DATA ARCHIVAL

Richard J. Olson and Raymond A. McCord

Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831-6407

Abstract. A data archive is a permanent collection of data sets with accompanying metadata such that secondary users can readily acquire, understand, and use the data. Although data archival for ecology is in its infancy and there are a limited number of permanent data archives for ecological data, ecologists can manage their data in ways that facilitate data sharing and prepare their data for eventual archival. In this chapter, incentives for archiving data are presented, components and functions of data archives are reviewed, and future directions for data archival are discussed.

INTRODUCTION

Traditionally, science involves making systematic observations that can be replicated. Over the past two decades, there has been a shift from traditional studies of isolated ecosystems toward more broad-scale modeling, synthesis, and assessment studies. Scientists are collecting data over the Internet in addition to doing field or laboratory work. As ecology moves toward regional and global multidisciplinary studies, mechanisms for sharing data with many disciplines (meteorology, hydrology, soil science, forestry, agriculture, botany, etc.) are needed. Consequently, submitting data to archives and acquiring data from archives are integral parts of today's scientific process. However, data archival is not yet given the attention, resources, or recognition required for it to become a routine part of the research and publication cycle.

A data archive is a permanent collection of data sets with accompanying metadata such that secondary users can readily acquire, understand, and use the data. An archive preserves data and metadata in an electronic form that will continue to be accessible as technologies change. An archive provides complete metadata so that secondary users with interests varying from watersheds to global change understand inherent limitations of the data and use the archived data properly in new applications. Whereas "archive" may imply simple preservation, the implicit goal is to facilitate data sharing to foster broader ecological discoveries. Archives are more than a long-term backup or an index or catalog with pointers to data sets stored elsewhere.

Data archival for ecology is in its infancy and there are a limited number of permanent data archives for ecological data. Often, different terms and functions are associated with data sharing and storage activities (Table 1). Informal sharing, repositories, or digital libraries may provide much or all of the functionality of an archive. The concepts presented here for data archives readily apply to other, less-formal data storage activities.

INCENTIVES FOR DATA ARCHIVAL

Although most ecologists may support the concept of data archival and even use data from archives in their research, they generally do not archive their own data. There is a trend for federally sponsored research announcements to require that the data generated by any proposed project be placed in an archive. In this section, we explore some of the factors that may contribute to avoidance of data archiving and some incentives to promote data archiving. In many ways, archiving data is equivalent to preparing a publication. The time and resources required to archive data can be equal to or exceed those for preparing, editing, and reviewing a publication. There may be uncertainty about the detail, format, and style of metadata (see Michener et al. 1997). In addition, there may be fears that providing immediate public access to data may result in others

preempting the contributing investigator’s opportunity to publish his or her findings

Table 1. Distinctions between data exchange and storage activities.

Venue	Data Exchange	Comments
Data custodianship	Data sharing by request, usually with colleague having technical expertise	Current expert, provides technical information, could authorize changes to the data, may be primary compiler or inherited role
Data stewardship	Data sharing by request	Gatekeeper, minimal knowledge of the data, may inherit data from custodian
Data repository	May limit data access	Usually project-level support; limited functionality, may cease to exist after project ends
Digital library	Public access	Broad subject area, limited expertise and user support, includes tabular and graphic data
Data archive	Public access	May have thematic emphasis, search and order, long-term commitment, packaged metadata

first. Unfortunately, the scientific community, especially at administrative levels, currently does not acknowledge well-documented data as equivalent to a publication.

In order to facilitate free and open exchange of data among ecologists and to create the reward structure necessary to encourage ecologists to share and exchange data (Porter and Callahan 1994), the ESA Data Sharing and Archive Committee is proposing that the ESA create a venue in which the Society will electronically publish peer-reviewed data papers (Ellison, A. M., personal communication Oct 24, 1997). Data papers are envisioned to include extensive data and metadata (basically an expanded materials and methods section without results and discussion sections). If approved, the Committee will develop guidelines for publishing data and metadata and an appropriate peer review process. The data and metadata would be maintained in a long-term archive, and contributors and users would be able to cite the published data papers as they now cite papers in other ESA journals.

Project leaders, sponsors, and science managers can provide the following incentives for investigators to archive data:

- establish a citation policy to give credit to data contributors,
- establish a citation policy to give credit to multiple contributors to integrated data sets,
- provide adequate resources for data management to investigators,
- involve data personnel in the initial project planning,
- provide guidelines and training for metadata preparation,
- produce high-visibility data products (e.g., CD-ROMs or hardcopy data products), and
- give credit (i.e., include in promotion and salary actions) to those who produce well-documented data sets.

DATA ARCHIVE COMPONENTS

Archives consist of more than data and metadata. Other key components are the data storage system, information system, network connections, a security and backup system, data analysis tools (optional), archive staff, and, most importantly, a user services support staff. Although we will not fully discuss the computer technology component of data archives, the advances in this area, especially PCs, networks, and the World Wide Web (see Porter, this volume), contribute greatly to the growth of data archives. Typically, staff are a mix of systems specialists, database administrators, user interface specialists, information specialists, and scientists.

The Environmental Sciences Division of Oak Ridge National Laboratory (ORNL) has a 25-year history of managing and archiving ecological data, starting with the data from the International Biological Program (IBP) in the early 1970's. Currently ORNL has four data archive centers (Table 2). Each center uses a different technology and organization; however, all emphasize the combination of computer specialists and scientists, provide useful metadata, and supply citation information so that the original data contributors can be correctly acknowledged.

We have found that it is essential that scientists be involved in the archive operations. Scientists play a critical role in the organization and presentation of data, quality assurance/quality control review of data and metadata, and development of value-added products. In addition, an advisory group can effectively represent the interests of data contributors, secondary users, and sponsors.

DATA ARCHIVE FUNCTIONS

The flow of data from a contributor to a publicly accessible archive is a multi-step, potentially time-consuming process. Although there may be many variations on the overall process, an initial step is to connect a contributor to the appropriate data archive. Most archives continually work to identify user needs and data availability and to establish priorities to acquire new data. Often, a data archive is associated with a specific program or has a thematic orientation and will actively seek selected data sets. The data archive provides guidelines to the contributor for formatting and submitting data and metadata to the archive. The contributor prepares data and metadata as completely as possible and submits them following archive guidelines.

Archive staff review the data and metadata and may reformat them to achieve consistency and completeness, making sure that the metadata supply citation information so that the original data contributors can be correctly acknowledged. Staff also review the quality assurance that was performed on the data as documented in the metadata. They may also select keywords based on the metadata to be used in search and order functions. Archive staff and contributors work together to resolve questions and review changes. After consensus is reached, metadata and data are entered into the archive for public access and long-term storage.

In addition to maintaining a long-term, secure data archive, the archive staff also provide post-project support, such as answering user questions, informing users of updates and additions, and maintaining user statistics. Archives must also plan for the periodic upgrading of storage media. Archives can perpetuate the growth and value of their data holdings by including a strategy for incorporating data updates, value-added products (especially from synthesis and modeling applications), and user feedback. Staff can collaborate with scientists to determine useful enhancements to data sets (e.g., add common variables, aggregate to common units, or calculate uncertainty) on the basis of user needs. It is also crucial for the staff to promote the availability of the data by interacting with the user community, through attendance at professional meetings and use of Web marketing techniques.

Table 2. Data Archives at ORNL.

Data Archive / Web Address	Focus	Special Features
Carbon Dioxide Information Analysis Center (CDIAC) *http://cdiac.esd.ornl.gov*	Acquire, compile, quality-assure, document, archive, and distribute information on greenhouse gases and climate change in support of the US Department of Energy's (DOE) Global Change Research Program. Since 1992, CDIAC has hosted a component of the World Data Center-A for Atmospheric Trace Gases of the International Council of Scientific Unions to store and manage data on radiatively active trace gases and their concentrations.	Special emphasis on quality assurance, documentation, and derived, integrated products. User community includes many thousands of researchers, policymakers, educators, students, and corporate officials around the world. User services office.
Atmospheric Radiation Measurement (ARM) Program Archive *http://www.archive.arm.gov*	Improve radiative transfer functions in General Circulation Models (GCMs) and the parameterization of cloud properties and formation in GCMs as part of the DOE's ARM Program.	Stores massive amounts of data: >2 million files (>1500 GB) and 70,000 new files (70 GB) per month. Provide large volumes of data to scientists: ~22,000 files (20 GB) per month, ~400 registered users.
Distributed Active Archive Center for Biogeochemical Dynamics (ORNL DAAC) *http://www~eosdis.ornl.gov*	Acquire, quality-assure, document, and archive multidisciplinary data for terrestrial ecosystems and provide access to the global change research community, policymakers, and educators, as part of the National Aeronautics and Space Administration Earth Observing System.	Web-based data search and order interface, browse/view data before ordering, multiple distribution media, free and ready access, user services office.
Oak Ridge Environmental Information System (OREIS) *http://www~oreis.cad.ornl.gov:8080/oreis/help/oreishome.html*	Develop a consolidated database to support environmental cleanup activities on the Oak Ridge Reservation for DOE.	Full relational database management system with links to statistical and geographic information system (GIS) tools, 5 million records added since 1994 (password-protected).

FUTURE DIRECTIONS FOR DATA ARCHIVAL

Ecological synthesis and assessment studies that address long-term regional and global ecological issues will continue to expand and use data from data archives. Sharing and archiving data can be more efficient if the following general principles are considered in the overall project planning and operations: (1) establish the flow of data from investigator to a long-term data archive as part of the work plan; (2) process data to achieve consistency and completeness of data and metadata; and (3) institute policies to give data producers adequate credit for their data archival efforts.

To more fully share data, we suggest the scientific research community embrace the following actions:

- provide incentives for sharing and archiving data,
- recognize data sets with metadata as valuable research products,
- establish a universal citation policy for data,
- establish guidelines for metadata,
- develop data distribution and archive centers, and
- ensure long-term financial and institutional support.

As ecologists, we have an opportunity to educate and lobby science administrators, program managers, and agencies about the data archival process, its intrinsic value, and required resources.

ACKNOWLEDGEMENTS

We would like to thank S. W. Christensen and B. T. Rhyne for their helpful reviews of this manuscript. This work was sponsored by the National Aeronautics and Space Administration (Distributive Active Archive Center project) under Interagency Agreement DOE No. 2013-F044-A1 and by the US Department of Energy (Atmospheric Radiation Measurement Archive project). Oak Ridge National Laboratory is operated by Lockheed Martin Energy Research Corp. under contract DE-AC05-96OR22464 with the US Department of Energy.

LITERATURE CITED

Michener, W. K., J. W. Brunt, J. Helly, T. B. Kirchner, and S. G. Stafford. 1997. Non-geospatial metadata for the ecological sciences. Ecological Applications 7(1): 330-342.

Porter, J. H. and J. T. Callahan. 1994. Circumventing a dilemma: historical approaches to data sharing in ecological research. Pages 193-202 in W. K. Michener, J. W. Brunt and S. G. Stafford editors. Environmental Information Management and Analysis: Ecosystem to global scales. Taylor & Francis, Bristol, PA.