Data Management: Could and Should Libraries Breach the Bibliographic Barrier?

Janet Webster
Guin Library
Hatfield Marine Science Center
Oregon State University
Newport, Oregon 97365

ABSTRACT

Marine libraries serve as access points to a multitude of bibliographic information which is vital to the pursuit of science. Yet of growing importance is access to data by researchers. Is there a place for the library in the management of scientific data? In April 1990 the United States National Science Foundation sponsored a workshop on data management at biological field stations and marine labs. The workshop's report, due in October 1990, could redefine the NSF approach to data management and include libraries in the process.

*****

I find comfort in a book - a nice package of two hard covers and a text block that has some heft. Even better if it has some nice illustrations. Monographs, journals and serials are our mainstays. But as we meander into the realm of electronic information, we stretch our concepts of what should be in a library. Many of us have bibliographic datasets - CD-ROMs, OPACs, diskettes - and we have online access to others[1]. But do many of us have other than bibliographic datasets - the data analyzed for some graduate student's thesis, the spreadsheet of species distribution, the tidal and weather data for our locale?

Frankly, I hadn't thought much about data. Bibliographic material provides plenty of food for thought as well as needing plenty of management. But in April, I had an eye-opening plunge into the world of data management. The United States National Science Foundation sponsored a workshop on data management at biological field stations and marine labs and I had the opportunity to attend. Fred Lohrer of Archibold Biological Station and an IAMSLIC member and I were the two librarians of the thirty-five or so participants. The workshop held at Kellogg Biological Station brought together data managers, field biologists, librarians and station administrators. Its purposes were to review a report generated from a similar workshop held in 1982[2] and, to generate recommendations to site administrators and NSF on the administration of data, on data standards, and on hardware. The 1990 workshop report, due in late 1990, is a group effort. The participants, divided into three working groups, drafted the report. Since April, several of us have reviewed the drafts and added to them. The workshop, set up along the Dahlem model[3], was a successful forum for contemplating the data management question.

The discussions of that week made me think hard about the role of the library *vis a vis* data management. Librarians are involved in data management but, in general, we manage data that has been formatted for us. Do we have a responsibility to provide access to data, to archive it and/or to organize it? Do we want to accept that responsibility and do we have the resources even if we wanted to?

Since returning from Michigan, my continuing thoughts have focused on service concerns, hardware decisions, and policy choices.

The library is a service-oriented institution. It is the obvious access point to a variety of information. Visiting researchers often start at the library to find out what's been done at a particular site before and by whom. Resident researchers use the library to stay current and to explore past work.

There are times when access to a station bibliography, to species lists, to site descriptions or to tidal data would be helpful. Access to such data might reduce the amount of time a researcher takes to find what he or she really wants. Access to another site's species list might be crucial to the study of particular ecosystems.

The service implications of data management are large. How could we provide electronic access to another library's dataset? Who pays for the access? How do we know what's out there? These aren't new questions for librarians but potential demand for access to datasets just bring them up again.

In addition to access, preserving data could become another challenge to library service. There is no doubt that potentially valuable data is continually lost. A graduate student reformats a disk and hence erases the data used in a dissertation. Yes, it has been published but what of the researcher ten years later who tries to reconstruct what that student did. The data is gone and only the conclusions drawn from it remain. There is a need for continuity of data. What is valuable to us may not have looked important to a researcher some years ago. Libraries traditionally play a role as a storehouse for information. Do we want to add datasets to our collections? If not us, who will preserve the record?

Datasets come in all configurations. So, the hardware problems alone are daunting. There is the access question raised earlier. How can you provide local access to data that comes on various size diskettes, or even on magnetic tape? What if it uses a software your library does not own? Egad, what if it is not for a

PC but needs a Macintosh? Again these are not new questions but their solutions are crucial to accessing data.

Remote access is important not only so data can be shared between sites, but so data can be jointly compiled by researchers at different sites. It is not impossible to transfer datasets but people need to be informed on the ins and outs of the process.

We are getting better at coping with hardware variety. Standards in documentation, hardware, software and format are possible (or at least we all hope they are so). Workshop participants in Michigan observed that "in the context of databases, the basic purpose of standards is to enhance communication among users of the data be they the original or future investigator."[4] Librarians are accustomed to standards. Our MARC format was mentioned at the NSF workshop as a splendid example of a data standard that worked. Simple, nonintrusive standards would be useful and would alleviate many of the problems of data access, transfer and interpretation.

Policy choices perhaps provide the biggest opportunity for librarians to be involved in data management. We have experience selecting, organizing, and communicating information. We work with researchers and have a glimmer as to their information needs. We work with and for our administration, and hence, have some familiarity with the institution's mission, structure, and priorities. Policy choices for data management involve all of the above.

In order for data management to be successful, the administration must be supportive both financially and philosophically. Librarians are very familiar with this: if an institution is not committed to its library, that library does not flourish. Administrators could involve librarians in the process of incorporating data management into an institution's structure. This does not mean a commitment to have the library as the center of the data universe; but it does imply a recognition of the librarian's role in information issues.

Data management is dependant on researcher's data. If that group of people is not convinced of the importance of making their work accessible or at least comprehensible, data management will not be successful. Most researchers use libraries and communicate with librarians somewhat regularly. This relationship could be used to explore the issues of data management and lead to its incorporation in the research process. Is it that unreasonable for the librarian to ask a departing graduate student for a copy of his or her data, even if only doe back-up or archival purposes, - especially if their experiments were set up in a previously unstudied area?

Quality of datasets is problematic. Selection of serials and monographs is at times highly subjective. Datasets too will need to be examined, selected and catalogued for access and archival purposes. Preserving the integrity of datasets is another quality issue. Just as someone can razor blade color plates out of a folio, someone could alter data with a few quick key strokes. Researchers' concerns over protecting, updating and interpreting data are valid. Quality can rarely be guaranteed.

As librarians, we have expertise that could make us invaluable in the development of data management at our individual institutions. We deal with service issues, technology choices and policy decisions daily. NSF is considering funding for more data management related projects. The Foundation is very concerned with how field stations handle their data. Currently the LTER sites (longterm ecological research sites) have major data management components which utilize geographic information systems. NSF sees this development as a potential model for smaller sites and several such sites are implementing GIS[5]. But NSF does not currently see librarians as players in data management. I think the Foundation should reconsider its position. As an active member of a field station, the librarian should play a role in data management.

At the April workshop, I made sure that libraries were recognized as possible access points to data. I tried to communicate that librarians are concerned with the scientific process and that just as we are involved in the management of bibliographic data, we could manage other types.

I am not an evangelist for librarians as data managers. I like what I do as a librarian. But I am intrigued by the issues raised by data management. Libraries need to remain vital and relevant to science and this area is going to be of growing importance.

My compatriots at the NSF workshop have drafted the following:

> "Science may be described as a process whereby facts about the real world are coaxed out of data sets which were designed to store observations of real world patterns and processes. Various methods may be employed in the scientific process but all ultimately rely on the availability of high quality and well documented data. All scientists participate in data management activities to varying degrees. Data management may therefore be viewed as a critical component of the scientific process."[6]

What I would like all of you to contemplate is this: where does a librarian fit in the scientific process? And given some of the issues I've raised, what is our responsibility to breach the bibliographic barrier and address data management as a library concern?

REFERENCES:

(1) Parker, Joan. survey in these proceedings.

(2) *Data Management at Biological Field Station: Report of a Workshop May 17-20, 1982, W.K.Kellogg Biological Station, Michigan State University.* National Science Foundation. 46 pp.

(3) The Dahlem Workshop model was developed to provide an interdisciplinary approach to exploring and understanding scientific problems and issues. Participants are divided into three or four interdisciplinary working groups. Background papers are selected and serve as the basis for discussion. During the workshop, each group interacts with the other groups and then generates a report of either answers to questions posed in the background papers, or insights gained on the particular topic.

At the Michigan workshop, participants were divided into three groups and all had read the 1982 report as background. By the end of the week, all three groups had draft reports addressing specific questions posed by the workshop organizers and NSF. Additional drafts were written by the group leaders with input by participants during the ensuing months. The finished report is expected by the end of 1990.

(4) Brunt et al. "Outline for Data Standards Group: Second Draft -We're Getting Closer." p.1. Unpublished draft from 1990 workshop data standards group.

(5) Bodega Bay Marine Lab and Archibold Biological Station are two examples of small stations implementing GIS.

(6) Michener et al. "Administration Highlights." p.4. Unpublished draft from 1990 workshop data administration group.

FURTHER READING:

Brenneman, Judy and Tawny Blinn (eds.) *Long-term Ecological Research in the United States: A Network of Research Sites 1987.* 4th ed.,rev. Corvallis, Oregon: Forest Science Department, Oregon State University, 1987. 39 pp.

Data Management at Biological Field Stations and Marine Labs: Report of a Workshop held at the W.K.Kellogg Biological Station, April 22-26, 1990. National Science Foundation. In press.

Data Management at Biological Field Station: Report of a Workshop May 17-20, 1982, W.K.Kellogg Biological Station, Michigan State University. National Science Foundation. 46 pp.

Experimental Ecological Reserves: A Proposed National Network. Washington, D.C.: The Institute of Ecology; June 1977.

High Performance Computing and Networking for Science. Washington, D.C.: Government Printing Office, 1989. GPO #052-003-01164-6. 48 pp.

Moritz, Tom. 1989. "Towards Community Standards in Systematics." *Spectra* 16(2):18,20.

Research Needs of Biological Field Stations: Report of a Workshop at the W.K. Kellogg Biological Station, October 21-25, 1984. National Science Foundation. 42 pp.

POSSIBLE DATA MANAGEMENT ACTIVITIES FOR LIBRARIANS

Basic Activities

Develop a fact sheet describing the site.
Standardize place names and abbreviations.
Maintain accurate site bibliography.
Collect all published research which references the site.

Policy/Management Activities

Assist with the development of sites data management policy.
Develop collection policy which includes datasets.
Prioritize datasets for archival and purchase purposes.
Develop guidelines for archival storage of datasets.
Formulate procedure for use of datasets.

Acquisitions Activities

Collect basic data sets (temperature, rainfall, etc.).
Accumulate copies of datasets for security/archival purposes.
Develop guidelines for collecting student datasets.
Request all datasets that relate to site's locale or mission.
Catalog datasets.

Reference Activities

Develop expertise with local data and research
Identify existing datasets and ones that are needed.
Develop procedures for exchanging datasets between sites.
Be familiar with software/hardware needed to access datasets.

EXAMPLES OF DATASETS USEFUL TO FIELD STATIONS

Bibliographic Datasets

Published/unpublished papers of onsite researchers

Published/unpublished papers which reference the site

Student papers

Library/reading room holdings

Administrative and anecdotal site histories

Scientific Datasets

Temperature ranges

Elevations and geologic formations

Stream flows

Species lists and habitats

Tidal ranges

Researchers' Datasets

Site plans

Species distribution

Lab results

Long term monitoring of habitats