Transition from a Legacy Databank to an Integrated Ecological Information System

Donald L. Henshaw
USDA Forest Service, Pacific Northwest Research Station
Corvallis, Oregon 97331, USA

Gody Spycher
Department of Forest Science, Oregon State University
Corvallis, Oregon 97331, USA

Suzanne M. Remillard
USDA Forest Service, Pacific Northwest Research Station
Corvallis, Oregon 97331, USA

ABSTRACT

Many tasks and issues are encountered in the process of converting a scientific databank containing multiple legacy and long-term study databases into an integrated data production and distribution system. Metadata issues include questions of structure, translation from legacy to new content standards, and connecting spatial with non-spatial metadata. The authors review the history of the Forest Science Data Bank and examine many aspects related to this latest transition to a more comprehensive and better-integrated information management system. The system is designed to accommodate new and legacy study databases, comply with emerging standards for ecological information, and enable dynamic discovery and access to multiple information products over the Internet.

Keywords: information management systems, ecological metadata, information access, data archive, ecoinformatics, Long-Term Ecological Research (LTER)

INTRODUCTION

Information Management Systems operating at various complexities and on multiple computer platforms have been used to manage the environmental databases residing within the Forest Science Data Bank (FSDB) for nearly 25 years. The FSDB was established to house data generated by participating scientists in the National Science Foundation's (NSF) Long-Term Ecological Research (LTER) program at the Andrews Experimental Forest site as well as contributed data sets from other collaborating researchers [1], [2]. Dedicated to the long-term preservation and availability of environmental databases, the FSDB features a rich and diverse repository of data and metadata for over 250 ecological studies [3]. This rich legacy of long-term databases and accompanying metadata poses significant challenges when new requirements necessitate changes to the Information Management System. Transition requires careful consideration and evaluation of new computing technologies, choice of computer platform and software, researcher and client needs, standards for ecological information, and existing system requirements. Consideration of both personnel and financial resources is also critical in determining the scope of the new system and a timetable for system implementation. In the case of the FSDB, limited resources coupled with the quantity of legacy information have resulted in a three-year-plus transition period from initial planning to complete implementation.

HISTORY OF FSDB INFORMATION MANAGEMENT SYSTEMS

Advancing information technologies coupled with scientific demand for easy discovery, access, and integration of research study databases have led to multiple evolutionary stages of the FSDB. From an early mainframe tape library to a PC-based metadata system on a Local Area Network (LAN) to the employment of more powerful tools such as Relational Database Management Systems (RDBMS) on high-speed database servers, the FSDB has evolved with computing technology, researcher demands, and emerging new standards for the management of information. Table 1 summarizes the primary developmental stages and Information Management Systems employed by the FSDB.

The need to manage scientific information arose with early data collection efforts at the Andrews site by the U.S. Forest Service Pacific Northwest Research Station (PNW) beginning in the 1950's followed by the International Biological Program (IBP) in the 1970's, and the LTER program beginning in 1980. The IBP efforts focused on the development of documentation forms to capture critical study abstract and data set description information, and set the stage for the formal creation of the FSDB. The first information management system was established in 1981 with the consolidation of mainframe computer data files into a magnetic tape library, and the development of the first FSDB database catalog [4]. An automated bookkeeping system was used to track the storage requirements and documentation status of study databases and computer programs, and an automated data retrieval system was installed. The interactive retrieval system allowed researchers to obtain study databases from mainframe magnetic tapes and provided security from unauthorized use of the data files. This increased data security and tape backups, as well as significantly lower mainframe storage costs, provided strong incentives for researcher participation in the FSDB. A data verification system allowing two different persons to independently enter each data set and alert the second operator of discrepancies initiated data quality assurance.

Table 1. Primary developmental stages and Information Management Systems employed by the Forest Science Data Bank (FSDB) including the Andrews Experimental Forest LTER site over the past 25 years.

Period	**Platform**	**Metadata storage**	**Data storage**	**Primary tool**
1980's	Mainframe	Paper forms	ASCII	Fortran
Transition period 1988-1991				
1990's	Local Area Network (LAN) File Server	Desktop RDBMS	ASCII/ Desktop RDBMS	SAS/ Desktop RDBMS
Transition period 1999-2002				
2000's	Database Server/ UNIX-based Web Server/ PC-based Web Server	RDBMS Server	RDBMS Server	SQL Server/ Desktop RDBMS

While the FSDB provided an advanced system for this era, change was inevitable with the common occurrence of personal computers and powerful desktop software. Seeing the limits of further extensibility to the mainframe system, FSDB personnel moved the data from the mainframe tape library to a PC-based LAN and housed the metadata in a desktop RDBMS. Central database catalogs and standard metadata tables for each individual study database formed the basis for a quality assurance system and other generic data production tools such as writing data documentation and error reports, automatic creation of data entry forms, seamless data import and export (ASCII<-->RDBMS), and eventually automatic webpage creation for study data [5]. This new system was a vast improvement over the original system while still providing strong incentives for participation and preserving the positive features of the earlier system. In particular, improvements included the quality assurance system, which eliminated a major deficiency of the original system, the automation of paper copy metadata, and the ease of local access to the LAN-based system.

STIMULI FOR CHANGE

This information system gave stability to the FSDB throughout the 1990's, was invaluable in the improvement of data set quality, and proved to be extensible to the introduction of new web technology to accommodate an LTER mandate in 1994 to put research databases online. However, web-database applications were still in their infancy, and this original approach to distribute pre-positioned metadata and data files for downloading introduced new redundancies and a workload related to updating these static system files whenever changes in the underlying databases occurred. The need for planning a web interface for dynamic compilation of metadata and data and for a high-performance RDBMS to replace the desktop DBMS for storage and delivery functions was understood.

Compliance with emerging metadata standards for ecological data, [6], [7], [8], [9] [10], provided another strong signal to undertake a system redesign and include additional metadata elements. Other known flaws included pervasive redundancies in personnel data, keyword lists, site descriptions, and attribute and domain descriptions, which existed in asynchronous versions. Bibliographies, spatial data, and personnel were all maintained in separate, stand-alone structures without the ability to establish connections between them, and items such as keywords or people associated with research projects always existed but were not suitably structured for productive searches. While this legacy system continued to provide a strong tool for managing conventional databases, the introduction of an expanded ecological metadata content standard, and user demands for easy discovery and access to databases, publications, and other kinds of information, offered a unique opportunity to develop a more comprehensive and better-integrated information management system.

DEVELOPMENT OF AN INTEGRATED ECOLOGICAL INFORMATION SYSTEM

Spatial data, research publications, models and software, collections, maps, images, photographs, grants, assorted documents, and, the latest arrival, web content, have been added in recent years to the suite of objects to be covered by a scientific information system originally geared exclusively to managing conventional databases. Access to most of these objects depends on convenient and efficient searches of their shared domains of keywords, people, places and species. As these shared domains apply to databases as well as all the other products, it seemed reasonable to assemble information products and associated domains into a single extensible system.

The initial step was designing the new system schema by organizing metadata content into a normalized structure (Figure 1). The content generally conforms to new metadata standards and the design integrates the various information components. Normalization removes all model structures that provide multiple ways to know the same fact, and is a method of controlling and eliminating redundancy in data storage [11]. The design allows databases, publications, and other components to share the common domains of people (as well as projects, organizations, and funded grants), keywords (theme, place, and taxonomic), place descriptions, taxonomic systems, and even enumerated domains of data set attributes. The design grew naturally from a catalog of existing information products and the shared domains that serve to classify the products.

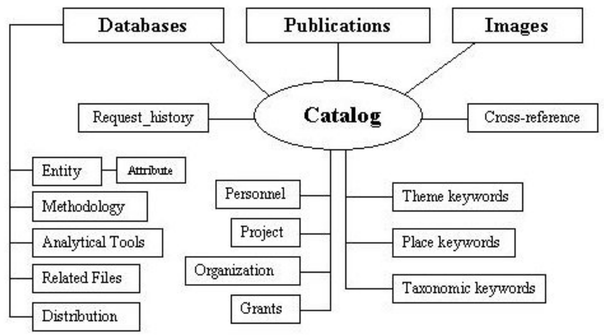

Figure 1. Simplified view of the normalized metadata structure.

The metadata system was originally designed to comply with described ecological standards for metadata [6], [7], [8], but design modifications were necessary to comply with the newer emerging NBII Biological Data Profile [9] and the Ecological Metadata Language (EML) [10]. This "moving target" for metadata standards certainly complicated the design process, but the resulting structure accommodates all of the leading ecological standards as well as existing FSDB metadata. In general these standards have provided content specification with no guide for their construction in relational databases. The modular structure of EML (implemented in eXtensible Markup Language (XML)) would have been useful in this regard, but emerged too late in the design process to be of significant help. Dynamic generation of EML by "cross-walking" corresponding FSDB metadata elements into the content standard is planned for EML compliance. In particular, applications currently under development [12] are being explored to generate native XML from the FSDB relational metadata schema and employ the XML Stylesheet Language (XSL) to map the native XML into EML module elements.

The new metadata standards represent an expansion to previously existing FSDB metadata and new high-level, or data abstract elements, and spatial data elements were added for compliance. Conversely, site-specific metadata elements not included in the standard content exist in the system structure for purposes of local management including database request histories, funded grants, and research theme classifications. Additionally, elements to capture user feedback, review history, and quality assurance and progress reports for study databases are included to represent more subjective aspects of database quality beyond the more explicitly described standard elements.

DESIGN AND IMPLEMENTATION ISSUES

The following discussion addresses selected topics of general interest that emerged during the design phase and as the implementation of the system and its associated web interface progressed. These topics include aspects of metadata structure, transition problems, and some production and maintenance issues for which realistic solutions are presently lacking.

Personnel. Personnel and associated tables were implemented first, since all components are associated with people and specific roles (e.g., author, investigator, data set contact, etc.). Inclusion of all publication authors and study database investigators provide a basis for searching these products by a person's name. Additionally, key information is maintained for Andrews LTER personnel, and applications are provided to dynamically generate personnel web pages and create local mailing lists. A web interface was developed allowing all LTER members to update this personal information.

Publications. All Andrews LTER publication metadata were imported from desktop bibliography software. This is problematic, as the RDBMS system does not currently provide equivalent features and enhancements, thus accommodations were made for importing and exporting into this bibliographic software. While this process of importing from other software is not seamless, inclusion of the bibliography in the information system allows searches by personnel and keywords, affords direct linkage to online publications, and provides connections to study databases directly related to a publication.

Theme keywords. A controlled vocabulary of preferred keywords for publications and databases was developed by a committee of local scientists and structured hierarchically. Non-preferred keywords are also listed with links to preferred keywords as a way of maintaining legacy keywords. Hierarchic structuring of keywords imposes additional maintenance overhead when adding new keywords, but also provides for improved search capability over the simple list. While the development of a specific controlled vocabulary was time consuming, existing vocabularies (e.g., Global Change Master Directory [13]) were found to be too general for local purposes.

Place keywords. Study sites often encompass multiple projects and therefore databases and/or publications.

Place keywords reside in sharable entities that will provide description of study sites, searches for products within a given set of bounding coordinates, and links to spatial data entities. Places may be classified as a type of place (e.g., meteorological stations, reference stands, research natural areas, etc), have attributes that reflect their specific nature, and serve as domains for database attributes of site codes. Hierarchic structuring of place keywords was rejected as placing was difficult with many arbitrary overlaps occurring among geographic and administrative units.

Taxonomic keywords. Like theme keywords, taxonomic keywords form another hierarchy. Although these are unambiguous, they remain editable as reclassification can occur. Generally the taxonomic lists only include taxa for groups of organisms that occur in our databases, and the table serves as a quality assurance check for attributes with species code domains. These lists also form the basis for searching for relevant publications and databases. However, multiple sets of codes are in use for individual taxonomic groups, and we have imported the list of all Pacific Northwest plant taxa from the USDA plants database [14], structured it hierarchically, and merged it with our local taxonomic reference [15]. This allows the support of both the new USDA plant codes and the Garrison codes that appear in the databases. Updates from national species lists will have to be done periodically, but eventual dynamic use of national systems such as the Integrated Taxonomic Information System (ITIS) [16] might be possible to provide a common framework for taxonomic data.

Study database metadata. The choice of normalizing metadata dictates a single table that lists all distinct *attributes* of all databases in the system. The system was originally structured to allow sharing of attributes among study databases. In practice shared attributes turned out to be fairly rare and they impose a significant maintenance overhead. The system does support sharing of attributes among tables within a study database. Additionally, *enumerated domains* of attributes are sharable across databases, but as with shared attributes, the incidence of code sets shared across databases is fairly low. Similarly, all distinct *methodologies* (e.g., field, laboratory, statistical, processing procedures) reside in one single table and are shareable across study databases. Additionally, methods can be described and shared at both the data abstract level, or more specifically for attributes.

The insertion of database metadata into the new framework has proved to be a formidable task, but also an opportunity to review, expand, and better organize critical database documentation. To accomplish this move, a special application was developed in the desktop RDBMS to allow assembly, editing, and reassignment of study metadata into the appropriate content elements for the new metadata system. For each study database, programmed inserts and remote views from the desktop to the database server were used to populate the new system's metadata tables. The ability to use the existing desktop RDBMS as a front-end to the database server was essential in this transition process.

Study data. Similarly, porting study data into the database server afforded an opportunity to examine the structure of the individual databases and restructure as needed. The comprehensive quality assurance system [5] was run before uploading to assure transfer of the cleanest possible data. This quality assurance system and other existing procedures are also being adapted, as generically as possible, into the new system primarily using the desktop RDBMS with remote connections to the database server. The generic production tools featured in the previous information systems will be adapted or redeveloped to ensure the continued use of metadata as the basis for both production and distribution of ecological data and with the perspective to minimize the need for data set-specific programming.

The tabular data for all FSDB study databases are stored as individual tables by entity in a single database, separate from the metadata database. The data tables are generally maintained in a semi-normalized state reflecting the "data sets" as produced and used by scientists. Although simplifications and efficiencies can be gained in restructuring the study databases, cost considerations have thus far prevented full data normalization due to the sheer number of legacy data sets.

Metadata and data as well as other information products are obviously connected. As a production issue this implies that in addition to quality control for metadata and data, the system should provide a mechanism for ensuring that metadata and data are congruent. For example, changes in an attribute's length, nullability, or enumerated domain, should not invalidate the integrity of the metadata in describing the actual data. A metadata-driven quality control system is helpful in this area but falls short of guaranteeing the integrity of the data-metadata whole. An obvious solution would be to manage databases through their metadata adding another layer of complexity to a hypothetical, full-fledged production interface, and is only under consideration at this time.

While both *non-spatial and spatial data* will reside together in the high level FSDB database catalog, we have yet to find a way to provide seamless programmatic connections between them. Proprietary Geographic Information System (GIS) databases and software are now resident on the database server, but metadata is managed autonomously. Many attributes of the tabular (non-spatial) data are associated with GIS spatial layers, but are not documented within the GIS. Compiling metadata for a spatial database that includes tabular entities will require a way of merging metadata from both systems. One possible solution might be creating the compilation in the XML-based EML again using XSL to map the XML-based GIS metadata elements into EML. Additionally, design provisions allow database searches, including searches by spatial coordinates, to return appropriate GIS layers as well as associated tabular data.

Web interface. *Internet access* to the Andrews LTER (http://www.fsl.orst.edu/lter) now provides FSDB databases and information through dynamic web applications. Various mechanisms for searching for publications and data are provided. The database server is used to manage, maintain and track the LTER web pages through two database tables that also serve as the basis for the web site map and web search engine. One table includes comprehensive documentation for all web pages and controls the origin of page content, the page display template, page images, page author, title, meta-tags, and dates. Another navigation table controls the side and top navigation panels, navigation text, display elements, and web page URL's. Control of the website through the database, along with the use of navigation and page templates, improves the ease and efficiency of maintenance.

The *data distribution system* has been rebuilt entirely to support searches and dynamic web access to the study data and metadata. Metadata web pages are created from web application programs using RDBMS stored procedures. Metadata output in EML structured modules has been successful in limited testing and is planned for future implementation. Comma-delimited datasets are also dynamically created for every entity within each study database. Users are requested to complete a one-time only registration form that will allow them to login and have free access to all available data. Users, intended purpose, and instances of data download are automatically tracked into a request history table within the system.

The creation of a robust system for metadata entry and editing is one of the more difficult tasks remaining. Given multiple databases and multiple owner-curators with varying skills, the interface must be highly generic and assume a low skill-level of users. Interface features to ease the burden of providing metadata are essential. Examples might include choices of personnel and keywords from drop-down lists, or selection of sharable descriptions of study sites or methodology. However, certain databases, especially long-term databases, invariably require special features.

LESSONS AND CONCLUSIONS

In our experience, the planning, design, development, and implementation of an Information Management System may take years to accomplish given limited resources and depending upon the new system's scope and complexity. [Note that this latest transition of the FSDB required significant time of two permanent staff members and was supported by three $25K NSF supplemental grants, which provided contract personnel, hardware, and software.] This task is complicated by the need to maintain and support the existing system, and the transition to full implementation tends to become a stepwise process as modular aspects of the legacy system are replaced. System enhancements continue after implementation before the legacy system can be completely dropped and a period of relative stability can begin. Even periods of stability require considerable maintenance, upgrades and occasional design changes as new technologies emerge and system requirements change. For example, the emergence of new technologies such as those recommended by the World Wide Web Consortium [17] (i.e., XML, XSLT, XPath, XMLSchema) have already altered our thinking on the presentation, export, and exchange of metadata.

The selection of technology and software is constrained to "mature" tools considering the size of the FSDB and available resources. The selection of relational database software capitalizes on existing expertise and preserves the continuity of many of the existing data production tools. The ability of the desktop RDBMS to communicate through remote connections with the full-featured RDBMS server made the task more efficient and allowed easier adaptation of existing system features. The selected RDBMS server and web server are also compatible with long-term plans of the larger research enterprise enabling sharing of costs and staff.

One of the primary goals of this development effort is to improve integration of previously disparate information sources, and utilize a web interface to realize this potential for new discovery. This is illustrated by the ability to do a keyword search for a database, discover all related publications, directly link to those publications or other related files or websites, and link to pertinent personnel biographies. In considering the extension of the metadata model to multiple information products, limiting the metadata system exclusively to databases does not significantly decrease structural complexity or make the transition any easier. The inclusion of other managed information products added essentially no complexity other than the content tables and their relations to the shared domain tables, and greatly improves the integrated nature of the model.

This transition to a new information system provides a unique opportunity to evaluate legacy data and metadata, and in some cases "deactivate" study data with questionable quality or documentation. Metadata content is reviewed for accuracy, reassigned for consistency, and in many cases improved with newly edited abstracts. Study data is also reviewed and in some cases restructured to normalized forms. However, the reality of metadata production in contrast to stated needs of information delivery is another difficult issue. A good example is connecting databases or publications with grants (or publications with databases) implemented with a very simple table in the metadata database model. Capturing existing relationships and ensuring continued, reliable maintenance are still problematic. Given the complexity of metadata content standard, it is critical that the information system be designed with user needs and requirements in mind, and that in return is supported through long-term research planning.

The Information Manager cannot simply reside in the trenches battling data sets independently from the larger research enterprise. The need for collaborating with the research scientists, offering rewards for cooperation, and providing mechanisms for the broader group to help share workload has often been discussed [18], [19]. The move to a more controlled RDBMS environment, together with an ambitious new metadata standard, have made these goals, if anything, more elusive. Typically, high quality study data and metadata are achieved only through the diligent efforts of the data provider or a conscientious data manager. The entry and maintenance of metadata by databank users (i.e., researchers, graduate students, and other data providers) remains limited, and the challenge of establishing a robust metadata interface is daunting. Data production tools supporting quality assurance and web publishing of data sets will be necessary incentives for research scientist participation. Data access and distribution on the other hand have proven to be positive benefits of the new system. The interoperability of metadata content exported to the EML standard should also offer considerable value through discovery of information resources and sharing of general tools in support of ecological science.

ACKNOWLEDGEMENTS

The USDA Forest Service, Pacific Northwest Research Station and the National Science Foundation Grant DEB-9632921 provide support for this work. The authors wish to acknowledge the contributions of James Tucker in the design and implementation of the relational metadata database.

REFERENCES

[1] 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.

[2] Stafford, S. G.; Spycher, G.; Klopsch, M. W. 1988. Evolution of the Forest Science Data Bank. Journal of Forestry. 86(9): 50-51.

[3] Henshaw, D. L.; Spycher, G. 1999. Evolution of ecological metadata structures at the H.J. Andrews Experimental Forest Long-Term Ecological Research (LTER) site. In: Aguirre-Bravo, Celedonio; Franco, Carlos Rodriguez, eds. North American science symposium: toward a unified framework for inventorying and monitoring forest ecosystem resources; 1998 November 2-6; Guadalajara, Mexico. Proceedings RMRS-P-12. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station: 445-449.

[4] FSDB staff. OSU Forest Science Data Bank Newsletter. 1981. Department of Forest Science, Oregon State University. 2p.

[5] Spycher, G.; Cushing, J. B.; Henshaw, D. L.; Stafford, S. G.; Nadkarni, N. 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.

[6] Porter, J. H.; Henshaw, D. L.; Stafford, S. G. 1997. Research metadata in Long-Term Ecological Research (LTER). In: Second IEEE metadata conference; 1997 September 16-17; Silver Spring, MD. [Online]. Available: http://computer.org/conferen/proceed/meta97/list_papers.html [1999 February 2].

[7] Federal Geographic Data Committee (FGDC), USGS. 1998. [Online]. Available: http://www.fgdc.gov/fgdc/fgdc.html [2002, March 7].

[8] Michener, W. K.; Brunt, J. W.; Helly, J. J.; Kirchner, T. B.; Stafford, S. G. 1997. Nongeospatial metadata for the ecological sciences. Ecological Applications. 7(1): 330-342.

[9] NBII Biological Data Profile. 2001. [Online]. Available: http://www.nbii.gov/datainfo/metadata/standards/index.html [2002 May 17].

[10] Ecological Metadata Language (EML). 2001. [Online]. Available: http://knb.ecoinformatics.org/software/eml/ [2002 May 9].

[11] Date, C.J. 2001. An introduction to database systems. 7th ed. Addison Wesley.

[12] Center for Environmental Studies, Arizona State University. 2002. [Online]. Available: http://caplter.asu.edu/bdi/ [2002 May 17].

[13] Global Change Master Directory. 2002. [Online]. Available: http://gcmd.gsfc.nasa.gov/ [2002 May 17].

[14] USDA, NRCS. (2001). The PLANTS Database, Version 3.1. [Online]. Available: http://plants.usda.gov/ [2002 April 5].

[15] Garrison, G. A.; Skovlin, J. M.; Poulton, C. E.; Winward, A. H. 1976. Northwest plant names and symbols for ecosystem inventory and analysis. 4th ed. Gen. Tech. Rep. PNW-46. Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest. 263 p.

[16] Integrated Taxonomic Information System (ITIS). 2001. [Online]. Available: http://www.itis.usda.gov/ [2002 May 17].

[17] World Wide Web Consortium (W3C). 2002. [Online]. Available: http://www.w3.org [2002 May 17].

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

[19] Porter, J. H.; Callahan, J. T. 1994. Circumventing a Dilemma: Historical approaches to data sharing in ecological research. In: Michener, W. K.; Brunt, J. W.; Stafford, S. G., eds. Environmental Information Management and Analysis: Ecosystem to Global Scales. London: Taylor & Francis: 193-202.