MANAGEMENT OF A LONG-TERM WATER QUALITY DATABASE: FLATDAT FOR THE FLATHEAD LAKE BIOLOGICAL STATION

Melissa E. Holmes and Geoffrey C. Poole

The University of Montana, Flathead Lake Biological Station,
311 Bio Station Lane, Polson, MT 59860-9659.

Abstract. Long-term monitoring databases present data management challenges that are unique. The development of an information management system must be carefully planned to determine expectations of the system in terms of use, output and longevity. Data and metadata must be adequate for accurate future analyses. A system must evolve to address an organization's changing needs and take advantage of new technology.

INTRODUCTION

Long-term monitoring databases present unique data management challenges. First, the personnel who collect and manage monitoring data may change over time, often resulting in inconsistencies in the ways data are collected, analyzed, and stored. Second, techniques used to collect monitoring data may change over time due to improvements in data collection methodologies. Third, archiving and documenting data sets that result from ongoing long-term monitoring are difficult because there is often no "final product." Instead, the database is continually growing and represents the current and ever-changing status of the monitoring program.

For more than two decades, the Flathead Lake Biological Station (FLBS) has been monitoring water quality in Flathead Lake and its catchment. In order to address the data management challenges presented by this monitoring program, we began to develop a digital information management system in 1992 called *FlatDat*. By providing a central repository for the FLBS monitoring data, FlatDat helps to ensure that: a) data are collected, entered, and archived in a consistent manner; b) any changes in standard procedures in the field or laboratory are documented; and c) the current status of each project is accurately represented in a location where FLBS researchers can access the data and track progress.

FlatDat provides a total data management solution for the acquisition, calculation, retrieval, and archival of data generated by the analysis of water samples at the station. It tracks the status of each water sample brought into the lab, automates all calculations in our analytical lab by generating different types of electronic worksheets for each lab methodology, archives data in a form that is accessible to researchers, and generates billing reports for accounting purposes.

DEVELOPMENT OF FLATDAT

The development of a data management system requires a significant investment of resources. For the Flathead Lake Biological Station, the largest portion of this investment is in personnel assigned to the project. Approximately half of one full-time Data Manager's time is spent on evaluation, development, and maintenance of the FlatDat system. Several thousand dollars are spent each year on purchasing and upgrading computer equipment and software, and on the personnel required to evaluate and maintain FlatDat. Thus far, the development of the system has been funded entirely by the Flathead Lake Biological Station. The current implementation of FlatDat was written using Microsoft® Foxpro® for Macintosh®.

The management of ecological data has received well-deserved scrutiny in recent years (e.g., Michener et al. 1994). FlatDat was designed based on four premises that arose from such scrutiny: 1) electronic data are most flexible and powerful when stored in the rawest form possible; 2) data must be secure, yet accessible; 3) computerized databases should work the way people work; and 4) data management should be inexorably linked to existing tasks and jobs. Each of these premises is discussed below.

Electronic data are most flexible and powerful when stored in the rawest form possible.

In many ecological databases, only the final data values are entered. These data are often the result of a variety of calculations, manipulations, and transformations that are not evident in the numbers themselves. This can lead to several problems. First, the limitations inherent in a data collection methodology are apt to be ignored and the data are more apt to be misapplied or misinterpreted. Since only the final product (i.e., the final value) is available, people generally *assume* a high level of accuracy in the methodology. Even if the database user questions a particular datum, the only choices are "take it" or "leave it" and, frequently, no means are available to check the datum by tracking its genesis.

On the other hand, if raw values are entered into the database and all calculations are programmed and automated, the database can be used to investigate the genesis of any particular datum. Readings from field or lab instruments are available for scrutiny as are the calculations used to convert raw data into finalized values. Additionally, in the event an error is discovered in the calculations, or if an investigator wishes to calculate final values using a different method (for instance, to compare numbers to another study that used an alternative method), then the new calculations can be programmed into the computer and the entire database recalculated automatically. Programming calculations into the database ensures consistency in calculations over time and generates an accurate record of any changes in methodology.

In FlatDat, raw data (e.g., readings from analytical equipment) are entered into electronic laboratory worksheets. These worksheets are built automatically by FlatDat based on the status of samples that have been collected and logged into the computer by field personnel. If necessary, electronic worksheets can be altered easily by the chemical analyst to delay analysis of some samples or include quality control samples if necessary. The analyst then prints out the blank worksheet, writes down the raw readings on the hard copy as the analyses are run (thereby providing a hard-copy for future reference), and enters the values from the hard copy back into the electronic worksheet. The computer performs necessary calculations as the data are entered and prints a final copy of the worksheet when saved. The hand-written hard copy and final hard copy are compared to check for errors, stapled together and filed for future reference. Through the use of a "quality control number" ("qc number") assigned to each value resulting from the worksheet, any value in the database can be traced back to a specific worksheet generated in the analytical lab.

This provides the ability to track down errors when unlikely values are discovered during data analysis and ensures consistency over time. However, consistency over time only extends as far back as raw data are entered. In order to provide this level of accuracy, consistency, and scrutiny to our entire period of record, all *raw* lab readings from 1979 through the first implementation of FlatDat in 1992 have also been entered into FlatDat. During the ~20 year period of record, several methodological changes were made in the lab, including changes in detection limits and purchases of new instruments. FlatDat is able to accommodate such changes because each analysis for each sample is linked, via the electronic worksheet and qc number, to the methodology used at the time it was run. To date, FlatDat contains the results of over 200,000 individual water quality analyses from tens of thousands of water samples collected in the Flathead Basin.

Data must be secure, yet accessible.

In order to be most useful, long-term monitoring data must be easily accessible to researchers in formats that are flexible and useful. However, unrestricted access to the database risks corruption or loss of the data. FlatDat provides a simple query window that allows even novice users to formulate complex queries. The results of the query can be saved to external files that can be imported into statistical software or spreadsheets. FlatDat employs several levels of security as well. First, users must log into the FlatDat program to use the database. Users are granted privileges that are appropriate to the level of access required. Most users can only view and query data. Field technicians can log samples into the computer and edit samples that have not yet been analyzed in the lab. Chemical analysts can edit electronic worksheets. Some especially powerful tasks (such as altering the pre-programmed methodologies or recalculating the database) are only available to the database administrator. Any time a change is made to a sample description or electronic worksheet, FlatDat records the date of the change and the person who made the change.

Assigning appropriate access privileges to the networked disk drive where the data are stored provides additional security. Additionally, we designed a database "maintenance" utility that checks the database for logical errors and rebuilds indexes associated with each database table. Finally, nightly incremental backup of the database ensures that we can restore the database to the state in which it existed at any particular date in the past. On a monthly basis, a backup copy of the database is stored off-site to protect against catastrophic loss.

Computerized databases should work the way people work.

A computer program should always save time. It should never make the user's job more difficult. When an information management system is being designed, the number of users and their skill levels need to be considered. FlatDat utilizes a graphical user interface with windows and forms making the program easy to use and preventing users from inappropriately seeing or changing the underlying data files. Controls in these windows and forms limit the information that can be entered into each field. For instance, some controls only allow users to enter a data element from a list of available elements, force the user to enter data in a particular format, or automatically default values such as dates and user names. Restrictions like these help to ensure ease of use and accuracy, but reduce flexibility for more advanced users. In developing FlatDat, we worked closely with the users to design a system that strikes a balance between data security, ease of use, and flexibility. During development, users were encouraged to provide feedback regarding what they liked and disliked about the database and whether or not solutions were efficient or inefficient. To the extent possible, the program was modified to incorporate this information while ensuring data integrity.

Data management should be inexorably linked to existing tasks and jobs.

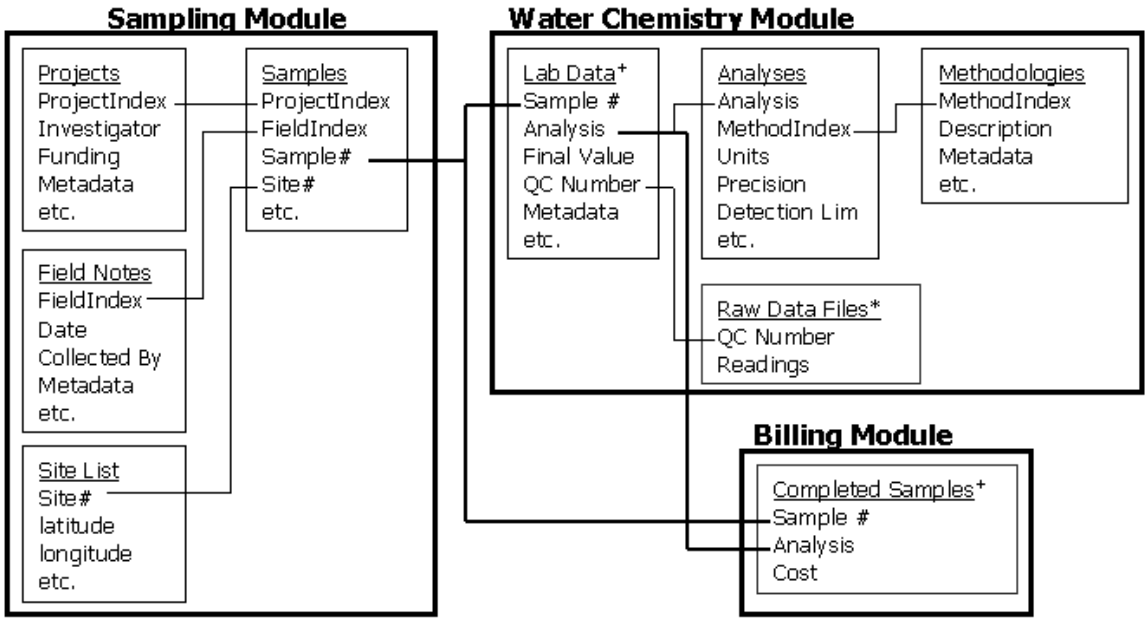
A primary goal of FlatDat was to allow individual employees to manage the data they need to manage in order to do their job. Additionally, we intended to reduce duplication of effort, improve communications regarding data resources, and increase productivity (*sensu* Michener and Haddad 1992). This was sometimes difficult, since some employees are inexperienced with or even resistant to using a computerized database. However, since field personnel most accurately know what happens in the field and lab personnel know what occurs in the lab, the database will be most accurate if both groups are empowered to manage the data. To accomplish this, data entry screens were designed to look like the field and laboratory sheets already in use at the time when FlatDat was implemented. This provided a comfortable and familiar electronic "environment" for personnel

who were not necessarily skilled computer users. The system is entirely menu- and mouse-driven to accommodate the novice, yet liberal use of "short cut" keystroke and "hot-buttons" allows rapid navigation by experienced users. Again, the use of forms and windows that limit what a person is allowed to enter or change is critical. However, ample opportunity to enter field and lab notes also provides a means of recording critical information in a more flexible format.

DATA FILES AND RELATIONSHIPS

FlatDat consists of three modules: the Sampling Module, the Water Chemistry Module, and the Billing Module (Figure 1). The Sampling Module contains information about water samples that have been collected. Each sample belongs to a project that has been developed by a particular investigator or group of investigators. Sample collection sites for all projects are stored in the same "site list", thereby encouraging cross-compatibility between projects over the long term. The "field notes" table contains data and metadata that describe the field sampling trip and the conditions under which the samples were collected.

Figure 1. Relationships diagram for the FlatDat system. Fine lines indicate linked data fields within a module. Bold lines indicate linked data fields between modules.



*Each methodology has its own raw data file; the specific raw data file associated with a particular analysis is determined by the methodology used.
+Each sample can have multiple records in this database; one record for each chemical analyses performed

The Water Chemistry Module contains data and metadata describing laboratory analyses of water chemistry. Any number of analyses may be run on each sample. The "lab data" table stores the results from each analysis, along with quality control information and lab notes (metadata). The raw data used to calculate the final values are stored in the Raw DB Files. Since each methodology run in the lab generates different raw data, each methodology has a separate and unique raw data file.

A unique sample number is assigned to each sample as it is collected. This sample number is stored in both the "samples" table and the "lab data" table and is used to link data in the Sampling Module to data in the Water Chemistry Module.

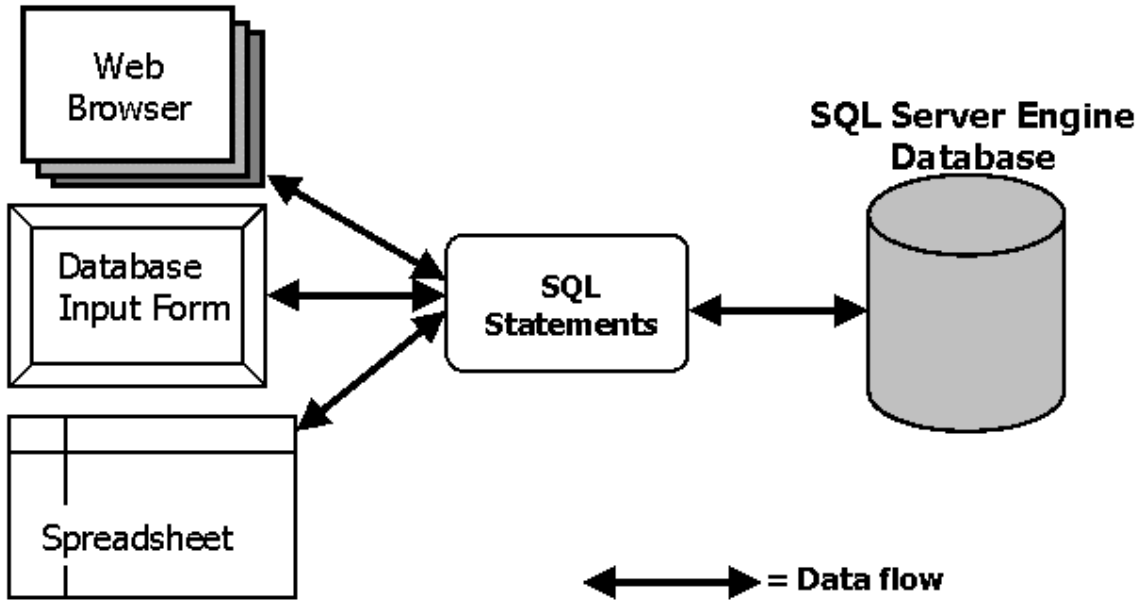
The Billing Module uses data from the "samples" and "lab data" tables to determine which water samples are complete (i.e., all requested analyses are run and entered into FlatDat). Information about completed samples is pulled into the Billing Module, where an invoice report is generated and the sample is marked as having been billed.

FUTURE IMPROVEMENT FOR FLATDAT

The advent of a mixed-platform environment at the FLBS, along with the needs for Internet access, better database tools, and better integration of metadata (*sensu* Ingersoll et al. 1997) have influenced the design of the next version of FlatDat. Although calculations and methodology have remained much the same, software and storage needs have changed entirely.

In the next version of FlatDat, the data will be stored in a centralized Structured Query Language (SQL) server that will be able to provide query results to a variety of applications including web browsers (Figure 2). This will allow more flexibility in matching the capabilities of existing software to requisite tasks, thereby reducing the amount of custom programming required. For instance, database forms can be used to log samples, spreadsheets can be used to create electronic worksheets, and the database can be queried via a web browser allowing cross-platform access to the data. In a manner similar to Ingersoll et al. (1997) we plan to make specific portions of the data available to the public on the World Wide Web. We are currently in the process of evaluating specific software tools for use in the next version of FlatDat.

Figure 2. Data flow in the FlatDat system. The data will be stored in a SQL server that will provide query result to a variety of applications, including web browsers and spreadsheets.



Recently, issues surrounding the importance of managing and incorporating metadata have received overdue attention (e.g., ESA 1995, NRC 1995). Having finished compiling and entering our historic water chemistry data (i.e., data from the period 1979-1992), the need for organization

and integration of our metadata has become evident. While long-term FLBS employees may know where to access metadata in old field-logs and journals, newer employees do not know what metadata are available or where to find it, thereby reducing the utility of our consistent and meticulously organized data set. The new version of FlatDat will include fields for tighter integration of metadata and, as we did for analytical data, all existing field notes and other types of metadata will be entered for the entire period of record. The new version will also feature a more robust billing module, including the capability to easily bill multiple funding sources for a particular project.

CONCLUSIONS

When preparing to build an information management system, system requirements and design are extremely important and should be completed before implementation begins. The developer should work with management and key users to decide what information needs to be tracked and what is expected of the system in terms of use, output, and longevity. As the needs of an organization change and computers and software become more advanced, an information management system must evolve to address those needs and take advantage of new technology.

The expenses associated with the maintenance and evolution of information systems must be an *a priori* acknowledgement and planned for accordingly over the long term. With only occasional exception, biologists and ecologists can no more be expected to manage data effectively than information specialists could be expected to design and conduct ecological research. While research staff, students, and scientists must be involved in the data management process in order for it to be effective, there is no substitute for hiring and retaining staff who are trained in information system management to oversee the data management process.

The development of FlatDat has been a learning process for everyone involved. Our progress to date has increased the utility of our data set and the productivity of investigators at the FLBS. We expect that the next implementation of FlatDat, based on the needs and principles outlined above, will fulfill our data management needs well into the next millennium.

LITERATURE CITED

Gross, K.L., C.E. Pake, E. Allen, C. Bledsoe, R. Colwell, P. Dayton, M. Dethier, J. Helly, R. Holt, N. Morin, W. Michener, S.T.A. Pickett, and S. Stafford. 1995. Final report of the Ecological Society of America Committee on the future of long-term ecological data (FLED). Volume I: Text of the report. (*http://www.sdsc.edu/~ESA/FLED/FLED.html*).

Ingersoll, R.C., T.R. Seastedt, and M. Hartman. 1997. A model information management system for ecological research. BioScience 47: 310-316.

Michener W. K., J. W. Brunt, S. G. Stafford, editors. 1994. Environmental information management and analysis: ecosystem to global scales. Taylor and Francis, London, UK.

Michener, W.K., K. Hadadd. 1992. Chapter 1-Database Administration. Pages 4-14 in J.B. Gorentz, editor. Data management at biological field stations and coastal marine laboratories. Report of an invitational workshop, April 22-26 1990, W.K. Kellogg Biological Station, Michigan State University, East Lansing, MI.

[NRC] National Research Council. 1995. Finding the forest in the trees: the challenge of combining diverse environmental data. National Academy Press, Washington, DC.