats required for large-scale statistical analyses and modelling.

The first database system developed by the CRSP used Apple IIe computers and a small commercial database called General Manager. Although the system could store the data, it was slow and cumbersome, particularly when reorganizing data for preliminary analyses.

The Current Database System

In 1985, the Program Management Office reexamined the options for the CRSP database system which could be implemented using existing equipment. Most of the field stations were still using Apple IIe computers with 128K memory but IBM Personal Computers (IBM PC) had replaced the Apple IIe at one of the stations and several of the universities.

A new system was designed based on commercially-available software. Spreadsheets were selected for data entry because they were relatively easy to use and were similar to tabular data sheets. VISICALC was selected for use on the APPLE IIe whereas LOTUS 1-2-3 was selected for use on the IBM PC. The system was expanded to include MULTIPLAN and APPLEWORKS after VISICALC was removed from the market.

When data are submitted from many locations, the opportunities for misinterpretation are increased. To minimize this, the spreadsheet column headings are complete (i.e., not abbreviated) or use only unambiguous abbreviations (Table 1). Column headings include the units of measurement and an indication of expected magnitude and desired precision (e.g., XX.XX). Another advantage of using complete headings is that preliminary data reports can be printed directly without much additional formatting.

Each line in a spreadsheet table contains a complete set of identifiers (e.g., pond number, date, site, experiment number and season). This organization is very flexible and allows the number of ponds sampled on a given date to vary. Although repeating identifiers on each line is redundant, alternative data arrangements would have required many more tables. Also, the standard data organization of many statistical packages requires lines of data with complete identifiers. Entry of the identifiers is expedited by using the spreadsheets' copy commands.

The spreadsheet tables group data according to data type and collection schedule. The system uses twelve tables that correspond closely to the data collection sheets. The data are entered by pond (i.e., horizontally across a row) or

[1]Pond Dynamics/Aquaculture CRSP contribution CRSP 87:24. This article is an update on Hopkins, K.D., J.E. Lannan and J.R. Bowman. 1987. A Data Base Management System for Research in Pond Dynamics. CRSP Research Reports 87-1, Oregon State University, Corvallis, Oregon, USA. The first article was presented at the World Mariculture Society meeting in Reno, Nevada in January 1986.

[2]The Pond Dynamics/Aquaculture CRSP is funded by grant number DAN-4023-G-SS-7066-00 from the United States Agency for International Development.

[3]Budget reductions caused the number of countries to be decreased to three in September 1987: Panama, Rwanda and Thailand.

Table 1. Example of CRSP data entry template used for diurnal data.

Site	Date Day XX	Month XX	Year XXXX	Time of day XXXX	Pond	Oxygen Top mg/l XX.X	Mid mg/l XX.X	Bot mg/l XX.X	Pond temperature Top deg C XX.X	Mid deg C XX.X	Bot deg C XX.X	pH XX.X
H	21	3	1986	618	B1	6.5	6.4	6.4	20.1	20.1	20.0	7.7
H	21	3	1986	621	C2	6.9	6.8	6.8	19.8	19.8	19.8	7.2
H	21	3	1986	943	B1	6.8	6.6	6.5	20.4	20.3	20.2	8.1
H	21	3	1986	946	C2	6.4	6.1	5.8	20.0	19.9	19.8	7.7
H	21	3	1986	1,311	B1	8.6	7.9	7.5	22.0	21.1	20.8	8.6
H	21	3	1986	1,317	C2	7.5	7.0	6.4	21.2	20.8	20.4	7.7

by variable (i.e., vertically down a column). Each table contains data on all of the ponds at a given station. This arrangement allows each set of observations to be entered without having to shift between several different tables, each containing the data for one pond.

Although spreadsheets were well suited for data entry and preliminary analysis, they could not handle the large quantity of data in the consolidated database. A relational database, RBase System V, installed on an IBM PC-compatible computer, is used for consolidating the data from the field stations. RBase was selected because it could directly read ASCII, LOTUS, MULTIPLAN and VISICALC files and translate them into RBase database format. The consolidated database is stored on a 20 + 20 Bernoulli Box, a data storage device which uses two removable diskette cartridges. Each diskette cartridge has a capacity of 20 megabytes.

Apple-formatted diskettes could not be read directly on IBM PC diskette drives. Data on APPLE-formatted diskettes could be transferred to an IBM PC using a communications package or the APPLE disks could be translated into MicroSoft Disk Operating System format (MS-DOS or PC-DOS format). The latter option was selected and an IBM PC was equipped with an Apple Turnover accessory board and accompanying software. This enabled the IBM PC to read Apple-formatted disks and translate the data into PC-DOS format. APPLE VISICALC and MULTIPLAN files could be directly translated using APPLE TURNOVER but APPLEWORKS files had to be transformed into ASCII fixed-field format before translation to PC-DOS format.[4]

The data are made available to CRSP researchers at several locations. The data are generally distributed using PC-DOS formatted diskettes. If the data are required on computer tape, data are transmitted to a Control Data Corporation CYBER mainframe using a local area network and then are written to tape. Because of the relatively high cost of using a mainframe computer, all sorting and reorganization of data are accomplished using RBASE on the IBM PC, whenever possible.

Suggestions for Database Design

The CRSP experience in establishing its database system emphasized several points which should be considered in designing a database:

1. The system is only as good as its weakest component. The first system's weakest component was the cumbersome nature of the General Manager database for conducting preliminary analyses. Consequently, the field research staff considered data entry into General Manager to be an administrative chore.
2. Administrative chores are not of high priority to researchers. The database system should be easy to use by both the field staff and the home office.
3. Data input forms should use complete headings, not abbreviations, whenever possible.
4. Units of measurement should be included in the data table headings. Although the computer can convert data from one unit to another, conversion can be time consuming for a large number of files. Standardized units of measurement greatly facilitate consolidation of data sets.
5. The system should be accepted by the field staff before data are collected. If the system is not acceptable, it should be improved or replaced before a backlog of data develops.
6. Entering a backlog of data into a database is difficult because collecting new data is much more interesting than filing old data.
7. Incompatibilities between the various computers and between the software packages were the major impediment to a smoothly operating CRSP data management system. Each time data are translated from one computer or software package to another, opportunities for error increase. Standardization of basic equipment within the data management system is highly desirable.

The Pond Dynamics/Aquaculture CRSP encourages other scientists to contribute data to the CRSP database. Cooperating scientists will be allowed access to the entire CRSP database. Data sets from cooperators need not be as complete as those collected by the CRSP projects to be of value. Instructions for Data Entry and templates are available from the author.

[4]Translation of files with a large number of variables into ASCII fixed-field format should be avoided whenever possible. The reason is the typical methods of creating ASCII fixed-field files use print procedures which usually limit line width to less than 300 characters.