A Brief History of Data Sharing in the U.S. Long-Term Ecological Research Network

John H. Porter, University of Virginia

In the abstract, the advantages of sharing data are manifest. No individual scientist, or even small group of scientists, can collect all the data that is needed to address the major ecological research questions of today. The integration of data from diverse sources opens new opportunities for research to the common benefit of the ecological research community and humankind in general (National Research Council 1997, Porter 2000). However, data sharing remains more the exception than the rule throughout most of ecology. The U.S. Long-Term Ecological Research (LTER) Network has had some successes in promoting the sharing of data, but it was not always so. For that reason, it is of interest to projects hoping to promote data sharing, to review the steps the LTER Network took in achieving those successes.

The U.S. Long-Term Ecological Research (LTER) Network has a strong policy to promote the sharing of data, both inside the LTER Network and with the larger scientific community. The data access policy (Table1) was adopted in 1997 by the LTER Coordinating Committee (the governing body of the U.S. LTER Network), but it is only the most recent of a series of efforts to make needed data available to researchers. The LTER Network now has over 3,000 data sets in its on-line data catalog.

Table 1: DATA ACCESS POLICY FOR THE LTER NETWORK (1997) (http://lternet.edu/data/netpolicy.html)

1) There are two types of data: Type I (data that is freely available within 2-3 years) with minimum restrictions and, Type II (Exceptional data sets that are available only with written permission from the PI/investigator(s)). Implied in this timetable, is the assumption that some data sets require more effort to get on-line and that no "blanket policy" is going to cover all data sets at all sites. However, each site would pursue getting all of their data on-line in the most expedient fashion possible.

2) The number of data sets that are assigned TYPE II status should be rare in occurrence and that the justification for exceptions must be well documented and approved by the lead PI and site data manager. Some examples of Type II data may include: locations of rare or endangered species, data that are covered by copyright laws (e.g. TM and/or SPOT satellite data) or some types of census data involving human subjects.

However, in the late 1980s, researchers within the LTER Network were no more likely to share data than were other researchers. For example, during a visit to discuss connections to the (then newly available to researchers) Internet in 1989, the lead PI of one LTER site forcefully stated (literally banging his fist on his desk for emphasis): "If being connected to the Internet means people can get access to our data, **we don't want it!**" His concerns were somewhat ameliorated by a discussion of Internet security (few of us would want to

expose the entire contents of our computers to the world), but he also expressed concerns about "stealing data" or being "scooped" on publications.

To address these, and other, concerns, at the 1990 LTER "All-Scientists Meeting" an *ad hoc* committee was convened. With substantial input from the data managers and investigators at the LTER sites, the group stopped short of recommending a network-wide policy for data sharing, and instead developed guidelines for site information policies at individual sites (Table 2).

Table 2: LTER GUIDELINES for SITE DATA MANAGEMENT POLICIES (1990)

Each Long-Term Ecological Research (LTER) site should develop its own data management policy in consultation with key investigators and higher administrative units. The following provides general guidelines and rationale, but each site should be prepared to defend its own policy through the site and peer review process. The general policy of the Division of Biotic Systems and Resources, National Science Foundation (NSF), is that the data are public property one year after termination of the relevant grant.

General Guidelines:

The management policy should include provisions that assure:

- The timely availability of data to the scientific community;
- That researchers and LTER sites contributing data to LTER databases receive adequate acknowledgement for the use of their data by other researchers and that sites receive copies of any publication using that data;
- That documentation and transformation of data is adequate to permit data to be used by researchers not involved in its original collection;
- That data must continue to be available even if an investigator leaves the project through transfer or death;
- That standards of quality assurance and quality control are adhered to;
- That long-term archival storage of data is maintained;
- That researchers have an obligation both to contribute data collected with LTER funding to the LTER site database and to publish the data in the open literature in a timely fashion;
- That costs of making data available should be recovered directly or by reciprocal sharing and collaborative research;
- That LTER data sets not be resold or distributed by the recipient; and
- That investigators have a reasonable opportunity to have first use of data they collected.

In addition to the guidelines, the also included a sample policy, addressing four different classes of data, that would meet the requirements of the guidelines (Table 3).

Table 3: Sample policy meeting the criteria listed in the guidelines (1990)

The following is an example of a policy that will meet these guidelines with respect to data sharing:

Data Type I. Published data and meta data (i.e., data about data).

Policy: Data are available upon request without review.

Data Type II. Collective data of the LTER site (usually routine measurements generated by technical staff).

Policy: Data are available for specific scientific purposes one year after generation.

Data Type III. Original measurements by individual researchers.

Policy: Data are available for specific scientific purposes two years after generation.

Data can be released earlier with permission of the researcher.

Data Type IV. Unusual long-term data collected by individual researchers.

Policy: The principal investigator of the LTER site can designate that such data can be withheld for longer periods Such action should be rare and justified in writing.

Why did the committee adopt guidelines and not a policy? A major reason was that, at that time, the whole notion of sharing data was foreign to most of the researchers in the LTER network. A dictatorial policy could have driven researchers away from the network, fearing that their data was being "stolen." Moreover, researchers active participation in the preparation of documentation (metadata) would be required if the data were to be usable by others (Michener et al. 1997). A second reason was that ecological researchers in general had little experience in crafting data sharing policies. It was not known what would work and what wouldn't. By adopting guidelines rather than a policy, the committee engaged the researchers at each LTER site in the design of guidelines that would work, at least at the level of the individual site.

Researchers can't ask for data they don't know exists, so the year 1990 also saw the first version of an LTER-wide data catalog (Michener et al. 1990) in printed form. This catalog consisted summary descriptions of datasets (but not the data themselves) for a minimum of 10 "core" datasets from each of the LTER sites. Even this catalog was an ambitious step in a time where both sites and researchers clung closely to their data.

However, once created, it helped alert the research community to the data resources and their potential to address a wide array of inter-site questions.

By 1993, the majority of LTER sites had site-specific data sharing policies in place. The resulting individual site policies were reviewed by (Porter and Callahan 1994). A simple additive model, based on the time devoted to different tasks, was used to examine "evolutionarily stable strategies" for promoting data sharing. One conclusion was that sharing data in the absence of policies that mandated attribution was unsound. Individuals who share their data need to be rewarded, either through receiving scientific credit via acknowledgement, citation or co-authorship, or by receiving financial remuneration (e.g., royalties, increased likelihood of future grant funding). Not surprisingly, many of the LTER site data policies echoed the sample policy (Table 3), with the identification of different classes of data and allowing data collectors to withhold data for designated time periods.

During the period of 1991 through 1993, a revolution in information technology occurred with the advent of Internet-based on-line systems (FTP, gopher and subsequently the World-wide Web). The 1990 guidelines and the site data sharing policies did not explicitly address the on-line sharing of data. In 1990, only one LTER site operated a dial-up "bulletin board" type system, but by 1993 most of the sites had on-line capabilities in one form or another. With substantial urging by the National Science Foundation, in 1994 the LTER Coordinating Committee mandated that by the end of the year, each site should have at least one dataset available online.

The trickle of on-line datasets started in 1994 turned into a flood in subsequent years as LTER sites competed with one another in seeing who could make the most information available. Reviewers of funding proposals increasingly look at the value of the data being made available by an LTER site in deciding whether to continue funding. For this reason, the adoption in 1997 of the simplified network-wide data sharing policy (Table 1) was relatively simple. The experience of the LTER network in the intervening period is that sharing data has done a great deal more good than harm. Although data from LTER sites is widely used by researchers or students working on class projects, there has yet to be an identified case where LTER data was used under false pretenses. Usually the users of the data are more than happy to properly give credit to the providers of the data in their publications, or even to include them as co-authors.

What are some of the lessons learned during this process? The first is that it is important to engage the researchers or institutions that will be providing data in the formulation of data policies. Only if they are comfortable with the provisions in the policy will they contribute. A second lesson is that all the "stakeholders" in the data sharing process have responsibilities. The data collector is responsible for collecting high-quality data and supporting metadata and for providing access to it. The manager of the on-line system managing the data has the responsibility for archiving and preserving the data and metadata and to make sure that it is only accessible to users within the context of the policy (e.g., authorized users). Finally, and perhaps most importantly, the data user has the responsibility for properly acknowledging or citing their use of the data.

The final lesson is that it is not necessary to do it all at once. A critical aspect of building confidence in the concept of sharing data was the stepwise approach, which allowed data contributors to develop a level of comfort at each stage before moving to the next step.

The LTER Network is not the only group that has been working to promote data sharing. (Arzberger et al. 2004) lay out a set of operating principles for data access. Various governmental entities such as the NASA Distributed Active Archive Centers and the NOAA National Ocean Data System

Literature Cited

Arzberger, P., P. Schroeder, A. Beaulieu, G. Bowker, K. Casey, L. Laaksonen, D. Moorman, P. Uhlir, and P. Wouters. 2004. An international framework to promote access to data. Science **303**:1777-1778.

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

Michener, W. K., A. B. Miller, and R. W. Nottrott, editors. 1990. Long-Term Ecological Research Network core data set catalog. Belle W. Baruch Institute for Marine Biology and Coastal Research, University of South Carolina, Columbia, SC.

National Research Council. 1997. Bits of Power: Issues in global access to scientific data. National Academy Press, Washington, D.C.

Porter, J. H. 2000. Scientific databases. *in* W. K. Michener and J. Brunt., editors. Ecological Data: Design, Processing and Management. Blackwell Science Ltd., London, UK.

Porter, J. H., and J. T. Callahan. 1994. Circumventing a dilemma: historical approaches to data sharing in ecological research. Pages 193-203 *in* W. K. Michener, S. Stafford, and J. W. Brunt, editors. Environmental Information Management. Taylor and Francis, Bristol, PA.