Ecological Applications, 7(1), 1997, pp. 330–342

NONGEOSPATIAL METADATA FOR THE ECOLOGICAL SCIENCES

WILLIAM K. MICHENER,[1] JAMES W. BRUNT,[2] JOHN J. HELLY,[3] THOMAS B. KIRCHNER,[4] AND SUSAN G. STAFFORD[5]

[1]*Joseph W. Jones Ecological Research Center, Route 2, Box 2324, Newton, Georgia 31770 USA, and University of Georgia, School of Ecology, Athens, Georgia 30602 USA*
[2]*Sevilleta Long-Term Ecological Research Program, University of New Mexico, Department of Biology, Albuquerque, New Mexico 87131-1091 USA*
[3]*San Diego Supercomputer Center, P.O. Box 85608, San Diego, California 92186-9784 USA*
[4]*Colorado State University, Natural Resources Ecology Laboratory and Department of Range Science, Fort Collins, Colorado 80523 USA*
[5]*Department of Forest Science, Oregon State University, Corvallis, Oregon 97331-7501 USA*

Abstract. Issues related to data preservation and sharing are receiving increased attention from scientific societies, funding agencies, and the broad scientific community. Ecologists, for example, are increasingly using data collected by other scientists to address questions at broader spatial, temporal, and thematic scales (e.g., global change, biodiversity, sustainability). No data set is perfect and self-explanatory. Ecologists must, therefore, rely upon a set of instructions or documentation to acquire a specific data set, determine its suitability for meeting specific research objectives, and accurately interpret results from subsequent processing, analysis, and modeling.

''Metadata'' represent the set of instructions or documentation that describe the content, context, quality, structure, and accessibility of a data set. Although geospatial metadata standards have been developed and widely endorsed by the geographical science community, such standards do not yet exist for the ecological sciences. In this paper, we examine potential benefits and costs associated with developing and implementing metadata for nongeospatial ecological data. We present a set of generic metadata descriptors that could serve as the basis for a ''metadata standard'' for nongeospatial ecological data. Alternative strategies for metadata implementation that meet differing organizational or investigator-specific objectives are presented. Finally, we conclude with several recommendations related to future development and implementation of ecological metadata.

Key words: data archive; data lineage; data management; information science; metadata; quality assurance.

INTRODUCTION

Historically, ecological data have been collected primarily by single or small groups of investigators in plots of ≤ 1 m^2 over relatively short periods of time (Kareiva and Anderson 1988, Brown and Roughgarden 1990). Increased societal and scientific interest in issues such as global change, biodiversity, and sustainability are, however, causing ecologists to question how ecological patterns and processes vary in time and space, and to understand the causes and consequences of this variability (Levin 1992). Many such questions require far more data than could feasibly be collected, managed, and analyzed under the auspices of a single investigator or project (group of investigators). Consequently, ecologists are increasingly using data that have been collected by other scientists from numerous disciplines including ecology, often for different purposes. In addition to the scientific ''question-driven'' impetus, funding agency mandates (e.g., National Science Foundation 1994) and the interests of state and federal agencies (e.g., National Biological Survey; National Research Council 1993) and scientific societies (e.g., Ecological Society of America; Colwell 1995) have focused increased attention on preserving, sharing, and promoting the understanding of valuable data sets.

Because of the interdisciplinary nature of the science, ecologists are generally accustomed to freely sharing data with expert colleagues in order to address specific questions. Unfortunately, few, if any, ecological data sets are perfect or intuitive. Thus, even in cases where recently collected data are shared with a colleague who is associated with the same institution or project and who is reasonably familiar with the research and resulting data set(s), a modest set of instructions is necessary to effectively use the data and accurately interpret the results. Highly detailed instructions or documentation may be required for scientists to accurately interpret and analyze historic or long-term data sets, as well as data resulting from unfamiliar research or complicated experimental designs. Numerous discussions of data management issues associated with global change, biodiversity, and sustainability have

Manuscript received 18 October 1995; accepted 28 February 1996.

highlighted a need for accepted protocols to assist scientists with preserving important data sets and providing guidelines for the supporting documentation that is necessary to interpret the data (National Research Council 1991, 1993, 1995*a*, *b*, Gosz 1994).

Metadata, i.e., data documentation, may be defined as representing the higher level information or instructions that describe the content, context, quality, structure, and accessibility of a specific data set. Ideally, metadata comprise all information that is necessary and sufficient to enable long-term secondary use (reuse) of the data set by the original investigator(s), as well as use by other scientists who were not directly involved in the original research efforts. Thus, objectives for metadata implementation include facilitating: (1) identification and acquisition of data for a specific theme, time period, and geographical location; (2) determination of the suitability of data for meeting a specific objective; and (3) data processing, analysis, and modeling.

Significant progress has been made during the past decade in developing metadata standards for geospatial data. For example, numerous spatial data transfer standards incorporate a metadata component (Digital Geographic Information Working Group 1991, Defense Mapping Agency 1992, National Institute of Standards and Technology 1992). More recently, a comprehensive set of Content Standards for Digital Geospatial Metadata has been released that defines standard geospatial metadata descriptors related to data availability and accessibility, determination of fitness for use, and processing and utilizing a set of data (Federal Geographic Data Committee 1994).

The important role of metadata in facilitating ecological research has been recognized since the 1980s (Kellogg Biological Station 1982, Michener et al. 1987, Kirchner et al. 1995), and several practical approaches to metadata management have been presented (Stafford et al. 1986, Conley and Brunt 1991, Brunt 1994). However, metadata standards for nongeospatial ecological data currently do not exist in any standard format beyond individual studies and experiments. Objectives of this paper are to: (1) examine potential benefits and costs associated with developing and implementing metadata for nongeospatial ecological data; (2) propose a set of generic metadata descriptors that could serve as the basis for a ''metadata standard'' for the ecological sciences; and (3) present alternative strategies for metadata implementation that meet differing organizational or investigator-specific objectives. Finally, we conclude with several recommendations related to future development and implementation of ecological metadata.

Benefits and Costs Associated with Metadata Implementation

Scientists often refer to the rows and columns of numeric or encoded observations as raw data. Raw data are useful only when they can be framed within a theoretical or conceptual model. Relating raw data to underlying theoretical or conceptual models requires understanding the types of variables that were measured, measurement units, potential biases in the measurements, sampling methodology, and other pertinent facts not represented in the raw data, i.e., the metadata. The combination of raw data and metadata within a conceptual framework produces information.

Information can be lost through degradation of the raw data or the metadata. Such loss is unavoidable. Technological advances can make data collected in earlier times obsolete. Automated data collection procedures can now overwhelm our ability to effectively store, retrieve, manage, and analyze data (Stafford et al. 1994), which has sometimes necessitated the implementation of procedures to purposely discard some data. It is the premature loss of useful data, such as long temporal sequences or irreplaceable data, that is a major scientific concern. The preservation of metadata is particularly problematic because metadata encompass a diverse and variable collection of facts that are often not recorded in any systematic way, including some facts that may reside only within the minds of the researchers.

Many processes can lead to the loss of information through time (Fig. 1). Some of these processes operate continuously, such as the gradual degradation of storage media containing the data, whereas others can be categorized as discrete events, such as the retirement or death of the scientist who collected the data, obsolescence of storage technology, or the loss of storage media through catastrophic events. Although loss of metadata can occur throughout the period of data collection, the rate of loss is likely to increase after project results have been published or the study has been terminated. Specific details are likely to be lost first, due to abandonment of data forms and notes in lieu of digitally preserved data and to loss from the memory of the investigator. Over longer time periods, degradation of storage media and further memory losses can reduce the information about general details not covered in relevant publications. Retirement or other major career changes may lead to the physical loss of records and hamper access to the investigator's recollections regarding data. Bowser (1986), for example, documents many of the problems associated with archiving, metadata, and quality assurance that were encountered during attempts to reanalyze data collected from 1926 to 1941 by E. A. Birge, C. Juday, and collaborators in Wisconsin lakes.

Ecologists also lose information through the loss of conceptual models used to help interpret the data. Such models are often simple and can be expressed using statistical models to represent relationships among variables. However, some data sets, particularly long time series, are associated with hypotheses involving complicated nonlinear relationships that are best rep-

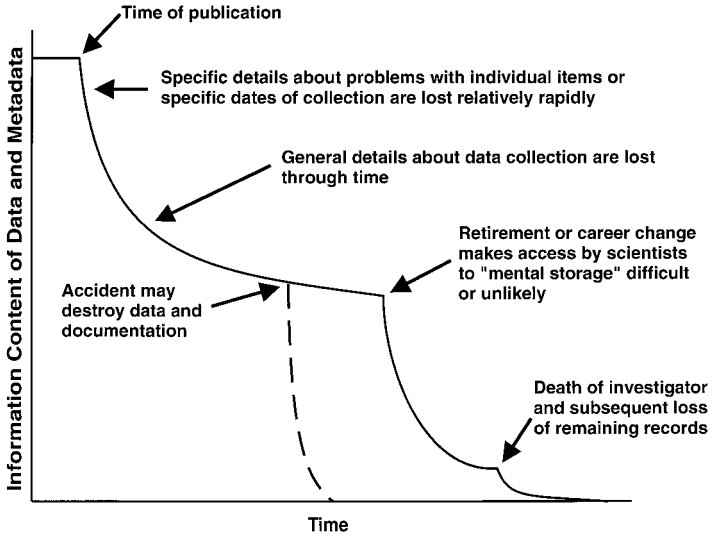

FIG. 1. Example of the normal degradation in information content associated with data and metadata over time (''information entropy''). Accidents or changes in storage technology (dashed line) may eliminate access to remaining raw data and metadata at any time.

resented by complex simulation models. Thus, preservation of the information about a set of data may also involve preservation of the simulation model and its associated input and output files (Kirchner 1994). Peer-reviewed publications featuring simulation models tend to focus on the results and the conceptual and mathematical foundations for the model. Because simulation models tend to be modified through time, preservation of the model code and input files is likely to be critical if model experiments are ever to be truly reproducible.

Both benefits and costs accrue during the development, implementation, and maintenance of metadata. In the following discussion, we present some of the benefits and costs associated with metadata implementation. An example from the International Biological Program (IBP) illustrates many of the difficulties encountered in attempts to reinvigorate extant data sets, and highlights the importance of well-conceived and adequately maintained metadata.

Benefits

The most important reason to invest time and energy in developing metadata is that *human memory is short.* If data are to undergo any secondary usage, then adequate metadata will be required even if that secondary usage consists of reuse by the data originator. Scientists have long recognized the importance of preserving information, but have often focused only on preserving the results of their synthetic activities through publication. Publication typically preserves some of the metadata, but often only a subjectively selected portion of the metadata needed to relate the data to a specific hypothesis. To aggravate this scenario, ecological data sets are often extremely complex. Missing values, midcourse modification of sampling or laboratory procedures, addition or deletion of study parameters, personnel turnover, plot or habitat modification by disturbances (natural and anthropogenic) or changing environmental conditions, and numerous other factors leading to data anomalies are commonplace. Adequate documentation (metadata) of sampling and analytical procedures, data anomalies, and data set structure will help to insure that data can be correctly interpreted or reinterpreted at a later date. Twenty years is often established as the objective for having data usable by scientists unfamiliar with the data and their collection (''the 20-yr test''; Webster 1991, Strebel et al. 1994).

In addition to the limitations of human memory, significant changes in the scope of ecology further underscore the critical role of metadata in supporting science. For example, the life-span of a typical ecological data set that was collected 10 yr ago may have been very short, lasting from data set conception to publication, roughly corresponding to the average funding cycle of two to three years. At best, many such data sets met their resting place as dusty file folders of poorly documented data relegated to the bottom drawer of a filing cabinet. History and personal experiences are ripe with examples in which data became useless because relevant metadata were missing or unavailable (National Research Council 1995*a*). More recently, however, increased interest in long-term ecological research (Franklin et al. 1990), comparative studies (Pace 1993), and expansion of the spatial, temporal, and thematic scales of basic and applied ecological studies (Levin 1992) have resulted in data sets being used for multiple purposes, often repeatedly over long periods of time.

Metadata provide the information that is critical for expanding the scales at which ecologists work. Comparative studies including temporal comparisons among sites, statistical replication, and comparisons within and among sites all depend upon the availability of sufficient metadata. For example, calibration and intercalibration (measurements of similar parameters by different methods or instruments) of methods and

instruments should be well-documented in order to confirm data integrity and proper use of experimental methods and data acquisition. Similarly, ground-based reference data from multiple sites are frequently used to calibrate or support analyses of remotely sensed data, thereby expanding the spatial domain from the site to the landscape, region, or globe. Metadata are critical for combining physical, chemical, and biological data sets that contain different parameters but share common spatial or temporal domains. Many short-term studies serendipitously evolve into long-term studies. In some cases, relatively short- to medium-term time series data (possibly from different investigators or research programs) are integrated into a single long-term record. Metadata are essential for maintaining a historical record of long-term data sets that have resulted from such integration efforts, as well as documenting changes in personnel, methods, and data anomalies in ongoing long-term projects. Over the course of a long-term project, field and laboratory equipment are frequently replaced by other instruments that offer better precision or improved data acquisition methods (e.g., remote data loggers, etc.). Bowser (1986) and Strayer et al. (1986) discuss the importance of method intercalibration, quality assurance, and metadata for supporting reliable long-term data sets.

Synthesis and modeling projects are often hindered by the lack of high quality data and metadata. For example, ecological modelers routinely extract parameters from publications. Frequently, publications do not provide sufficient information pertaining to the data distribution, requiring many assumptions by the modeler about data ranges, frequency distributions, percentiles, etc. Ideally, raw data would be available for the modeling project, as well as the metadata that are critical for describing data collection objectives and methods, scale relevance of the data, and other potential limitations for secondary usage. For example, data collected under the auspices of IBP at the Andrews Experimental Forest in Oregon's Cascade Range during the 1970s were published as data summaries in internal technical reports (Emmingham and Lundburg 1977). Currently, entire Long-Term Ecological Research (LTER) data sets collected at the Andrews Experimental Forest are accessible on–line via the World Wide Web.

Costs

High costs, primarily in personnel time, can be associated with the initial development of metadata. For short-term projects, the metadata file size and level of effort expended in developing metadata may exceed the physical size of the raw-data file and the efforts expended in data collection. Real costs are associated with editing data and metadata and making them available to the scientific community in hard-copy or electronic formats. Research grants and other existing funding mechanisms are often insufficient to support development of a comprehensive set of metadata.

Stewardship and a continuing need for curation and maintenance of the data and metadata represent real cost burdens that are not often factored into project budgets. After a study is completed, who bears the responsibility of informing the user community of changes to the data set and newly discovered anomalies? Furthermore, critical details can often be overlooked in even the most comprehensive metadata. However, the role and appropriateness of funding for data originators as metadata consultants have received little attention and should be considered.

Data and metadata entropy through time: an example

Consider the history of data collected for the IBP Grassland Biome. Data were collected by investigators at several sites within the biome and sent to the Natural Resources Ecology Laboratory (NREL) at Colorado State University (CSU) for processing and management. Most of the data were recorded on standardized paper forms that were color coded for convenience of the investigators. Data were transcribed from the forms to punched cards by professional keypunch operators. Much of the data was then transferred to seven-track magnetic tapes for storage. Metadata were distributed among technical manuals (which specified methods for data collection), peer-reviewed publications, and the data form (which specified items such as units and species codes and included room for comments that were typically not keypunched when the forms were transcribed). The cards, tapes, and many data forms were preserved by the NREL after the end of the IBP project, but without the benefit of active maintenance, such as the periodic replacement of magnetic tapes. When the CSU computer center made the transition from seven-track to nine-track tapes, scientists obtained funding from the National Science Foundation to transfer as much data to the new tapes as was feasible without extensive effort. Those tapes that could not be read were abandoned, and an attempt was made to recover the data from the card decks. However, the combination of old card decks and antiquated card readers meant that some data could not be recovered. In addition, some data sets were stored in machine-specific, packed binary files that could no longer be decoded. Card decks were discarded after the CSU computer center abandoned card readers.

At the start of the Central Plains Experimental Range LTER project, an attempt was made to recover data stored on the nine-track tapes and to preserve the IBP data forms using microfiche. Budgetary limitations restricted preservation to those data specific to the Pawnee Site, although some data from other sites were also preserved if convenient or of immediate interest to an LTER scientist. Some of the tapes had degraded to a point at which they could no longer be read. Numerous

problems were encountered with the microfiche process, including the fading and bleeding of inks on the forms and difficulty in getting clear photographs from the darkly colored forms. Although attempts were made to assemble metadata, those for any specific data set were often incomplete or absent, since publication histories were not linked to the data sets, and many of the original investigators were no longer able to provide requisite information. Thus, although most of the raw data associated with the CPER site were successfully preserved after considerable effort, it is also true that critical metadata were lost, thereby reducing the value of many of the historic data sets.

Standard Ecological Metadata Descriptors

The previous discussion highlighted the importance of metadata for providing scientists with the information necessary to reuse previously collected data. The IBP example further documented many problems associated with preserving data and metadata over a single decade. Assuming that some ecological data sets are inherently valuable and that we desire to preserve them and provide the relevant metadata required for sound secondary usage, we are still left with the problem of determining what metadata are essential. Fortunately, the process that scientists might normally follow in acquiring and utilizing existing data sets provides a guide to what metadata may be required. For example, once a scientist recognizes the need for specific data, several questions (or steps) must be addressed in an orderly fashion: (1) What data sets exist that might meet specified objectives? (2) Why were those data sets collected, and are they ''fit'' for my particular use? (3) Can these data sets be obtained? If so, how? (4) How are the data organized and structured? (5) Is there additional information that would facilitate my use and interpretation of the data?

The five steps hypothetically followed during identification, acquisition, and utilization of data serve as the basis for the five classes of metadata descriptors listed in Table 1 (based on Michener et al. 1987, 1990, Federal Geographic Data Committee 1994, Kirchner et al. 1995). [Note that the term ''data set'' used in this discussion is synonymous with ''data table,'' which frequently appears in computer science literature.] Class I includes basic attributes of the data set (data set title, associated scientists, abstract, and key words) that are frequently included in hard-copy and electronic data set catalogs (e.g., Michener et al. 1990). The purpose of Class I descriptors is to alert potential secondary users to the existence of data sets that fall within specific temporal, spatial, and thematic domains. Preliminary determination of fitness-for-use by secondary users can be facilitated by incorporating in the abstract a brief discussion of the scientific context and a description of potential uses of the data set. In many cases, a short summary of the ''data set usage history'' (Table 1: V.G.; especially the data request history and questions and comments from secondary users) could be used to identify potential uses of the data and to highlight major strengths and weaknesses.

Class II includes all relevant metadata that describe the research leading to development of a particular data set. Two subcategories of research origin descriptors are essential. The first subcategory includes a description of the broader, more comprehensive project (e.g., LTER program at a specific site) that may have led to numerous spin-off projects from which individual data sets emerged (e.g., climate, primary production, decomposition, etc.). The purpose of the ''overall'' project description is to provide the broader scientific context for an individual study. If an individual data set emerged from a stand-alone project, then the ''overall'' project description is superfluous. The second subcategory includes all pertinent information related to the research origins of a specific data set. Site characteristics, experimental or sampling designs, research methods, and project personnel are described in detail. Two additional descriptors may be essential for some data sets. First, permits are required for research and collecting on public lands within the United States and for importation of specimens into the country. Thus, permit history may be especially critical for museums that archive physical specimens, especially if museum personnel were not involved in the research and do not have the permit records. Second, many environmental monitoring and compliance data sets are generated in response to legal and organizational requirements. In such cases, it is important to document relevant laws, decision criteria, compliance standards, and other factors that may affect study design and data collection (Eagan and Ventura 1993).

Class III metadata describe the status of the data set and associated metadata, as well as information related to data set accessibility. Data set accessibility is affected by numerous factors that should be fully documented in the metadata. In some cases, copyright or other legal restrictions (e.g., state or federal laws restricting access to maps of endangered species locations, etc.) apply to specific data sets. In other cases, various proprietary restrictions may apply. For example, research laboratories, universities, and funding agencies frequently require appropriate citation of the grant that funded the research or the institution or site where the research was performed.

Class IV metadata describe all attributes related to the structure of the data file. All variables should be labeled, defined, and characterized as to data type and format. Finally, all known data anomalies (e.g., missing data, etc.) are fully documented.

Class V metadata include all other related information that may be necessary for facilitating secondary usage, publishing the data set, or supporting an audit of the data set. In some cases, for example, a scientist may find it necessary to review raw data forms, quality assurance/quality control procedures, computer pro-

grams and algorithms, and publications resulting from the data set. In addition, it may prove necessary to examine existing field notebooks or physical specimens. Pertinent data for physical specimens may include references to accession records/numbers (e.g., the transfer of a group of voucher specimens to a museum), specimen numbers assigned by the collector or the collection, and linkages among different forms of physical vouchers (e.g., sound recordings, chemical analyses, etc.) with different parts of physical specimens. Archived maps and photographs may facilitate resampling of a specific site.

Metadata may also serve as a vehicle for user feedback and reporting data anomalies. A ''data set usage history'' (Table 1: V.G.) may greatly facilitate long-term utilization of important data sets. There is no unique minimal and sufficient set of metadata for any given data set, since sufficiency depends on the use(s) to which the data are put. Because uses may vary, problems with data and metadata should be recorded and retained as a usage history, analogous to attaching ''post-it notes'' to the data to alert subsequent users to idiosyncracies within the metadata or anomalies within the data. However, direct modifications of the data should only be made by the data set owner/originator.

Metadata Implementation

A primary objective of metadata development and implementation is to facilitate data reuse by the data originator as well as to support research activities by other scientists (Briggs and Su 1994). Fig. 2 illustrates how metadata may evolve during the course of a specific project and how secondary users (a modeler, in this example) might interact with the data and associated set of metadata. Specifically, hypotheses/questions and generic descriptions of the experimental or sampling design may be incorporated into the data set description (I.). Other more detailed aspects of the project design, including field and laboratory procedures, would be included in the ''specific subproject'' description (II.B.). Information related to data collection, data entry, and quality assurance/quality control (QA/QC) is relevant to data set status and accessibility (III.), data structural descriptors (IV.), and supplemental descriptors (V.). Descriptors related to the ''complete'' digital data set are relevant to all metadata classes except perhaps those related to the research origin (II.). Finally, information obtained during analysis, synthesis, modeling, and publication may be incorporated into the supplemental descriptors (V.) in order to facilitate secondary utilization.

Subsequent to completion of the original project, a modeler may become aware of the existence of a particular data set via perusal of an organization's metadata database or data catalog, which may contain ''generic'' data set descriptors (Fig. 2: I.). The decision to acquire and understand a particular data set as well as the mechanisms for doing so would require that the modeler have access to all relevant metadata that describe the origin of the data set (II.), its status and accessibility (III.), structure of the data set (IV.), and possibly other miscellaneous details (V.). During the course of model execution, hypothesis testing, and model validation, new information about the data set that could benefit other secondary users may come to light. This information, plus a listing of publications resulting from secondary use of a data set, would ideally be incorporated into the data set usage history (V.G.) to facilitate additional secondary usage.

The development and maintenance of metadata can be a very costly endeavor. Thus, it may be important to attempt to match the level of metadata *content* and *format* to the needs of anticipated users. As an example, we have identified three levels or types of secondary data utilization (Table 2; see also Kellogg Biological Station 1982). This categorization is derived from the identification of at least three types of data reusers and from the recognition that the metadata content must increase at each level. The first, a Level I data reuser, may be a colleague or collaborator with technical expertise in the subject area and adequate knowledge of data collection, analytical, and processing procedures. Thus, such an individual may require only a basic description of the data set, as well as more detailed data structural descriptors, in order to effectively use the data set. A Level II data reuser may be someone who is searching a metadata catalogue for reference or comparative data, and would be using the data ''in-the-blind'' (i.e., without direct contact with the data originators). In addition to data set and data structural descriptors, such an individual would require much more detail about research origins and data set status and accessibility. Finally, a Level III data reuser might be conducting an audit of the data for ethical or environmental litigation, or conducting a peer review for a citable publication involving second-party reproduction of computational results. Satisfying this objective may require that the individual have access to the most comprehensive set of metadata, including all supplemental descriptors. These examples illustrate the relationship between several types of secondary usage and the variability in requisite metadata content.

With high levels of projected or actual secondary data usage and increasing metadata content, the utility of data is improved by adding structure to the metadata. Expanding upon the example presented in Table 2, we define three levels of format or structure (low, medium, high) that also roughly correspond to the degree of formalization and the level of effort involved in adding that structure (Fig. 3). The lowest level of metadata structure may simply consist of a hard-copy document or free-format ASCII text in narrative form. This low level of metadata structure may be suitable for exchange with expert colleagues, but inadequate for electronic data set publication or other uses. A medium level of structure may encompass mixed format or par-

TABLE 1. Standard ecological metadata descriptors and examples (based on Michener et al. 1987, 1990, Federal Geographic Data Committee 1994, Kirchner et al. 1995).

Descriptors	Examples
Class I. Data set descriptors	
A. Data set identity	Title or theme of data set
B. Data set identification code	Database accession numbers or site-specific codes used to uniquely identify data set
C. Data set description	
1. Originator(s)	Names and addresses of principal investigator(s) associated with data set
2. Abstract	Descriptive abstract summarizing research objectives, data contents (including temporal, spatial, and thematic domain), context and potential uses of data set
D. Key words	Location (spatial scale), time period and sampling frequency (temporal scale), theme or contents (thematic scale)
Class II. Research origin descriptors	
A. ''Overall'' project description	[Note: this section may be essential if data set represents a component of a larger or more comprehensive database; otherwise, relevant items may be incorporated into II.B.]
1. Identity	Project title or theme
2. Originator(s)	Name(s) and address(es) of principal investigator(s) associated with project
3. Period of study	Date commenced, date terminated, or expected duration
4. Objectives	Scope and purpose of research program
5. Abstract	Descriptive abstract summarizing broader scientific scope of ''overall'' research project
6. Source(s) of funding	Grant and contract numbers, names and addresses of funding sources
B. ''Specific subproject'' description	
1. Site description	
a. Site type	Descriptive (e.g., short-grass prairie, blackwater stream, etc.)
b. Geography	Location (e.g., latitude/longitude), size
c. Habitat	Detailed characteristics of habitats sampled
d. Geology, landform	Soils, slope/elevation/aspect, terrain/physiography, geology/lithology
e. Watersheds, hydrology	Size, boundaries, receiving streams, etc.
f. Site history	Site management practices, disturbance history, etc.
g. Climate	Descriptive summary of site climatic characteristics
2. Experimental or sampling design	
a. Design characteristics	Description of statistical/sampling design
b. Permanent plots	Dimension, location, general vegetation characteristics (if applicable).
c. Data collection period, frequency, etc.	Information necessary to understand temporal sampling regime
3. Research methods	
a. Field/laboratory	Description or reference to standard field/laboratory methods
b. Instrumentation	Description and model/serial numbers
c. Taxonomy and systematics	References for taxonomic keys, identification and location of voucher specimens, etc.
d. Permit history	References to pertinent scientific and collecting permits
e. Legal/organizational requirements	Relevant laws, decision criteria, compliance standards, etc.
4. Project personnel	Principal and associated investigator(s), technicians, supervisors, students
Class III. Data set status and accessibility	
A. Status	
1. Latest update	Date of last modification of data set
2. Latest archive date	Date of last data set archival
3. Metadata status	Date of last metadata update and current status
4. Data verification	Status of data quality assurance checking
B. Accessibility	
1. Storage location and medium	Pointers to where data reside (including redundant archival sites)
2. Contact person(s)	Name, address, phone, fax, electronic mail
3. Copyright restrictions	Whether copyright restrictions prohibit use of all or portions of the data set
4. Proprietary restrictions	Any other restrictions that may prevent use of all or portions of data set
a. Release date	Date when proprietary restrictions expire
b. Citation	How data may be appropriately cited
c. Disclaimer(s)	Any disclaimers that should be acknowledged by secondary users
5. Costs	Costs associated with acquiring data (may vary by size of data request, desired medium, etc.)
Class IV. Data structural descriptors	
A. Data set file	
1. Identity	Unique file names or codes
2. Size	Number of records, record length, total number of bytes, etc.
3. Format and storage mode	File type (e.g., ASCII, binary, etc.), compression schemes employed (if any), etc.

TABLE 1. Continued.

Descriptors	Examples
4. Header information	Description of any header data or information attached to file [Note: may include elements related to ''variable information'' (IV.B.); if so, could be linked to appropriate section(s)]
5. Alphanumeric attributes	Mixed, upper, or lower case
6. Special characters/fields	Methods used to denote comments, ''flag'' modified or questionable data, etc.
7. Authentication procedures	Digital signature, checksum, actual subset(s) of data, and other techniques for assuring accurate transmission of data to secondary users
B. Variable information	
1. Variable identity	Unique variable name or code
2. Variable definition	Precise definition of variables in data set
3. Units of measurement	Units of measurement associated with each variable
4. Data type	
a. Storage type	Integer, floating point, character, string, etc.
b. List and definition of variable codes	Description of any codes associated with variables
c. Range for numeric values	Minimum, maximum
d. Missing value codes	Description of how missing values are represented in data set
e. Precision	Number of significant digits
5. Data format	
a. Fixed, variable length	
b. Columns	Start column, end column
c. Optional number of decimal places	
C. Data anomalies	Description of missing data, anomalous data, calibration errors, etc.
Class V. Supplemental descriptors	
A. Data acquisition	
1. Data forms or acquisition methods	Description or examples of data forms, automated data loggers, digitizing procedures, etc.
2. Location of completed data forms	
3. Data entry verification procedures	Procedures employed to verify that digital data set is error free
B. Quality assurance/quality control procedures	Identification and treatment of outliers, description of quality assessments, calibration of reference standards, equipment performance results, etc.
C. Related materials	References and locations of maps, photographs, videos, GIS data layers, physical specimens, field notebooks, comments, etc.
D. Computer programs and data-processing algorithms	Description or listing of any algorithms used in deriving, processing, or transforming data
E. Archiving	
1. Archival procedures	Description of how data are archived for long-term storage and access
2. Redundant archival sites	Locations and procedures followed
F. Publications and results	Electronic reprints, lists of publications resulting from or related to the study, graphical/statistical data representations, etc.
G. History of data set usage	
1. Data request history	Log of who requested data, for what purpose, and how data set was actually used
2. Data set update history	Description of any updates performed on data set
3. Review history	Last entry, last researcher review, etc.
4. Questions and comments from secondary users	Questionable or unusual data discovered by secondary users, limitations or problems encountered in specific applications of data, unresolved questions or comments

tially parameterized information fields that could be searched easily (electronically) by a third party. For example, a medium level of structure may minimally support search and retrieval of Level I descriptors. High levels of structure may be used to store information in fixed format or highly parameterized fields such as those associated with the more sophisticated database management systems (DBMS). Some DBMS software supports development of executable and searchable metadata databases. Although this high level of structure is good practice for projects that require periodic data audits, it may be excessive for other objectives.

Increased metadata structure is beneficial for at least two reasons. First, the checklist character of structured metadata provides the data originator with a memory aid for what is important to record about the data to enable his/her own reuse as well as to facilitate utilization by others. Second, increased structure facilitates verification of results and development of searchable catalogues and database interfaces, making the data available to a larger potential population of users with a wider range of processing software. As an example, the NOAA Earth System Data Directory (Barton 1995) and the U.S. Geological Survey's Global Land Information System (GLIS; Scholz and Smith 1995) utilize Directory Interchange Format (DIF) as a mechanism for ensuring that a minimum level of metadata is available during searches for data sets.

Although increasing metadata structure (i.e., format definition) reduces the burden on data reusers, it sig-

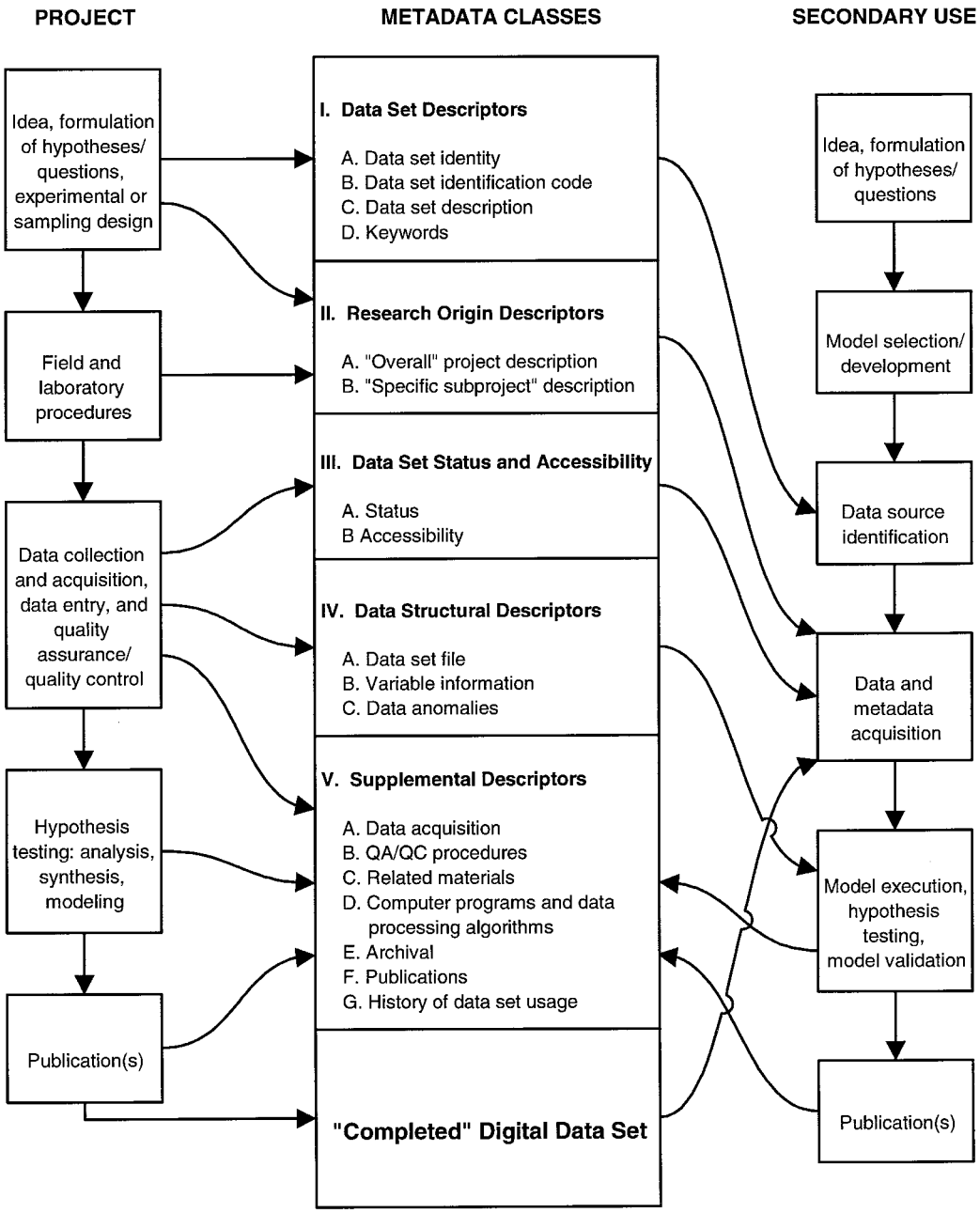

FIG. 2. Metadata development (middle column, refer to Table 1 for more information) in relation to project design and implementation (left column), and relationship of project metadata with subsequent secondary data set utilization in a modeling project (right column).

nificantly increases the burden on the data originator. Although one may argue that the burden should be on the data reuser to ferret out the relevant details, this is frequently impossible. Thus, data reuse is frequently based on intelligent and well-intentioned guesses. For example, if the data originator is still alive, it may be that he/she does not remember what quality assurance procedures or analytical algorithms were used, since the relevant information was never documented, or programmers or knowledgeable technical personnel have since left the project. Ultimately, the burden falls on both the data originator and secondary users to apply good practices and minimize the propagation of errors arising from unintentional misuse of the data. The Carbon Dioxide Information Analysis Center, for example, emphasizes the value-added component of data sets resulting from joint participation of scientists and users in metadata preparation, rigorous QA/QC processing, peer review of data and metadata, ''beta testing'' of data sets prior to general release, and incorporation of

TABLE 2. Content of metadata (refer to classes in Table 1) associated with three levels of secondary data utilization.

Metadata descriptor classes	Levels of secondary data utilization and associated metadata content		
	Level I: exchange with expert colleague	Level II: searchable and third party data reuse	Level III: publishable and auditable
I. Data set descriptors	X	X	X
II. Research origin descriptors		X	X
III. Data set status and accessibility		X	X
IV. Data structural descriptors	X	X	X
V. Supplemental descriptors			X

user feedback into its data packages (Boden 1995, National Research Council 1995*a*).

CONCLUSION AND RECOMMENDATIONS

Basic and applied ecological research depend upon the availability of high-quality data. If a priori consideration is paid to the development of high-quality data sets and accompanying metadata, then individual scientists and organizations can focus valuable time and effort on performing appropriate analyses with the requisite high-quality data. As metadata and metadata standards are developed and implemented, individual scientists and organizations can further benefit by being able to easily reuse data developed for other applications.

Further progress in development, adoption, and implementation of nongeospatial, ecological metadata standards depends upon data and metadata being recognized as representing an integral component of the scientific process. Study repeatability, comparative ecological studies, attempts to scale up domain-specific studies to broader spatial, temporal, and thematic domains, ecological simulation modeling and model validation, and more applied ecological research (e.g., restoration ecology, ecological risk assessment, research into sustainable development, etc.) all depend upon the availability of archived data and, equally importantly, upon the ability to understand those data via the metadata. Data are more frequently being reused by data originators and being utilized by other scientists who often were not involved in the data collection. Thus, the scientific value of being able to reuse data and to utilize data for multiple objectives that may not have been foreseen by the data originator(s) may far exceed the perceived value associated with publications resulting from the original study.

All data should be accompanied by metadata. The completeness of the metadata governs the length of time and the extent to which data can be reused by the original investigator(s) and utilized by other scientists,

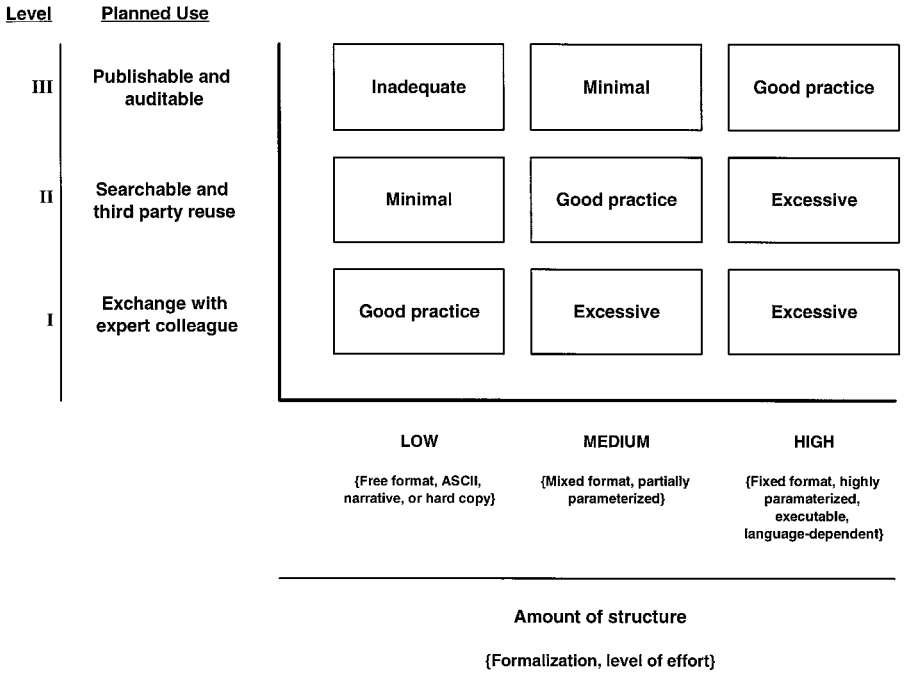

FIG. 3. Degree of metadata format/structure sufficient for three levels of projected secondary data utilization.

resource managers, decision makers, and other potential users. Just as the data and information contained in a manuscript support peer review of the publication and the conclusions reached therein, metadata support peer review of the data and facilitate secondary utilization. Ideally, the metadata should be physically linked as closely as possible to the data. For example, non-imagery data and associated metadata collected under the auspices of the U.S. Department of Energy's Atmospheric Radiation Program are integrated and stored in Network Common Data Format (netCDF) structure (Melton 1995).

If state and federal agencies, scientific societies, and academic institutions perceive the value of data sets collected by grantees or members of their organizations, then appropriate value should be placed on the publication of data and metadata, in addition to more traditional peer-reviewed publications (National Research Council 1995*a*). Perhaps data and accompanying metadata for ''irreplaceable'' or otherwise valuable ecological data sets could be published in an as-yet-to-be-developed electronic journal and then submitted to a data archive. Such data sets would then be citable in the scientific literature. Ultimately, however, successful incentives will rely upon organizations placing appropriate value on data/metadata publications during the scientific merit review process.

Agencies and scientific societies should promote metadata development and standardization (National Research Council 1995*a*). For example, the geographical sciences community has developed spatial data transfer standards (National Institute of Standards and Technology 1992) and metadata standards (Federal Geographic Data Committee 1994) that are widely endorsed by federal and state agencies, scientific societies, and academic institutions. The National Biological Service (American Institute of Biological Sciences 1995) and a subcommittee of the Federal Geographic Data Committee (M. Nyquist, *personal communication*) are developing additional extensions that would comprise supersets of the existing geospatial metadata content standards (Federal Geographic Data Committee 1994), and would be appropriate for biological resource, cultural, and demographic metadata. When appropriate, existing metadata standards (e.g., spatial metadata standards) should be endorsed and promoted by the ecological sciences. Where metadata standards are incomplete or do not exist, attention should focus on developing, endorsing, and adopting appropriate standards.

Data and metadata should be independent of hardware and software to the fullest extent possible (Conley and Brunt 1991). Proprietary data storage formats inevitably change through time or are replaced by new formats. Thus, the life-span (long-term utility) of data and metadata may be severely degraded when data/metadata conform to a proprietary standard rather than a more generic ''industry-wide standard.'' Agencies and institutions may find it beneficial to collaborate in development and support of digital library services for data and metadata archiving.

Funding agencies, scientific societies, and research institutions should recognize that there are costs, as well as benefits, associated with archiving data and developing and maintaining the requisite metadata. Thus, enhanced levels of funding to support these ancillary activities should be recognized as being necessary and appropriate. Similarly, funds would be required to resurrect valuable historic data and metadata. However, it should also be understood that historic data are frequently more readily retrieved (resurrected) than are the essential metadata, as demonstrated by the IBP example.

Metadata descriptors included in Table 1 may serve as the basis for initially developing metadata for individual scientists, laboratories, and projects until formal metadata standards emerge and supporting software is developed. As ecologists attempt to develop metadata for their numerous and diverse subdisciplines, it is likely that additional metadata descriptors will be required to fully document specific data sets. Small groups of scientists focused on a specific research objective, such as synthesizing data on a particular topic, may benefit significantly from metadata implementation efforts. Although compiling metadata for such a project would be an extremely daunting challenge for any one individual, it might be possible to plan one or more workshops whereby scientists exchange metadata, review metadata for completeness and comprehensibility, and fill in missing gaps. Successful completion of a comprehensive, synthetic database, including both data and metadata, may lead to new and innovative experiments and analyses that could be tailored to fit the existing, developing database.

As ecologists address the complex issues associated with metadata standardization, long-term data and metadata archiving, and secondary data utilization, a cautionary note from the geographical sciences may be in order. Specifically, Chrisman (1994) asserts that ''all the standardized procedures in the world cannot ensure that the product actually satisfies the user's needs.'' He emphasizes the joint responsibilities of users and providers in relation to spatial data use and documentation, the need to incorporate spatial statistics more fully into GIS, research leading to a better understanding of error propagation in GIS, and, importantly, the critical need to develop ''procedures that can handle large differences in resolution, accuracy and other key properties.'' Ecology, like geography, is interdisciplinary by its very nature. However, a review of interdisciplinary environmental research and assessments conducted by the Committee for a Pilot Study on Database Interfaces concluded that ''the existing missions and attendant reward systems of research organizations act to inhibit the data sharing, mutual support, and interdisciplinary mind-set needed for successful data interfacing'' (Na-

tional Research Council 1995*a*). The increasing reliance on long-term, broad-scale, and multi- and interdisciplinary data to address issues related to global change, biodiversity, sustainability, and other societal concerns highlights the need for retaining important ecological data sets in an accessible and understandable form. Increased attention to developing high-quality data sets and their attendant metadata; understanding how uncertainty, error propagation, and research and statistical assumptions affect the ''fitness'' of data sets for intended and unintended uses; and promoting a sense of stewardship for ecological data will certainly enhance the interdisciplinary nature of ecology as a science.

ACKNOWLEDGMENTS

This work was supported by an Andrew Mellon Foundation grant to Katherine Gross and was performed under the auspices of the Ecological Society of America's Future of Long-Term Ecological Data Committee that was chaired by K. Gross. Thoughtful comments by Katherine Gross, Mark Harmon, Don Henshaw, Judy Meyer, Scott Miller, Maurice Nyquist, Catherine Pake, Gody Spycher, and David Strayer improved the manuscript and are greatly appreciated.

LITERATURE CITED

American Institute of Biological Sciences. 1995. AIBS review to National Biological Service: content standard for non-geospatial metadata workshop. American Institute of Biological Sciences, Reston, Virginia, USA.

Barton, G. S. 1995. Directory Interchange Format: a metadata tool for the NOAA Earth System Data Directory. Pages 19–23 *in* R. B. Melton, D. M. DeVaney, and J. C. French, editors. The role of metadata in managing large environmental science datasets. Pacific Northwest Laboratory, Richland, Washington, USA.

Boden, T. A. 1995. Metadata compiled and distributed by the Carbon Dioxide Information Analysis Center for global climate change and greenhouse gas-related data bases. Pages 13–18 *in* R. B. Melton, D. M. DeVaney, and J. C. French, editors. The role of metadata in managing large environmental science datasets. Pacific Northwest Laboratory, Richland, Washington, USA.

Bowser, C. J. 1986. Historic data sets: lessons from the past, lessons for the future. Pages 155–179 *in* W. K. Michener, editor. Research data management in the ecological sciences. University of South Carolina Press, Columbia, South Carolina, USA.

Briggs, J. M., and H. Su. 1994. Development and refinement of the Konza Prairie LTER research information management program. Pages 87–100 *in* W. K. Michener, J. W. Brunt, and S. G. Stafford, editors. Environmental information management and analysis: ecosystem to global scales. Taylor and Francis, London, UK.

Brown, J. H., and J. Roughgarden. 1990. Ecology for a changing earth. Bulletin of the Ecological Society of America **71**:173–188.

Brunt, J. W. 1994. Research data management in ecology: a practical approach for long-term projects. Pages 272–275 *in* J. C. French and H. Hinterberger, editors. Seventh International Working Conference on Scientific and Statistical Database Management. IEEE Computer Society Press, Washington, D.C., USA.

Chrisman, N. R. 1994. Metadata required to determine the fitness of spatial data for use in environmental analysis. Pages 177–190 *in* W. K. Michener, J. W. Brunt, and S. G. Stafford, editors. Environmental information management and analysis: ecosystem to global scales. Taylor and Francis, London, UK.

Colwell, R. K. 1995. Ecological Society of America special committee on ESA communications in the electronic age. Bulletin of the Ecological Society of America **76**:120–131.

Conley, W., and J. W. Brunt. 1991. An institute for theoretical ecology? Part V: Practical data management for cross-site analysis and synthesis of ecological information. Coenoses **6**:173–180.

Defense Mapping Agency. 1992. Vector Product Format, Military Standard 600006. Department of Defense, Washington, D.C., USA.

Digital Geographic Information Working Group. 1991. DIGEST: a digital geographic exchange standard. Defense Mapping Agency, Washington, D.C., USA.

Eagan, P. D., and S. J. Ventura. 1993. Enhancing value of environmental data: data lineage reporting. Journal of Environmental Engineering **119**:5–17.

Emmingham, W. H., and G. A. Lundburg. 1977. Climatic and physiological data summaries for the H. J. Andrews reference stand network. Internal report Number 166. Coniferous forest biome, ecosystem analysis studies. U.S. International Biological Program, University of Washington, Seattle, Washington, USA.

Federal Geographic Data Committee. 1994. Content standards for digital geospatial metadata (June 8). Federal Geographic Data Committee, Washington, D.C., USA.

Franklin, J. F., C. S. Bledsoe, and J. T. Callahan. 1990. Contributions of the long-term ecological research program. BioScience **40**:509–523.

Gosz, J. R. 1994. Sustainable Biosphere Initiative: data management challenges. Pages 27–39 *in* W. K. Michener, J. W. Brunt, and S. G. Stafford, editors. Environmental information management and analysis: ecosystem to global scales. Taylor and Francis, London, UK.

Kareiva, P., and M. Anderson. 1988. Spatial aspects of species interactions: the wedding of models and experiments. Pages 35–50 *in* A. Hastings, editor. Community ecology. Springer-Verlag, New York, New York, USA.

Kellogg Biological Station. 1982. Data management at biological field stations. Report of a workshop held May 17–20, 1982. W. K. Kellogg Biological Station, Michigan State University, Hickory Corners, Michigan, USA.

Kirchner, T. B. 1994. Data management and simulation modelling. Pages 357–375 *in* W. K. Michener, J. W. Brunt, and S. G. Stafford, editors. Environmental information management and analysis: ecosystem to global scales. Taylor and Francis, London, UK.

Kirchner, T., H. Chinn, D. Henshaw, and J. Porter. 1995. Documentation standards for data exchange. Pages 5–8 *in* R. Ingersoll and J. Brunt, editors. Proceedings of the 1994 LTER Data Management Workshop. Long-Term Ecological Research Network Office, University of Washington, Seattle, Washington, USA.

Levin, S. A. 1992. The problem of pattern and scale in ecology. Ecology **73**:1943–1967.

Melton, R. B. 1995. Metadata in the atmospheric radiation measurement program. Pages 9–12 *in* R. B. Melton, D. M. DeVaney, and J. C. French, editors. The role of metadata in managing large environmental science datasets. Pacific Northwest Laboratory, Richland, Washington, USA.

Michener, W. K., R. J. Feller, and D. G. Edwards. 1987. Development, management, and analysis of a long-term ecological research information base: example for marine macrobenthos. Pages 173–188 *in* T. P. Boyle, editor. New approaches to monitoring aquatic ecosystems. ASTM STP 940, American Society for Testing and Materials, Philadelphia, Pennsylvania, USA.

Michener, W. K., A. B. Miller, and R. Nottrott. 1990. Long-term ecological research core data set catalog. Belle W.

Baruch Institute for Marine Biology and Coastal Research, University of South Carolina, Columbia, South Carolina, USA.

National Institute of Standards and Technology. 1992. Spatial Data Transfer Standard (Federal Information Processing Standard 173). National Institute of Standards and Technology, Gaithersburg, Maryland, USA.

National Research Council. 1991. Solving the global change puzzle: a U.S. strategy for managing data and information. National Academy Press, Washington, D.C., USA.

———. 1993. A biological survey for the nation. National Academy Press, Washington, D.C., USA.

———. 1995*a*. Finding the forest in the trees. National Academy Press, Washington, D.C., USA.

———. 1995*b*. Preserving scientific data on our physical universe: a new strategy for archiving the nation's scientific information resources. National Academy Press, Washington, D.C., USA.

National Science Foundation. 1994. Grant proposal guide. National Science Foundation, Arlington, Virginia, USA.

Pace, M. L. 1993. Forecasting ecological responses to global change: the need for large-scale comparative studies. Pages 356–363 *in* P. M. Kareiva, J. G. Kingsolver, and R. B. Huey, editors. Biotic interactions and global change. Sinauer Associates, Sunderland, Massachusetts, USA.

Scholz, D. K., and T. B. Smith. 1995. The Global Land Information System: the use of metadata on three levels. Pages 25–27 *in* R. B. Melton, D. M. DeVaney, and J. C. French, editors. The role of metadata in managing large environmental science datasets. Pacific Northwest Laboratory, Richland, Washington, USA.

Stafford, S. G., P. B. Alabach, K. L. Waddell, and R. L. Slagle. 1986. Data management procedures in ecological research. Pages 93–114 *in* W. K. Michener, editor. Research data management in the ecological sciences. University of South Carolina Press, Columbia, South Carolina, USA.

Stafford, S. G., J. W. Brunt, and W. K. Michener. 1994. Integration of scientific information management and environmental research. Pages 3–19 *in* W. K. Michener, J. W. Brunt, and S. G. Stafford, editors. Environmental information management and analysis: ecosystem to global scales. Taylor and Francis, London, UK.

Strayer, D. S., J. S. Glitzenstein, C. G. Jones, J. Kolasa, G. E. Likens, M. J. McDonnell, G. G. Parker, and S. T. A. Pickett. 1986. Long-term ecological studies: an illustrated account of their design, operation, and importance to ecology. Institute of Ecosystem Studies, Millbrook, New York, USA.

Strebel, D. E., B. W. Meeson, and A. K. Nelson. 1994. Scientific information systems: a conceptual framework. Pages 59–85 *in* W. K. Michener, J. W. Brunt, and S. G. Stafford, editors. Environmental information management and analysis: ecosystem to global scales. Taylor and Francis, London, UK.

Webster, F. 1991. Solving the global change puzzle: a U.S. strategy for managing data and information. Report by the Committee on Geophysical Data Commission on Geosciences, Environment and Resources, National Research Council, National Academy Press, Washington, D.C., USA.