ECOLOGICAL METADATA

William K. Michener

Joseph W. Jones Ecological Research Center, Route 2, Box 2324, Newton, GA 31770

Abstract. Metadata represent comprehensive documentation of the content, context, quality, structure and accessibility of a data set. In this chapter, relevant geospatial and non-geospatial metadata "standards" are reviewed, World Wide Web sources of information pertaining to metadata are identified, a metadata implementation "recipe for success" is proposed, and remaining challenges are discussed.

INTRODUCTION

Metadata are the information necessary to understand and effectively use data, and include documentation of the data set contents, context, quality, structure, and accessibility. From the perspective of the data originator, metadata are necessary to support further processing and analysis. When a scientist's goal is to re-use data collected by others, comprehensive metadata may be essential to support identification and acquisition of suitable data, as well as to facilitate additional processing and analysis.

Metadata are receiving increased attention by the scientific community. For example, ecologists, scientific societies, and state and federal agencies are recognizing the importance of high quality, well-documented, and securely archived data for addressing long-term and broad-scale questions (e.g., Gross et al. 1995). In addition, ecological data, such as those collected by individual and teams of scientists at field stations, marine laboratories, natural areas, parks, and preserves, represent a significant national resource that are essential for understanding and monitoring the health of the dynamically changing environment. Comprehensive metadata are required to counteract the natural tendency for data to degrade in information content ("data entropy" *sensu* Michener et al. 1997) through time.

PROGRESS IN METADATA STANDARDIZATION

Although all ecological data have a "spatial" element (i.e., are collected at one or more points in space), ecological data may be generally categorized as being either geospatial or non-geospatial. Geospatial data include those types of data that are explicitly associated with a geographical location. Examples include remotely sensed imagery, geographic information system (GIS) data layers, data derived from broad-scale sampling efforts (e.g., National Weather Service, National Atmospheric Deposition Program), as well as fine-scale sampling of spatially explicit patterns and processes. Non-geospatial data, on the other hand, and for the purposes of this paper, might include data from laboratory and micro- to mesocosm experiments, as well as other ecological data that are collected at a finite number of points. For these cases, precise geographic coordinates of the sampling sites are relatively unimportant and often unrecorded. Most metadata standardization efforts have, thus far, focused on geospatial data.

Geospatial metadata standardization

As part of the ongoing evolution of the National Biological Information Infrastructure (NBII) and standardization of geographical data in the Federal government, significant attention has focused on standardizing geospatial metadata. One of the most significant products to emerge has

been a document entitled "Content Standards for Digital Geospatial Metadata" (Federal Geographic Data Committee 1994) which contains a comprehensive list of geospatial metadata descriptors. Seven categories of metadata descriptors are included in the document: (1) identification; (2) data quality; (3) spatial data organization; (4) spatial reference; (5) entity and attribute; (6) distribution; and (7) metadata. Additional efforts are underway to add extensions to the Content Standards that are relevant to vegetation classification data, as well as cultural, demographic, and other types of geographical data. Additional information on Federal Geographic Data Committee activities, metadata generation tools (e.g., NBII MetaMaker), and related material can be found at the World Wide Web sites listed in Table 1.

Table 1. Geospatial metadata World Wide Web sites.

"FGDC Metadata FAQ" *http:07/28/98/geochange.er.usgs.gov/pub/tools/metadata07/28/98/tools/doc/faq.html*
"National Biological Information Infrastructure (NBII)" *http:07/28/98/www.its.nbs.gov/nbii/index.html*
"NBII MetaMaker Version 2.10" *http:07/28/98/biology.usgs.gov/nbii/metamaker/metamaker.html*
"Metadata Tools" *http:07/28/98/badger.state.wi.us/agencies/wlib/sco/metatool/mtools.html*
"Metadata Tool Evaluation" *http:07/28/98/www.fgdc.gov:80/metadata/mitre/task2/index.html*

Generic ecological metadata descriptors (non-geospatial)

Ecological studies often require the collection of an extremely diverse array of data including attributes that characterize and quantify the chemical and physical environment, organism physiology, population and ecosystem dynamics, community composition, landscape structure, as well as anthropogenic influences. It is unlikely that a single metadata standard, no matter how comprehensive, could encompass all types of ecological data because of this complexity. Consequently, a generic set of non-geospatial metadata descriptors were recently proposed for the ecological sciences (Michener et al. 1997). The list of metadata descriptors was suggested as a template that could serve as the basis for more refined subdiscipline- or project-specific metadata guidelines. Five categories of metadata descriptors were delineated: (1) data set descriptors; (2) research origin descriptors; (3) data set status and accessibility; (4) data structural descriptors; and (5) supplemental descriptors (Table 2).

IMPLEMENTATION STRATEGIES

Metadata may be recorded in a variety of forms ranging from free-flowing text to incorporation into a structured database management system (DBMS). Some of the most important metadata attributes are often recorded in the field using pencil and paper. "Natural history" observations are frequently critical for correct interpretation and analysis of field data. Field notes and other metadata can later be maintained in paper files or incorporated into word processing files, SAS programs, DBMS programs, or World Wide Web-accessible documents. The choice of metadata media is often dictated by availability of software, trained personnel, and time. Guidelines for

metadata structure and supporting technology (WWW forms, etc.) are currently being discussed and developed at the San Diego Supercomputer Center, National Center for Ecological Analysis

Table 2. Generic non-geospatial metadata descriptors for ecological research (adapted from Michener et al. 1997).

I. Data Set Descriptors
- **A. Data set identity**
- **B. Data set identification code**
- **C. Data set description**
  1. originator(s)
  2. abstract
- **D. Keywords**

II. Research Origin Descriptors
- **A. "Overall" project description**
  1. identity
  2. originator(s)
  3. period of study
  4. objectives
  5. abstract
  6. source(s) of funding
- **B. "Specific sub-project" description**
  1. site description
  2. experimental or sampling design
  3. research methods
  4. project personnel

III. Data Set Status and Accessibility
- **A. Status**
  1. latest update
  2. latest archive date
  3. metadata status
  4. entry verification
- **B. Accessibility**
  1. storage location and medium
  2. contact person(s)
  3. copyright restrictions
  4. proprietary restrictions
  5. costs

IV. Data Structural Descriptors
- **A. Data set file**
  1. identity
  2. size
  3. format and storage mode
  4. header information
  5. alphanumeric attributes
  6. special characters/fields
  7. authentication procedures
- **B. Variable information**
  1. variable identity
  2. variable definition
  3. units of measurement
  4. data type
  5. data format
- **C. Data anomalies**

V. Supplemental Descriptors
- **A. Data entry**
  1. data forms used
  2. location of completed data forms
  3. verification procedures
- **B. QA/QC procedures**
- **C. Related materials**
- **D. Computer programs and data processing algorithms**
- **E. Archival**
- **F. Publications**
- **G. History of data set usage**

and Synthesis, the Long-Term Ecological Research Network, as well as numerous other organizations. Regardless of the availability of tools that can facilitate metadata entry, storage, and retrieval, there are several non-technological activities that can be performed at the level of the individual investigator, field station, project, or "group" to facilitate successful metadata implementation.

Metadata: a recipe for success

The first and probably most important component of metadata implementation is to perform a site or project needs assessment. Such an assessment entails identifying data objectives (e.g., projected or desired data longevity, potential for re-use, value), establishing guidelines and procedures for data sharing and data ownership, assessing infrastructure (e.g., availability of hardware, software, people, funds), and categorizing and prioritizing metadata activities. For example, at a field station, meteorological data may receive a high priority for metadata implementation because of their perceived value to a large number of ongoing studies, historical usage patterns, and potential for repeated use over time. In contrast, infrequent field surveys performed as part of an undergraduate research project may receive a lower priority for archival and metadata implementation. Once categories of data are prioritized, it is necessary to either adopt an existing metadata standard (e.g., geospatial metadata standard (FGDC 1994)) or identify a set of minimal and optimal metadata descriptors that meet perceived needs.

The second recommended step in metadata implementation is to perform a pilot project using one to three relatively "simple" data sets. Based upon successes and difficulties encountered in the pilot project, it is useful to re-evaluate site needs and objectives. For example, a formal or informal cost-benefit analysis may facilitate future prioritization and balance completeness of metadata versus funding and personnel availability. Following this evaluation process it is necessary to formalize metadata activities. It may be desirable, for example, to develop relevant policies and procedures, identify available metadata tools or initiate programming efforts to develop appropriate tools, and establish a reward structure for providing comprehensive metadata. Metadata and other data management activities should be re-evaluated on a periodic basis to insure that they are meeting specified objectives. Several simple "rules of thumb" may facilitate successful implementation:

- Keep it simple! Start small and build upon successes. For example, the time and effort expended on a pilot project are usually paid back several-fold in the long run.
- Build consensus among scientists and data managers from the start. Data management initiatives, regardless of their potential benefits, are often unsuccessful when the "user community" is excluded from the process. Data management must be fully integrated into the research planning process and involve the scientific community it serves (Stafford et al. 1994).
- Data longevity is roughly proportional to metadata comprehensiveness. However, establishing a goal of complete metadata that can meet all future needs may be exorbitantly expensive and, ultimately, unattainable.
- Data and metadata should ideally be platform-independent. Hardware and software change frequently. Today's "standard" may be gone tomorrow. Thus, it pays to avoid proprietary storage formats whenever possible.
- The degree to which high-quality ecological data and accompanying metadata are securely archived and accessible for future research is directly related to the extent to which an ethic of data stewardship is promoted and rewarded (Porter and Callahan 1994).

FUTURE CHALLENGES AND OPPORTUNITIES

Flexible metadata tools that support entry, search, and retrieval are essential for facilitating metadata implementation. There is a significant need for research and development in this area. Many of the scientific benefits that are associated with the availability of high-quality data and metadata have been discussed here and elsewhere (Gross et al. 1995, Michener et al. 1997).

Although future research endeavors will inevitably pay more attention to metadata and other aspects of data management, the clock is running out on many extremely valuable long-term and unique ecological data sets. There is a significant need for established funding mechanisms and data archives to support metadata development and secure long-term storage of these irreplaceable data (Gross et al. 1995). Development of attendant reward systems (e.g., peer-reviewed data and metadata publications, equating database construction with publication efforts) will be essential for further promoting an ethic of data stewardship (Porter and Callahan 1994).

Much future discussion will likely focus on standardization issues. If ecological metadata are or should be standardized, then who decides on the standard? Should standardization occur at the level of the institution (e.g., field station, university), society (e.g., Ecological Society of America), discipline (e.g., litter decomposition), funding agency (e.g., NSF), or globe (e.g., International Long-Term Ecological Research Network)? What constitutes minimal and optimal criteria and standards? Like other standardization efforts, the true test of any emerging metadata standard will ultimately rest on whether the standard is simple to use and easily understood, and whether or not it makes our science better.

Finally, it should be reiterated that there are costs associated with metadata implementation, data archival, and other data management activities. Personnel costs associated with developing metadata can, in some cases, exceed data collection efforts. Issues related to long-term curation and maintenance of data and metadata cannot be dealt with effectively in most 1-, 2-, or 3-yr grant cycles. Devoting resources during a short-term project to data management (e.g., metadata) costs money and personnel effort and can result in fewer short-term publications. On the other hand, when high quality data and metadata are securely archived, they can be "mined" for many years or decades into the future. Proper balance of short-term costs versus long-term gain is an issue that warrants continued thought and discussion.

ACKNOWLEDGMENTS

This paper benefited from discussions with James Brunt, John Helly, Thomas Kirchner, and Susan G. Stafford.

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*).

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 7:330-342.

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.

Stafford, S. G., J. W. Brunt and W. K. Michener. 1994. Integration of scientific information management and environmental research. Pages 3-20 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.