Evolution of Ecological Metadata Structures at the H. J. Andrews Experimental Forest Long-Term Ecological Research (LTER) Site[1]

Donald L. Henshaw[2]
Gody Spycher[3]

Abstract—The success of any monitoring program depends on an information management system that supports the collection, quality control, archival and long-term accessibility of collected data and associated metadata. Intensive, research-driven site monitoring has been conducted on the H. J. Andrews Experimental Forest Long-Term Ecological Research (LTER) site since the 1950's. The resulting, diverse ecological databases are managed through the Forest Science Data Bank (FSDB) which features a metadata system to facilitate data production through the use of generic and database-specific tools. Increasing informational needs necessitate a system that is easily searchable and allows the integration of diverse types of information. Towards this end, FSDB personnel are developing an information system based on a normalized metadata database. The system consists of a catalog of research products such as databases and publications, and related tables to permit searching for these products by personnel, keywords, locations, and species.

Monitoring of forest ecosystem resources was initiated on the H. J. Andrews Experimental Forest shortly after its establishment in 1948. Early research efforts were conducted predominantly by the U.S. Forest Service Pacific Northwest Research Station (USFS PNW) and concentrated on forest watersheds, soils, and vegetation. With the inception of the International Biological Program/Coniferous Forest Biome (IBP/CFB) in 1969, university scientists began to play increasingly important roles. Long-term measurement programs that focused on climate, streamflow, water quality, and vegetation succession were established as part of the National Science Foundation (NSF)-funded Long-Term Ecological Research (LTER) program in 1980. The Andrews Forest LTER site now serves as a focal point for stream and forest ecosystem research, bringing together a community of over 50 university and federal research scientists. Building on these central themes and long-term research projects, research currently emphasizes predicting the effects of natural disturbance, land use, and climate change on ecosystem structure, function, and species composition.

The Forest Science Data Bank (FSDB) was created to house data generated from LTER scientists and other collaborating researchers (Stafford et al. 1984, 1988, Stafford 1993). The FSDB currently stores over 250 long-term and opportunistic databases from diverse scientific disciplines. The FSDB is funded by the USFS PNW, Oregon State University (OSU), and the LTER, and is supported by the Quantitative Sciences Group (QSG). The QSG is staffed by both OSU and USFS PNW personnel and provides database, statistical, software, and hardware support to the local research community. The FSDB has benefited greatly from the support and participation of the scientific community, and conversely, long-term measurement programs do not exist independently of information management systems that maintain and preserve measurement data for the long-term.

While the strategy and most of the FSDB data structures remain unchanged, the demand for rapid access to well-documented, high-quality long-term data has increased dramatically. This demand for information, coupled with the development of the Internet, web-based access tools, improved relational data management systems (RDBMS), spatial databases and accompanying tools, has signaled the need for changes. The FSDB is now in a transitional period as we design and begin the implementation of a more integrated infrastructure for managing scientific information. The intent of this paper is to review the evolutionary phases of the FSDB, and to report on the progress of its newest phase as the FSDB moves toward modern technologies of client-server architecture and a web-based user interface.

The Forest Science Data Bank (FSDB)

The Human Context

Ecological data management is dependent on a strong collaboration of data managers with this research community (Stafford et al. 1986). Lack of cooperation between research scientists and data management staff often leads to the creation of data sets that require restructuring, are inadequately documented, or are never submitted to the data archive. Costs and time requirements for data documentation and validation in centralized information management systems are often underestimated (Thorley and Trathan 1994). These problems can be offset by understanding and accepting the costs of information management, and by following a systematic approach where scientists involve data managers in research planning and developing sampling protocols (Stafford 1993). The FSDB employed this

[1]Paper presented at the North American Science Symposium: Toward a Unified Framework for Inventorying and Monitoring Forest Ecosystem Resources, Guadalajara, Mexico, November 1-6, 1998.

[2]Donald L. Henshaw is Statistician, USDA Forest Service, Pacific Northwest Research Station, located at the Forestry Sciences Laboratory, Corvallis, OR, USA.

[3]Gody Spycher is Senior Research Assistant, Department of Forest Science, Oregon State University, Corvallis, OR, USA.

systematic approach into its conceptual structure from the beginning (Stafford et al. 1984), and resulting activities have brought discipline to the collection and organization of the data and metadata (NRC 1995). Science involvement with information management becomes more important as the complexity of the research information increases.

FSDB History

Early data management efforts were initiated in the 1970's during the IBP program, the predecessor of LTER. The need to compare and quantify ecosystem processes among the different biomes led to methods for documenting data (Gross et al. 1995). The concept of a data set abstract was established, and Andrews Forest IBP data sets were documented for data structure, descriptive variable definitions, and descriptive codes. The data were stored on mainframe computers, access was exclusively available through the data manager, and documentation existed only on hardcopy forms. Nevertheless, a tradition of managing research information was born and carried into the LTER years starting in 1980.

Throughout the 1980's the FSDB evolved and experienced substantial growth as both legacy and new databases were added into the system. Improved hardware and software technologies enabled a fuller implementation of FSDB's conceptual structure. Commercially available Relational Database Management Systems (RDBMS) lead to development of relational database structures for housing study metadata including global database catalogs and detailed documentation of individual study data structures, variable descriptions, and study abstract information. Networked computer systems also allowed local researchers direct access to FSDB data sets.

The LTER program has always emphasized data management to fulfill its primary goals of long-term collaborative research including comparative, cross-site analyses (Franklin et al. 1990). The LTER network of data managers has been a tremendous asset to the program and to each individual site. The Andrews LTER has consistently participated within this network and has incorporated most of the recommended data protocols and metadata standards into its data management system (NSF 1984, Gorentz 1990).

The Current Metadata System

The FSDB houses databases that are "wide" rather than "deep" (Porter 1998). Whereas a "deep" database might specialize in one topical area and might contain large numbers of observations for one data type, a "wide" database contains many types of data with different structures, data from diverse ecological research topics, and with relatively few observations for each data type. Given this diversity within the data repository, the FSDB has concentrated on the development of generic tools that operate on metadata content and that can be used for all individual study databases. This approach has significantly reduced the time required for data production and permits the maintenance of multiple databases (Spycher et al. 1996).

The current FSDB metadata system includes database catalogs, table definition files, domain tables, and tables containing database-specific rules (Stafford 1993, Spycher et al. 1996). Maintaining complex metadata in relational database structures has many advantages (Stonebraker 1994). The FSDB quality control system itself consists of a set of simple procedures providing flexible, generic data validation. By maintaining standardized metadata structures for every FSDB database, mechanisms were developed to automatically perform validation checks based on standard metadata and specific database rules. Additionally, metadata tools are used (1) to guide users in understanding database content, (2) for global queries of the data catalogs, (3) for packaging data set documentation reports and (4) for other generic access functions such as webpage creation, automatic data entry form setup, and automatic import/export of ASCII files to RDBMS files (Spycher et al. 1996).

Limitations of the Existing System

The FSDB has traditionally housed conventional, non-spatial study databases. However, there is a need to manage more diverse information products such as Geographical Information System (GIS) coverages, remote sensing images, research publications, models, maps, photographs, and other documents including study plans, proposals, methods manuals, and web page documents. Information managers at the Andrews Forest have tended to maintain these other information products such as spatial coverage data and research publications as separate entities. As a result we cannot, for example, relate a publication with a database, a spatial coverage with a companion non-spatial data set, or a database contact person with the personnel directory.

Pervasive redundancies also exist within the system. Separate keyword lists are maintained for both research publications and databases; identical study sites are often described in multiple study data abstracts; species lists do not always reference master taxonomic databases; and the domain for widely used coded variables such as "decay class" may be described multiple times.

Web-based tools and navigational aids, not dependent on the computer literacy of the user, are necessary to facilitate data sharing (Günther 1998). In this regard, the Andrews Forest LTER has made many databases, models, personnel, and publication lists available on its webpage (http://www.fsl.orst.edu/lter). However, not only has the number of information products to manage increased, but the metadata context has also expanded to include personnel, location, keyword, and species data. Web-based tools are needed to allow and assist researchers in producing metadata, as well as to dynamically search and integrate metadata databases with information products.

A New Structure for Ecological Metadata

Recent publications have provided strong guidance on metadata content for spatial data sets (Federal Geographic Data Committee (FGDC)) and non-geospatial data (Gross et al. 1995, Michener et al. 1997). However, little advice has been provided on how this content might be structured for efficient management and access. The need to conform to developing metadata standards, and to manage all information products in an integrated, comprehensive information

Table 1.—Information product tables and descriptions maintained in the metadata database.

Information product table	Description of information product table
Study_data	Catalog of conventional, non-spatial databases
Remote_image	Catalog of remote sensing images, i.e., satellite images, scanned aerial photographs
Publication	Bibliography and abstracts of research publications
GIS_coverage	Catalog of Geographical Information System spatial coverage data
Document	Catalog of study plans, proposals, method manuals, and Web documents

management system, has motivated efforts to rationally structure and supplement the FSDB metadata database.

The objective was to design a normalized metadata database, "one thing in one place" (ERwin 1996), for ecological data objects as a foundation for an ecological information system. The main components or entities include a central CATALOG of information products, information product tables (STUDY_DATA, REMOTE_IMAGE, PUBLICATION, GIS_COVERAGE, and DOCUMENT) (See Table 1) and metadata tables (SPECIES, LOCATION, KEYWORD, and PERSONNEL) for finding and documenting information products. All components are linked through the central CATALOG (See Fig. 1). The implementation of a normalized database structure within the RDBMS will allow searches and linkages of entities through Structured Query Language (SQL).

In practice, CATALOG contains a list of all products as well as general information pertaining to the product, such as title, security restrictions or last revision date. The table CATALOG_TYPE indicates the type of data object or information product. The information product tables are metadata catalogs of that particular data object and contain information specific to that type of product. (Note that the information product tables do not include the actual data objects.) Additionally, each catalog item can be linked to appropriate keywords, locations, personnel, or species if applicable through associative tables such as CATALOG_KEYWORD or CATALOG_LOCATION. One CATALOG item can also be linked with any other through the RELATED_CATALOG table, allowing, for example, the capability to connect a study database with a companion spatial database or publication (See Fig. 2).

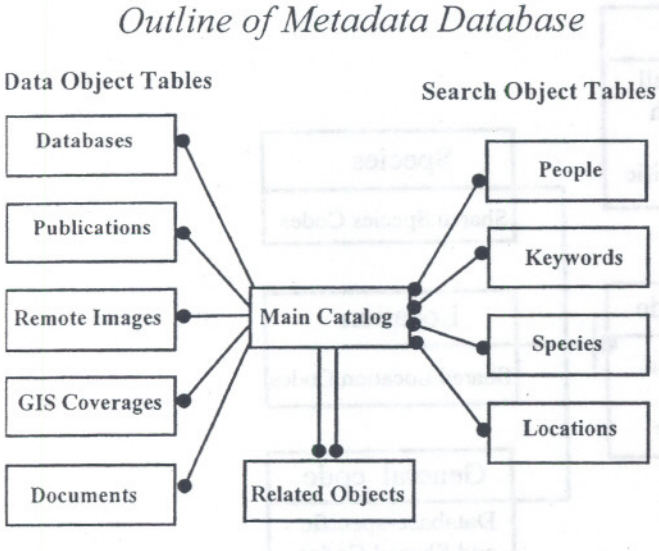

Figure 1.—A simplified structure including only the main entities and relationships of the FSDB Metadata Database (——• represents one-to-many relationships, •——• represents many-to-many relationships).

Of the five information product tables shown, only the metadata for conventional, non-spatial databases (STUDY_DATA) have been fully integrated (See Fig. 3). This subsystem provides for shared variables and codes. The VARIABLE table contains the attributes of all variables in the entire system. The variables may be generic (shared by several databases) or database-specific. In practice this means a reduction in the number of definitions for commonly used variables such as "percent vegetative cover". Similar redundancies are avoided for variable domains, which have been sub-typed, i.e., divided into mutually exclusive categories, into GENERAL_CODE, SPECIES, and LOCATION. For example, a coded variable such as "decay class" can be described in GENERAL_CODE and be shared among all study data using this same coding method. In addition, coded variables of species and location can directly refer to the master catalogs, and need not be redefined. SPECIES can be set up with the desired degree of taxonomic hierarchical detail. LOCATION is sub-typed into various groups such as point locations, watersheds or Research Natural Areas because the attribute sets are group-specific.

The metadata tables SPECIES, LOCATION, KEYWORD, and PERSONNEL serve multiple functions:

1. They can function as independent 'data', i.e., a comprehensive species list, a sub-typed location system, a keyword and a personnel facility
2. They provide the basis for finding data objects by single or multiple metadata categories
3. They serve as an integral part of the metadata for any given data object, permitting for example the compilation of complete documentation for any data object using SQL queries
4. They are actively used in generic quality control of conventional databases

Three of these metadata tables are recursive; that is, the table is hierarchically structured. In the case of SPECIES this is used to reflect taxonomic hierarchies. LOCATION has an indeterminate number of levels and permits the extraction of all locations (or a specified number of levels of

Logical Model of the Metadata Database

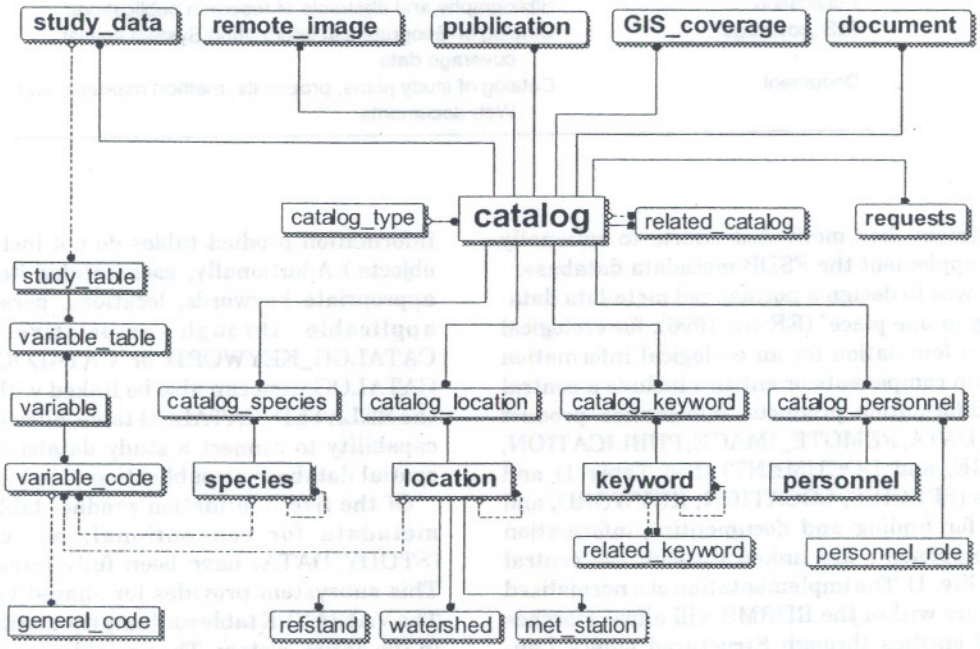

Figure 2.—Entity-Relationship diagram of the FSDB Metadata Database (——• represents one-to-many relationships; solid lines denote identifying relationships; dashed lines show non-identifying relationships, i.e., there is no identification dependency between two tables).

Study_data Section of Metadata Database

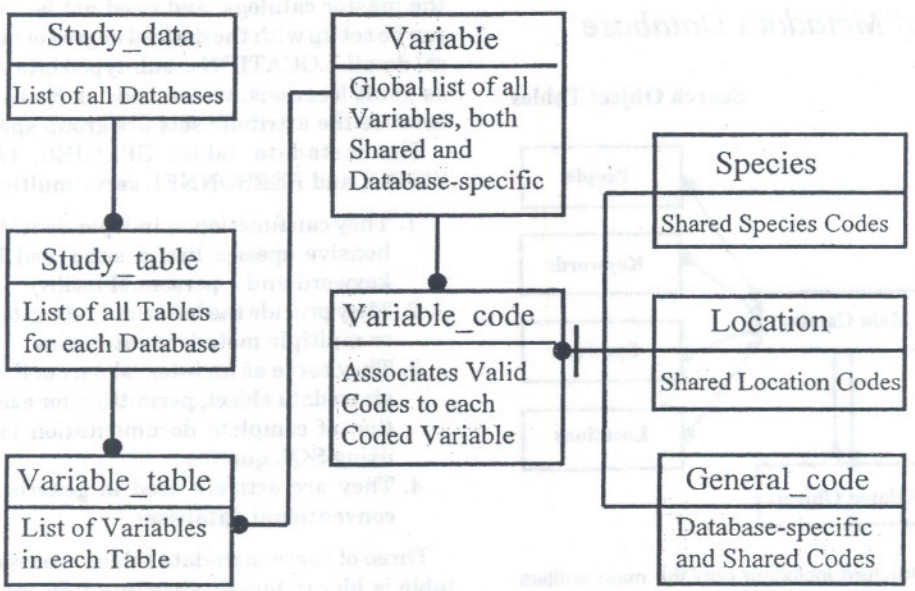

Figure 3.—The entities and relationships of the Study_data Section of the FSDB Metadata Database. Relationships are all one-to-many, and the variable_code table is sub-typed into three mutually exclusive tables.

locations) within a given location. For example, one can query for all locations within Oregon, all watersheds within the Andrews Forest, and all study plots within a watershed. KEYWORD has a 3-level hierarchy implementing a controlled keyword list developed locally for Pacific Northwest forest ecosystem studies. A RELATED_KEYWORD table also allows the ability to relate keywords with each other. The PERSONNEL table lists people and their attributes only once. The PERSONNEL_ROLE table associates functional roles for a given person in the associative CATALOG_PERSONNEL entity. These roles may include ownership of data objects, authorship roles in publications, contact persons, and others.

The REQUESTS table tracks secondary usage of information products, primarily the database products. The table logs information on users requesting the data and for what purpose. User feedback involving encountered limitations or problems with the data are stored here.

This normalized data structure is currently being implemented within the FSDB. It should be noted that all figures only show the simplified structure showing the main entities and relationships. Specific attributes within each table are not listed and are still evolving, but the intent is to conform to current standards for metadata content (FGDC, FLED). Other information product tables may be added such as models and photographs. Security features as well as research project descriptions including funding sources are also planned.

Conclusions

The success of any monitoring program depends on the implementation of data management strategies that support the collection, quality control, and long-term accessibility of generated information. Increasingly, funding agencies require online access to data soon after collection to facilitate intersite exchange of information. Information systems must also accommodate a growing number of information products including conventional databases, GIS spatial data coverages, remote sensing images, publications and other documents. These increasing demands necessitate information systems that are easily searchable and allow the integration of diverse types of information.

The FSDB is pursuing the development of a comprehensive metadata database to help meet these needs. A normalized metadata database structure was designed to avoid redundant information from multiple data sources. The design facilitates searches to locate information products using multiple categories of metadata. Personnel, keyword, location, and species databases are an integral part of the metadata and serve multiple functions. The new structure is complex and demands discipline in the collection and organization of the data and metadata. Strong support and involvement from the research community is an essential ingredient to success.

Acknowledgments

Support for the Forest Science Data Bank is provided by the U.S. Forest Service Pacific Northwest Research Station, Oregon State University Department of Forest Science, and the National Science Foundation LTER grant DEB9632921. The authors would also like to acknowledge Hazel Hammond for support of the Andrews Forest web pages and databases and for critical review of this manuscript. We also thank Chris Middour for instrumental support in the development of the metadata database schema.

Literature Cited

Committee for a Pilot Study on Database Interfaces, U.S. National Committee for CODATA

Commission on Physical Sciences Mathematics and Applications National Research Council [NRC]. 1995. The H.J. Andrews Experimental Forest Long-Term Ecological Research Site. In: The committee finding the forest in the trees: the challenge of combining diverse environmental data. Selected case studies. Washington, DC: National Academy Press: 46-55.

Franklin, J.F.; Bledsoe C.S.; Callahan J.T. 1990. The Long-Term Ecological Research Program: a contributor to ecological science. BioScience 40: 509-523.

Gorentz, John, ed. 1990. Data management at biological field stations and coastal marine laboratories: report of an invitational workshop; 1990 April 22-26; W. K. Kellogg Biological Station, Michigan State University, Lansing, Michigan.

Gross, K.L.; Pake, C.E.; Allen, E. [and others] [FLED]. 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, [Online]. Available: http://www.sdsc.edu/~ESA/FLED/FLED.html.

Günther, O. 1998. Environmental information systems. Berlin; Heidelberg: Springer-Verlag. 244 p.

Michener, William K.; Brunt, James W.; Helly, John J. [and others]. 1997. Nongeospatial metadata for the ecological sciences. Ecological Applications 7(1): 330-342.

National Science Foundation [NSF]. 1984. Data management at biological field stations: report of a workshop; 1982 May 17-20; W.K. Kellogg Biological Station, Michigan State University, Lansing, Michigan.

Platinum Technology, Inc. [ERwin] 1998. ERwin methods guide. Princeton, NJ: Platinum Technology, Inc. 96 p.

Porter, John H. 1998. Scientific databases for environmental research. In: Michener, William

K.; Porter, John H.; Stafford, Susan G., eds. Data and information management in the ecological sciences: a resource guide; 1997 August 8-9; Albuquerque, NM. Albuquerque, NM: LTER Network Office, University of Mexico: 117-122.

Spycher, G.; Cushing, J.B.; Henshaw, D.L. [and others]. 1996. Solving problems for validation, federation, and migration of ecological databases. In: Global networks for environmental information: Proceedings of Eco-Informa '96; 1996 November 4-7; Lake Buena Vista, FL. Ann Arbor, MI: Environmental Research Institute of Michigan (ERIM); 11: 695-700.

Stafford, S.G.; Alaback, P.B.; Koerper, G.J.; Klopsch, M.W. 1984. Creation of a forest science data bank. Journal of Forestry 82(7): 432-433.

Stafford, S.G.; Alaback, P.B.; Waddell, K.L.; Slagle, R.L. 1986. Data management procedures in ecological research. In: Michener, William K., ed. Research data management in the ecological sciences. The Belle W. Baruch Library in Marine Science No. 16. Columbia, SC: University of South Carolina Press: 93-113.

Stafford, Susan G. 1993. Data, data everywhere but not a byte to read: managing monitoring information. Environmental Monitoring and Assessment 26: 125-141.

Stafford, Susan G.; Spycher, Gody; Klopsch, Mark W. 1988. Evolution of the Forest Science Data Bank. Journal of Forestry 86(9): 50-51.

Stonebraker, M. 1994. Sequoia 2000—A reflection on the first three years. IEEE: 108-116.

Thorley, M.R.; Trathan, P.N. 1994. The history of the BIOMASS data centre and lessons learned during its lifetime. Southern Ocean Ecology: The BIOMASS Perspective. Cambridge: 313-322.

Abstract

Aguirre-Bravo, Celedonio and Carlos Rodriguez Franco, compilers. 1999. **North American Science Symposium: Toward a Unified Framework for Inventorying and Monitoring Forest Ecosystem Resources.** Guadalajara, Mexico (November 2-6, 1998). Proceedings RMRS-P-12. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fort Collins, CO USA. 533 p.

The general objective of this Symposium was to build on the best science and technology available to assure that the data and information produced in future inventory and monitoring programs are comparable, quality assured, available, and adequate for their intended purposes, thereby providing a reliable framework for characterization, assessment, and management of forest ecosystems in North America. Central to the syntheses delivered in this Symposium was the conclusion that a fundamental improvement in the approaches used for inventorying and monitoring ecosystem resources is required to meet current and future environmental uncertainties. Specific actions were proposed to address these challenges. These strategic actions are described in the last chapter of these proceedings.

***Editors's Note**: In order to deliver symposium proceedings to users as quickly as possible, many manuscripts did not receive conventional editorial processing. Views expressed in each paper are those of the author and not necessarily those of the sponsoring organizations or the USDA Forest Service. Trade names are used for the information and convenience of the reader and do not imply endorsement or preferential treatment by the sponsoring organizations or the USDA Forest Service.*

You may order additional copies of this publication by sending your mailing information in label form through one of the following media. Please specify the publication title and Proceedings number.

Fort Collins Service Center

Telephone (970) 498-1719

FAX (970) 498-1660

E-mail rschneider/rmrs@fs.fed.us

Web site http://www.fs.fed.us/rm

Mailing Address Publications Distribution
Rocky Mountain Research Station
3825 E. Mulberry Street
Fort Collins, CO 80524