SCI2002: Invited meeting sessions with papers on LTER information management

Title: The Ecoinformatics Challenge: Meeting Ecological Information Needs for the Site, Network, and Community
Organizers: Susan Stafford, John Porter, and Karen Baker for the LTER Information Management Committee
Meeting: International Institute of Informatics and Systemics; SCI2002
Dates: July 14-18, 2002
Location: Orlando, Florida, USA

The Ecoinformatics Challenge: Meeting Ecological Information Needs for the Site, Network, and Community – Invited Sessions

1. Baker, Karen S.; Brunt, James W.; Blankman, David (USA): Organizational Informatics Site Description Directories for Research Networks. p355-360

2. Bayard Cushing, Judith; Nadkarni, Nalini; Healy, Keri; Ordway, Erik; Delcambre; Lois; Maier, Dave (USA): Template-driven End-User Ecological Database Design. p361-366

3. Brunt, James W.; McCartney, Peter; Baker, Karen; Stafford, Susan G., (USA): The Future of Ecoinformatics in Long Term Ecological Research. p367-372

4. Henshaw, Donald L.; Spycher, Gody; Remillard, Suzanne M. (USA): Transition from a Legacy Databank to an Integrated Ecological Information System. p373-378

5. McCartney, Peter H.; Jones, Matthew B. (USA): Using XML-encoded Metadata as a Basis for Advanced Information Systems for Ecological Research. p379-384

6. Melendez-Colom, Eda C. (Puerto-Rico); Baker, Karen S. (USA): Common Information Management Framework in Practice. p385-389

7. Michener, William K.; Brunt, James W.; Vanderbilt, Kristin L. (USA): Ecological Informatics: a Long-Term Ecological Research Perspective. p390-395

8. Porter, John H.; Ramsey Jr, Kenneth W. (USA): Integrating Ecological Data Tools and Techniques. p396-401

9. Sheldon, Wade M.; Moran, Mary Ann; Hollibaugh, James T (USA): Efforts to Link Ecological Metadata with Bacterial Gene Sequences at the Sapelo Island Microbial Observatory. p402-407

10. Smith, Dan J.; Benson, Barbara J.; Balsiger, David F. (USA): Designing Web Database Applications for Ecological Research. p408-413

11. Stafford, Susan G.; Kaplan, Nicole E.; Bennett, Christopher W. (USA): Through the Looking Glass: What do we see. What have we Learned, What can we share? Information Management at the Shortgrass Steppe Long Term Ecological Research Site. p414-419

12. Vande Castle, John; Pennington, Deanna; Fountain, Tony; Pancake, Cherri (USA): A Spatial Data Workbench for Data Mining, Analyses, and Synthesis. p420-424

the Shortgrass Steppe Long Term Ecological Research Site. p414-419

Organizational Informatics: Site Description Directories for Research Networks

Karen S. Baker
Scripps Institution of Oceanography, University of California San Diego
La Jolla, California USA 92093-0218 USA

and

James W. Brunt
Department of Biology University of New Mexico
Albuquerque, New Mexico 87131-1091 USA

and

David Blankman
Department of Biology, University of New Mexico
Albuquerque, New Mexico 87131-1091 USA

ABSTRACT

A site description directory plays a central role as a catalog for a network of research sites. Such a directory represents a keystone element in an information management system. A directory contributes to community communications both through documentation of member information and relationships as well as through design feedback elicited from participants in the ongoing process of developing the catalog system. Presentation of a description directory for networked research sites via web interfaces permits distributed, remote site data input and access. There is a dual challenge in creating an extensible directory design: first to capture relevant content and second to incorporate such a system within the work practice of the community represented to ensure its continued evolution. We present here a directory designed for the Long-Term Ecological Research (LTER) Network community

Keywords: directory, metadata, organizational informatics, system design

INTRODUCTION

To describe, classify, and catalog are tasks fundamental to science. A site description directory extends the traditional catalogs of collections of objects (e.g., butterflies or rocks) and of data sets (e.g., temperature or biomass) to catalogs of systems such as research sites and networks. Although there is a growing understanding of the concept of data richness in the field of ecology today, there is less

Table 1: System and Sub-System Levels

Level	System	subsystem	Example	Organizational Principal
I. Cooperation	Individual	Data1...Data N	Researcher A	Related data
	Site	Data set1...Data set N	Site researchers A & B	Related projects
	Cross-site	Site1...Site N	LTER sites A & B	Related data sets & project
II. Federation	Discipline	Network1...NetworkN	LTER, NADP, OBFS	Related themes
	Domain	Partner1...Partner N	OBFS, NEON	Related domain
	Cross-Domain	Discipline1...Discipline N	Ecology, Earth Science	Related system

| II. Federation | Discipline | Network1...NetworkN | LTER, NADP, OBFS | Related themes |

recognition that each research site and network is information-rich at an organization level as well.

Like the "experimental unit" in statistics, the "atomic" unit of ecological information management is the data set. A data set often can be presented as a table and its associated metadata. The content of data sets in ecosystem science ranges in complexity from measurements of daily temperature at a particular location to measures of diversity for a particular ecosystem. A table entity is an abstracted description, either by direct measurement or by derivation of some aspect of a physical entity or interaction of physical entities. A site or network is also an entity acting as a source of data or the subject of metadata. Much of the data in this directory constitutes the metadata describing research programs. In this work, "site" is used to describe a research team comprised of some combination of individuals united by a common study through an information and social structure. A site may be a member of a network or an association of sites.

The LTER directory design builds upon the structural similarity and significance of systems associated with both subsystems and larger systems (see Table 1). A multi-tier directory schema of metadata assumes sampling regions associated with a Site, a Site as a member of (related to) Networks, and Networks as related to Federations (Sheth and Larson, 1990).

A prototype **site description directory** for research sites has been designed for use as a module within the Long-Term Ecological Research Network Information System (NIS, Baker et al, 2000; Brunt 1999). The working model gathers and displays descriptive site data in addition to responding to user queries. It is a multi-level (network, site, subsite) iterative schema with attention to portability. The current research network **site description directory** is a two-tier implementation with a centralized relational database back-end and web-based user interface for data input, modification, display and comparison. The data model is relational with some object-relational aspects and with categories and themes identified by scientific participants themselves.

DESIGN ELEMENTS

A site description directory provides answers to questions such as

"What biomes do the sites represent"
"What are the locations of the polar sites?"
"How large is the forest site?"

The needs met by a site description database include creating a repository of information that can be queried for a single site or across multiple sites, delivering easily accessible views of information in a common format and providing a mechanism for participant management of local information (add and/or modify). The choice of directory content material is important; it provides a common template that defines a site. Since local site definitions can be established independent of a network catalog, a site description directory can enhance organizational identity by making basic information about an association of sites available without detracting from local site autonomy.

Existing catalogs provide examples of working directory models. The initial LTER Network approach provides a list of links to each member's web page where content presentation is independent (member in this context means the site organization, not each individual associated with the network). The National Atmospheric Deposition Program (NADP; http://nadp.sws.uiuc.edu/nadpdata/), with more than 200 network participants, and the Organization of Biological Field Stations (OBFS; http://www.obfs.org/Members/StationList.html), with more than 400 research field stations, represent a range of technology capabilities. Each met the challenge of diversity by compiling responses to an online form requesting organizational information that is maintained in a centralized location and so is ultimately queriable. Currently, in partnership with the LTER Network Office and the National Center for Ecological Analysis and Synthesis, OBFS is in the process of moving from a static to a dynamic presentation of information.

The structure identified as meeting both LTER directory needs and design criteria consists of three interrelated categories of member information: organization, personnel, and descriptive material (see Figure 1) organized into tables and including category or look-up lists. Information about

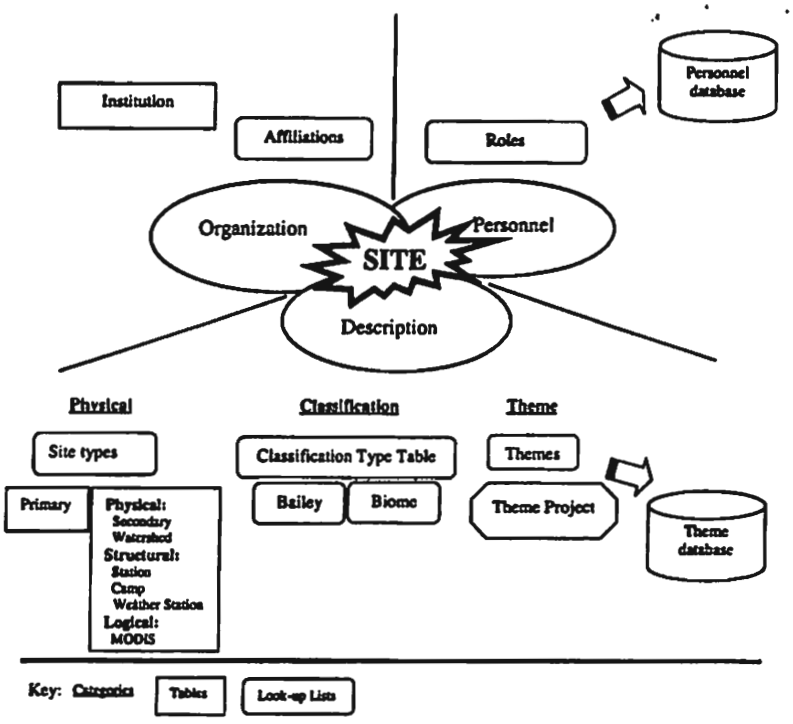


Figure 1. Site description directory components.

participating institutions such as their address and contact details along with their affiliations is presented in a directory as is personnel information including participant names, locations, email addresses and roles. Within the descriptive materials are subcategories of physical descriptors (e.g., latitude and longitude), classification tables (e.g., biome types) and theme lists (e.g., regionalization).

DESIGN APPROACH

The original design focused on simplicity. The data content represents a research site's general context. Discussions were held to engage and elicit the different participant views from within the organization so that the initial list of parameters could be shortened without losing critical information (or support). Information was eliminated that was either too detailed or too site specific in order ultimately to maximize site participation by minimizing time needed to complete forms. An early extension of the project was the addition of uniform resource locators (URL) in order to have available links to harvestable materials such as climate database files, site photos, and/or site maps.

Web based forms provide an interface for viewing and comparing entries. The main view, or site view, is the root module. Further views may be topic specific such as climate, vegetation, regionalization, and soils. Categories are defined to use emergent classifications such as the Terrestrial Ecological Monitoring Stations (TEMS), a pilot project of the Global Terrestrial Observing System (GTOS). The Profiles of Ecologists (1992) contains a survey of the membership with a classification scheme for fields of ecological research and areas of expertise; the categories are empirical and include research themes as well as land types. Three variables are included for site classification description including the Bailey ecoregion types (Bailey, 1998) included in the TEMS/GTOS database. The biome group (i.e. tundra) and region classification (city, state, upper Bavaria regionalization) schema definitions require further definition as does the physical classification system developed through multi-site discussions.

Multiple views are available from the descriptive component information: single member information (home view), selected information from multiple sites (element view) or participant information on selected topics (theme view). The ability to compare parameters across sites demonstrates the value of a database approach.

available links to harvestable materials such as

sites (element view) or participant information on

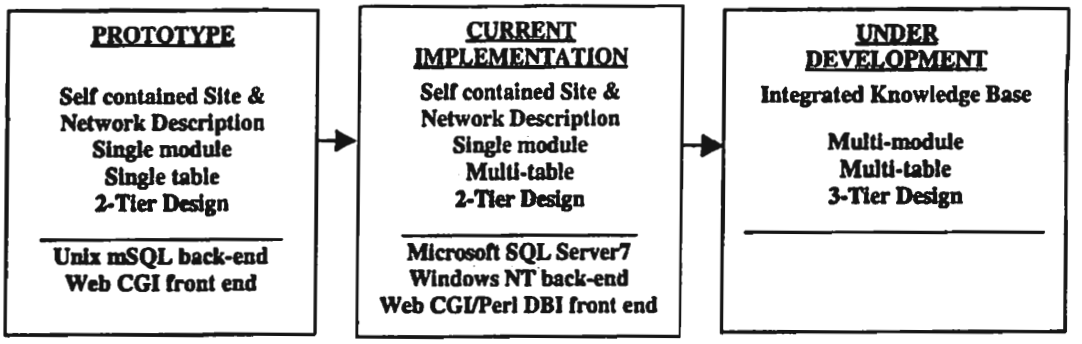


Figure 2. Three stage design development.

A set of web-based forms is used for adding, modifying and deleting from the directory. The entry forms are divided by view (theme): Site, Climate, Vegetation, Regionalization, and Soils. The modular entry forms approach breaks the information down into more easily digested units. A new subsite or theme may be added through the addition of a new table. Currently, the input tables include the site table with two forms or categories (description and URL), subsite tables with three categories (location, class, abiotic parameters), and research theme tables with a single category. Data entry is possible by designated site personnel with the site username/password login capability.

Design/Implementation Lineage

The design lineage of this project is outlined in Figure 2. The initial prototype design was driven by the design principle of simplicity, sacrificing flexibility, extensibility and modularity. The current implementation adds flexibility and some modularity at the expense of simplicity in order to broaden and deepen content. The current design is not easily portable with MS SQL Server 7 specific SQL calls along with mixed-case field and table names. While SQL Server (and Windows in general) is not case sensitive, the common, high-level programming language PERL (www.perl.com) is. Thus modification is required if the back-end is ported to Oracle or to non-windows platform.

The current implementation is also not interoperable with existing LTER database modules, specifically the LTER personnel database. For example, the current LTER personnel database uses an a Alphanumeric key to uniquely identify each person entered in the database. The directory database, however, uses a numeric key to identify people.

Current/Future Development

The design principles guiding current development efforts are:

Interoperability:

- Outward focus: Facilitate exchange with other network systems by using current and emerging metadata standards such as the Ecological Metadata Language (EML) for environmental data (http://www.informatics.org) and the Federal Geographic Data Committee (FGDC) standards for geospatial data (http://www.fgdc.gov).
- Inward focus: Develop XML web services to enable two-way communication between the centralized database and individual site databases that store similar information.
- Technical focus: Develop the ability to use different types of data stores (RDBMS, LDAP, XML Native) and exchange data with other metadata management systems such as the Metacat as part of the Knowledge Network for Biocomplexity (Jones et al, 2001).

Extensibility, flexibility and modularity:

- Extensibility: Develop the ability to use this information system for networks other than LTER or for sites within LTER that are also part of other networks.
- Flexibility: Develop the ability to add classification or descriptor systems from participant feedback without having to redesign the underlying database schema. This is being accomplished by i) Further normalization of the database schema; ii) Development of hierarchical representations within an RDBMS schema; Increased abstraction of the data model.

- Modularity: The original directory database is now a module in an integrated knowledge base that includes personnel, bibliography, document archive, event and meeting tracking, and grant tracking modules.

Scalability and portability:

- Portability: A three-tier design will be used to separate the user interface from the back-end data storage system. The system will be designed as a web service. The current plan is to use a J2EE compliant middle-tier using java data-typing, generic (platform independent) SQL calls with business rules defined in XML documents (Muench, 2001; OTN, 2002). This will allow the back-end storage system to be moved to a different database vendor or operating system without the need to modify the code on the user interface.
- Scalability: A three-tier system using lightweight data access objects makes the least demand on network and system resources. The platform independence discussed earlier also makes scalability possible.

Redesign is activated through testing of the site directory prototype in order to gain insights from user feedback. User feedback to date suggests addition of description information such as site directions, a biodiversity theme and capture of update times in addition to development history. Note, each addition preserves a bit of an organization's history.

Testing, a major task often neglected in the rush from design to production, requires time and support yet is essential to guarantee database module robustness. At best, practitioners participate in testing to ensure redesign usability and utility. To the extent that such a project incorporates local participation through identification of common information and classifications, organizational definition is enhanced through design anchored in practice.

CONCLUSION

Organizations, associations and partnership present the challenge of presenting member information that is manageable and accessible. A **site description directory** provides a mechanism to gather information about sites within a research network in a common format. The design permits sites to update records, to add tables as needed to an extensible schema, and to store site URLs. The importance of a directory effort is that its impact is immediate for an organization, providing infrastructure cohesion and presenting metadata for query.

A directory enhances group communications. There is an increasing emphasis on partnership science as an approach to conducting science research such as with recent discussions regarding a National Ecological Observatory Network (NEON, `http://www.sdsc.edu/NEON`). Such associations bring requirements for new methods to manage organizational information. We are at a point where both the maturity of communities and of technological tools supports innovation in establishment of organizational infrastructure. Focus on infrastructure development activates communications (Hutchins, 1995; Kies et al; Robbins, 1995). Such an effort provides a method to stimulate system self-definition and to explore cognitive ecosystem concepts (Tomlinson et al, 1998; Schatz, 1993; Star and Rhuleder, 1994) where the term 'cognitive ecosystem' is used to describe the interdependence of a community's distributed knowledge and its social process.

Given the LTER organizational paradigm of participatory governance in addition to a full and synergistic partnership between science and information management (Stafford et al, 1994), the LTER network can serve as a valuable test-bed for considering methods to optimize communication and management through collections of organizational information. The LTER provides an opportunity to consider how a directory design can incorporate elements of social design to enhance the availability of organizational information.

A group of associated research sites comprise a network or a federation when committed to common goals or characterized by interdependence. As routine adoption and predictability are replaced by speed and innovation, associations require dynamic, integrative systems capable of evolving with the discipline. Computer technology, first seen as a provider of powerful computational engines, has evolved to provide methods for data organization and delivery in addition to becoming a potential factor in social change. The emerging concept of **organizational informatics** with a focus on information and communication infrastructure

CONCLUSION

provides an enabling vision in community efforts to transform distributed elements into an integrated system.

ACKNOWLEDGEMENT

Recognition and thanks to the collaborative design and implementation by Weimin Li. NSF Grants OPP-96-32763, DBI-01-11544, and EIA-01-31958 (ksb) support this work as well as DEB-96-34135 and DEB-99-80154 (jwb, db). Emery Boose, Nicole Kaplan, and Susan Stafford provided helpful comments.

REFERENCES

Bailey, R.G., 1998. Ecoregions. New York, Springer-Verlag.

Baker, K.S., B. Benson, D.L.Henshaw, D. Blodgett, J.Porter, S.G.Stafford. 2000. Evolution of a Multi-Site Network Information System: the LTER Information Management Paradigm, BioScience 50(11): 963-978,2000.

Brunt J.W., 1999. The LTER network information system: a framework for ecological information management. Pages 435-440 in C. Aguirre-Bravo, C.R. Franco (eds) North American Science Symposium: Toward a Unified Framework for Inventorying and Monitoring Forest Ecosystem Resources; 2-6 Nov 1998; Guadalajara, Mexico. Fort Collins (CO): US Dept of Agriculture, Forest Service, Rocky Mountain Research Station. Proceedings RMRS-P-12.

ESA Profiles, 1992. Holland, M.M, D.M. Lawrence, D.J. Morrin, C. Hunsaker, D. Inouye, A. Janetos, H.R. Pulliam, W. Robertson, J. Wilson (eds). Profiles of ecologists: results of a survey of the membership of the Ecological Society of America, Washington, DC.

Hutchins, E., 1995. Cognition in the Wild, The MIT Press, Cambridge, MA

Jones, M.B., C.Berkley, J. Bojilova, and M. Schildhauer, 2001. Managing Scientific Metadata. IEEE Internet Computing Sep-Oct 2001. (http://computer.org/internet)

Kies J.K., R.C. Williges, M.B. Rosson, 1998. Coordinating computer-supported cooperative work: a review of research issues and strategies. Journal of the American Society for Information Science 49: 776-791.

Muench, S. and BC4J Development Team, 2001. Simplifying J2EE and EJB Development with BC4J. (http://otn.oracle.com/products/jdev/htdocs/j2ee_bc4j.html)

Oracle Technology Network, 2002. Querying XML in a Standard Way, on the emerging standards for querying XML: SQL/XML and XQuery. (http://technet.oracle.com/tech/xml/)

Robbins, R.J., 1995. Information infrastructure for the human genome project. IEEE Engineering in Medicine and Biology Magazine 14: 746-759.

Schatz, B.R., 1993. Building an electronic community system. Pages 550-560 in Baecker R, ed. Readings in Groupware and Computer-Supported Cooperative Work: Assisting Human-Human Collaboration. San Mateo, Calif.: Morgan Kaufmann Publishers, Inc.

Sheth, A.P. and J.A. Larson. 1990. Federated database systems for managing distributed, heterogeneous and autonomous databases. ACM Computing Surveys 22: 183-236.

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 (eds). Environmental Information Management and Analysis: Ecosystem to Global Scales.

Star S.L., K. Ruhleder, 1994. Steps toward an ecology of infrastructure: complex problems in design and access for large-scale collaborative systems. Pages 253-264. CSCW'94: Transcending Boundaries: Proceedings of the conference on Computer Supported Cooperative Work, 22-26 October, Chapel Hill, NC. New York: ACM Press.

Tomlinson, K.L., M.A. Spasser, J.A. Sanchez, and J.L. Schnase, 1998. Managing cognitive overload in the Flora of North America project. Proceedings of the Thirty-First Hawaii International Conference on System Sciences (HICSS-31)

Template-driven End-User Ecological Database Design

Judith Bayard Cushing, Nalini Nadkarni
Keri Healy, Erik Ordway
The Evergreen State College, Olympia WA 98505 USA

Lois Delcambre, Dave Maier
The Oregon Graduate Institute, Portland OR 97291 USA

ABSTRACT

Historically, ecologists have collected and stored data in individualistic ways, making data sharing among collaborators and subsequent data mining difficult. Integrating database technology into the research process makes data sharing across studies and access to analysis tools easier, but significant barriers to database use often prevent effective use of database technology. In this paper we identify some obstacles to use of information technology by ecologists, including the lack of systems for ecologists to design databases without hiring programmers. We describe a prototype system aka *DataBank* that aims to overcome these obstacles. The key system feature on which we focus is the reuse of domain-specific data types, which we call "templates".

Keywords: Ecosystem informatics, end-user database design, domain specific data structures, spatial databases.

1. INTRODUCTION

Because the collective analysis of data originally gathered by individuals can yield insight beyond a single data set 12,13], many advances in ecology will depend upon effective information management [11]. Database technology will be integral both to the management of ecological information and to the creation of new knowledge. Current database technology appears not adequate to this purpose. A workshop sponsored by the National Science Foundation, the USGS and NASA identified the need for a new "biodiversity and ecosystem (BDEI)" initiative. Noting challenges and opportunities for further research in acquisition, conversion, analysis and synthesis, dissemination of data and metadata (e.g., digital libraries, remote sensing, mobile computing, and taxonomies), workshop participants characterized ecological data and metadata as highly complex – ontologically, spatio-temporally and sociologically. Lack of harmonized protocols, resistance to depositing data and metadata in central repositories, and lack of expertise with informatics tools were also noted as contributing to the limited use of information technology [14] [http://bdi.cse.ogi.edu , http://bio.gsfc.nasa.gov].

The Canopy Database Project [2] is exploring how databases can be integrated earlier (than a final data warehouse) into the ecology research cycle. In this paper, we present the development of a prototype database design tool for a subdiscipline of ecologists. *The Canopy DataBank* provides a vehicle for canopy researchers to design, document, archive, and mine field databases more easily. As with our previous work [5], we emphasized not only physical connectivity, but agreement at the semantic level.

2. OBSTABLES TO ECOLOGISTS' USE OF INFORMATION TECHNOLOGY

In this section, we identify two major obstacles to the use of information technology by ecologists and hypothesize that database systems could help overcome these obstacles. Our ideas revolve around fostering end-user database programming that re-uses spatial database components; we also present design issues arising from this approach.

Although ecologists often consult web-accessible information, they typically enter data into private data stores that are rarely published or archived. Despite increasing pressure from funding agencies, the availability of several excellent ecological data archives [www.lternet.edu , http://www.ecoinformatics.org], emerging tools for recording metadata [20], and even opportunities to publish data in the prestigious Ecological Society of America archives [http://www.esapubs.org/esapubs/archive/archive_main.html], few data sets are published. **Documenting data for archival purposes is still perceived as a time-consuming process** and sometimes not even attempted [6,22]. Furthermore, even once data sets are archived and validated with adequate metadata, **idiosyncratic data representation makes cross study analysis difficult**, even for close collaborators [1]. These two obstacles prevent a semantically viable digital warehouse and global data integration.

Database technology would likely help overcome these obstacles if applied at all stages of the research cycle, just as it has in industry. While the Long Term Ecological Archives and other data publishers make

ontologically, spatio-temporally and sociologically. Lack

even for close collaborators [1]. These two obstacles

excellent use of sophisticated database technology, individual ecologists typically don't have the expertise or inclination to use current database technology, although some who are good programmers use sophisticated statistical programs and GIS, or write complex mathematical models [7,8]. It is not cost effective or even practical for ecologists to hire programmers to design, implement and maintain databases for field data sets. Even if this were the case, without certain **key features: 1) some "controlled" vocabulary, 2) common data structures, and 3) help generating and maintaining metadata**, integrating the resulting databases would still be difficult and costly. We thus decided to produce a prototype that would enable ecologists to program their own databases, and reasoned that experience with (1-3) would lead the way for later productivity tools such as field data sheets, data validation, visualization and analysis to compensate for effort required in database design.

3. *THE CANOPY DATABANK*

DataBank is a web-accessible database system designed for canopy researchers to integrate database technology more easily into their research. We chose canopy science as a representatiave subset of ecology and believe our results are applicable to ecology generally [3]. Goals are to help scientists increase research productivity, simplify sharing data with close collaborators, and facilitate data archiving. Our long term aim is that metadata acquisition be a natural byproduct of the research process, with archiving as easy as pushing the "publish my data" button.

In the Canopy *Databank*, a common vocabulary is fostered by the use of *The Big Canopy Database*, aka **BCD** [http://canopy.evergreen.edu/bcd]. Use of common data structures has been fostered by organizing our database design component around the concept of reusable database components that we call *templates*. Metadata generation and maintenance are being addressed in a separate component out side the scope of this paper. Issues we addressed were:

1. Can we define a sufficiently general set of database primitives (templates) that would be useful to a significant number of scientists to design field databases[1]?
2. How can templates be internally represented, with adequate information for composing several of them into a database design?
3. How can templates be composed into a database design and how can that design be refined across design sessions and used to generate database schema?
4. How can templates be presented to the ecologist in an intuitive interface?
5. Can a complex system such as this be implemented cost effectively as a web application (in our case, with one professional programmer and several undergraduate students)?

We have temporarily set aside issues such as metadata maintenance during the field work, visualization and analysis tools for the scientist, integration and validation of the database into an integrated archive, and cross-study data queries.

Our strategy for making field databases easier to document and comparable is to provide building blocks for database design and to use metadata source tables. Our strategy for the former is to reuse commonly recurring domain-specific data structures (what we call "templates") as building blocks for new databases, for importing data into a warehouse, and for composing cross study queries. When the field database is generated, we generate an access database and a first-cut metadata description of that database. Field databases designed with *DataBank* would be used in "single user" mode on a private workstation during fieldwork and analysis.

DataBank complements, not replaces, existing archives such as the canopy crane site databases and the LTER repositories. It differs from canopy crane databases in that we provide information spanning several sites. We differ from LTER repositories in that we specialize services for one community and provide help in research design and design. Thus, for example, citations are not limited to projects whose data are stored in *DataBank*, but are meant as community-wide references. Because we are compliant with the metadata requirements for data deposition at LTER sites, those who archive in *DataBank* could easily archive at an LTER site.

DataBank Functional Requirements

In this subsection, we describe *DataBank*'s three major functional requirements: field data repository, field database design, and data mining.

DataBank is a data and metadata repository for canopy research projects. It is modeled on the HJA LTER repository, with the capability of searching and viewing study metadata and field data, and downloading data sets for analysis. Security and privacy features allow researcher flexibility: to publish metadata only (no field data), or to make field data available only to selected colleagues, or viewable but not downloadable. Scientists may also forbid release of personal information.

To design and implement a field database, one has to add a project and study, design the study's database, and

[1] We distinguish between a collection of extensible user defined data types (aka *templates*) and a common conceptual data model. We believe templates do not fall prey to problems with a common conceptual data model. Discussion of this issue is outside the scope of this paper.

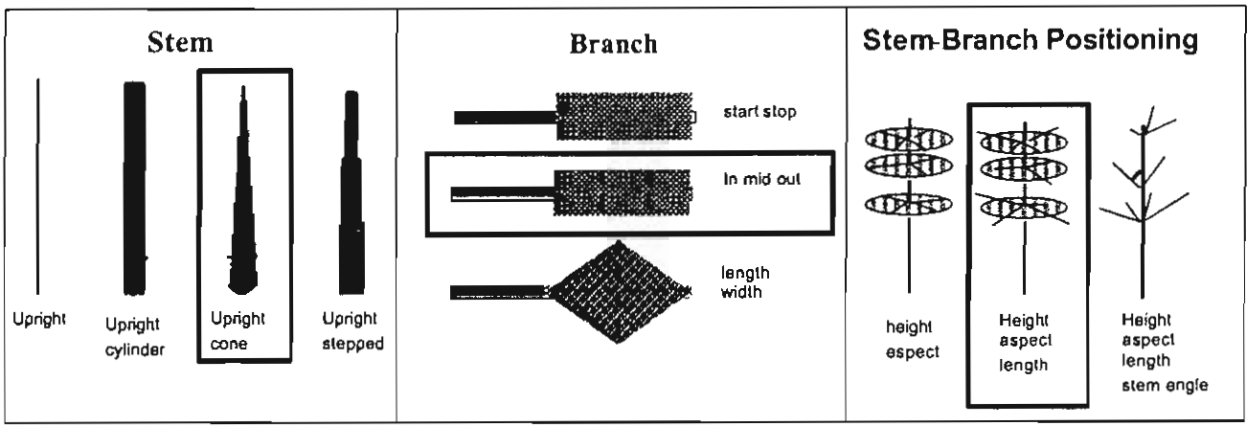


Figure 1. Selecting from a Series of Templates

download that database. To add a project, a *DataBank* archivist registers its Principal Investigator (PI) and creates a new project with that researcher as PI and possibly a second person as project arehivist. The PI or project archivist then adds other researchers to the project and creates studies associated with that project. A study's database is designed by clicking on data templates and associating reference material (such as a species) as source tables. He or she should also be able to view the resulting database, and add or modify attributes and attribute-level metadata such as range, and prepopulate the database with other site or study data. The database should be downloadable as an SQL or Excel database, and a populated field database later uploaded to the repository. Validation against metadata will be aecomplished via software and services from an LTER data center.

For data mining, we currently distinguish repository and warehouse features. The repository includes all field databases, as uploaded by the researcher. Databases in the repository can be searched via study-level metadata, e.g., "find all the studies condueted at the Wind River Canopy Crane Research Facility by Bob Van Pelt", and the contents of those databases viewed individually.

The *DataBank* warehouse, aka the *Master Field Database* (MFDB) integrates field studies into a single warehouse, loading a field database after ehecking it against the templates for schema differences. The MFDB allows cross-study data queries for parts of field databases that match a data template; its schema is the database schema that would be generated using every template. Data in the warehouse can be queried using data templates, e.g., "find the average diameter for trees with height greater than 20 m", or a combination of templates and metadata, e.g., "find the average diameter for trees with height greater than 20 m for studies in the U. S. Pacific Northwest".

DataBank Templates

A template represents data collected when measuring a particular physical object in the real world, e. g., a tree or branch, and appears to users as a conceptual database primitive – a domain-specific data type. Templates usually have absolute or relative spatial attributes. To a computer scientist, templates are collections of variables, each grouped as one or more relational tables, that can be composed into an end-user defined database. When more than one database table is generated, appropriate relationships between the tables are induced. Templates carry table-level metadata that can later be exploited for validation, archiving and query, but are transparent to the end user. Figure 1 presents our conceptual vision of how a researcher might select from a series of templates to produce a database design.

Consider an example where a researcher collects data about epiphytes as percent cover per branch-quadrat for each of several epiphyte species. To build a database for this study, the researcher would use a template for branch-quadrat. Data collected for each quadrat are epiphyte species and percent cover on the quadrat for that species, as well as the date of the observation. Since the data of interest are located on a tree branch, the researcher must also collect information about branches and trees, and hence includes a branch-template and a tree-template in her database. Because each tree must be located in space and that space described, the researcher will include plot- and site- templates.

We have designed templates for site, plot, stem aka tree, branch, and observation variables. The observation template defines generic forest functional observations such as light, temperature, and percent interception of rainfall at a particular location or on a forest entity.

Our current templates are derived from field databases for seven independent studies for which we designed databases. The studies range from those carried out by individual researchers working independently at separate sites to a group of very close collaborators using the same set of equipment and research sites.

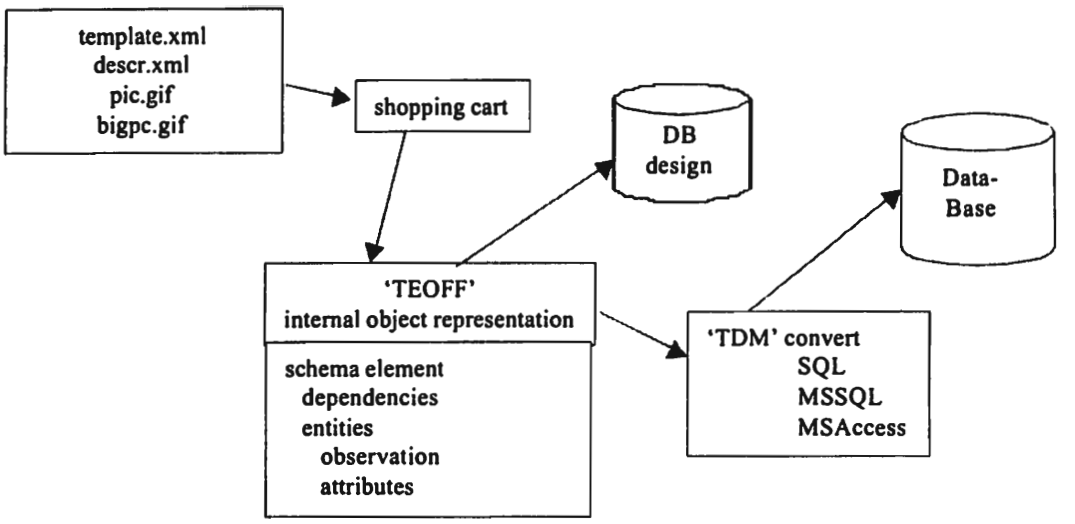


Figure 2 Canopy *Databank* Architecture

***DataBank* Implementation**

A prototype *DataBank* is currently implemented in Microsoft SQL Server, Java, JDBC, and Enhydra. The system currently allows creation of a database from a few prototype templates and the download of that database into an Access Database. We do not yet support refinement of a database design to add or delete attributes, or complex observation templates. We provide a separate metadata documentation tool (in Access and Excel) and a simple browser and viewing for study data that has been uploaded to the archive. Warehouse features are not yet implemented, and upload to the archive is currently by hand.

A template is represented as an xml document, with: 1) descr.xml, a description used by the application to explain a template to the ecologist, and 2) pic.gif, an icon of the template. As a user designs a database, templates are placed into a database design (using a shopping cart design metaphor). The system component *Template Entity Observation Framework* (**TEOF**) combines templates into an internal representation that checks dependencies for a particular template, e.g., a database that includes a tree should also have a species table. The TEOF representation is made persistent (DB Design) so that a user's design session can span several sessions, so that designs can be reused, etc. The *Template Database Model* (**TDM**) converts a TEOFF design into an SQL dialect and generates a database which can be downloaded to the user. Figure 2 shows the DataBank architecture and how the above system components relate to one another.

We identify three implementation issues: template representation, development platform, and effective end-user database design.

(1) Templates are currently represented as XML documents, with annotations about how elements are grouped into tables and how those tables are related. A collection of XML elements is subsequently mapped to Java objects and then to SQL tables and relationships. As

(2) we articulate templates, we also specify rules for how they can be composed into database designs for field databases. We seek more efficient representation and implementation of these rules and heuristics for suggesting additional data collection to expand the applicability of a database design.

(2) In our first prototype, we found HTML/ASP/SQL Server technology too brittle for flexible user interface. We then chose HTML/Java/Enhydra/SQL Server for the second version, but the web-accessibility requirement still limits user interface functionality.

(3) The appropriate level of abstraction for presenting templates and subsequent database design to the end user is still an open question. Database entities are more abstract, with greater normalization, than most researchers prefer. We provide high-level views for the SQLServer database, but the extent to which these will be implemented as non-normalized Access tables (and Excel spreadsheets) will be a matter of experimentation with users.

System design documents, such as conceptual data model of the metadata database, a conceptual view of how measurement data is organized, template and database design representations, and a current development schedule are available at http://scidb.evergreen.edu/cdbdocs-public/index.html.

The current system implementation can be seen at http://canopy.evergreen.edu/databank .

5. FUTURE WORK AND CONCLUSIONS

This section describes future work and outlines preliminary conclusions. Additional development includes refining the current implementation, adding productivity features and implementing the warehouse. More specifically, these involve: refine and field test data templates and determine how to manage template change; field test and improve the user interface; tools for data validation in the field; include additional field studies in the repository and increased scope of metadata source tables; provide metadata maintenance in the field. The data warehouse and cross study query capability are perhaps the most technically interesting of these.

One strategy for building *DataBank*'s warehouse is to upload the parts of a field database that match template structures into an integrated database, with a separate repository of individual databases each in its entirety. Rather than duplicate field databases, we are collaborating with Eric Simon's Caravel research team to investigate use of a virtual warehouse using *Le Select* [http://www-caravel.inria.fr]. To do this, individual database wrappers would articulate where schema match templates, and specialized client software would run queries using *Le Select*. We implemented a primitive *Le Select* client that allows a user to pose queries across a templates and executes those queries across several field databases.

Long term financial support and community contribution is more daunting than the technical challenges. To establish the canopy database in the real world, with real users, longer term issues remain, many of which are sociological. We identify three: 1) establishing a critical mass of users and data (including templates), 2) recruiting and training volunteer curators, and 3) finding long term funding for the operational system. We will work with the International Canopy Network (ICAN) and the Global Canopy Programme (GCP) to raise funds and consider chargeback to users, or provision of value-added services to the GCP foundation such as tracking research proposals and projects.

Though much remains to be done, we believe our preliminary work has shown the technical promise of using a relatively small number of domain-specific data structures (templates) to construct individualized field databases. We have shown that several such databases would be considerably more comparable than databases idiosyncratically designed, and that such a system is likely more practical than a global schema for ecology. We believe that end users could design effective databases using templates, and that productivity gains in research process, as well as easier data archiving and data mining, would ensue. We also now believe that templates would facilitate tools for visualization and analysis, and would significantly contribute to increased research productivity.

Although increasing researcher productivity is likely a necessary condition for ecologists to use database tools, it may not be sufficient. Integrating systems such as those we propose into the ecological research cycle will involve changes in the way ecology is practiced, and, as mentioned tangentially in this paper, rewards for archiving data sets are not yet generally perceived. Although such sociological changes are beyond the scope of this project, our work has suggested that both ecologists and computer scientists will be change agents as rewards for data archiving and integrative ecology are introduced into the scientific arena.

6. ACKNOWLEDGEMENTS.

We acknowledge many contributors, including student programmers James Tucker, Brook Hatch, Neil Honomichl, Peter Boonekamp and Mike Ficker; LTER information experts Don Henshaw, Gody Spycher and Susan Stafford; consultants Bonnie Moonchild and Jay Turner, and computer scientists Eric Simon and Dennis Shasha, and Phil Bernstein. Research technicians Steve Rentmeester and Bram Svoboda redesigned researcher spreadsheets into databases, and field researchers contributed data and advice, including Bob Van Pelt, David Shaw, Barbara Bond, Mark Harmon, Betsy Lyons, Hiroaki Ishii, Robert Mutzelfeldt, Roman Dial, Steve Sillett, Akihiro Sumida.

This work has been supported by the National Science Foundation grants and Research Experience for Undergraduate Supplements: BIR 9975510; BIR 9630316; BIR 9300771, INT 9981531.

7. REFRERENCES

[1] Bernstein, P.A. and E. Rahm, Data Warehouse Scenarios for Model Management, ER2000 Conference Proceedings. Springer-Verlag, 2000, pp. 1-15.

[2] Cushing, J. Nadkarni, N., Delcambre, L, Healy, K., Maier, D. and Ordway, E., The Development of Databases and Database Tools for Forest Canopy Researchers: A Model for Database Enhancement in the Ecological Sciences, in SSGRR2002W, L'Aquila, Italy, 2002

[3] Lowman, M. & N. Nadkarni, *Forest Canopies*. Academic Press, San Diego, 1995.

[4] Maier, D., J. B. Cushing, T. Keller and T. Marr, *Proxies in Practice: Object Architectures for Distributed Computational Workbenches,* Journal of the Brazilian Computer Society, 1996, pp. 3-1.

[5] D. Maier, J. B. Cushing, et al., *Object Data Models for Shared Molecular Structures,* in *Computerized Chemical Data Standards: Databases, Data Interchange, and Information Systems*, R. Lysakowski (ed), STP 1214, ASTM, 1993.

[6] Michener, W., J. Brunt, J. Helly, T. Kirchner and S. Stafford, Non-spatial metadata for the ecological sciences. *Ecological Applications* 7:330-342, 1997.

[7] Michener, W. , J. H. Porter, and S. Stafford (eds), Data and Information Management in the Ecological Sciences: A Resource Guide, 1998.

[8] Michener, W., J. Brunt (eds), *Ecological Data – Design, Management and Processing*, Blackwell Science Methods in Ecology Series, 2001.

[9] Moffett, M., *The High Frontier: Exploring the Tropical Rain Forest Canopy*, Harvard Univ. Press, Cambridge, Massachusetts, 1993.

[10] Nadkarni, N. & G. Parker, A profile of forest canopy science and scientists - who we are, what we want to know, and obstacles we face: results of an international survey. *Selbyana* 15:38-50, 1994.

[11] Nadkarni, N. & J. Cushing, Final report: designing the forest canopy researcher's workbench: computer tools for the 21st century. Intl. Canopy Network, Olympia, WA, 1995.

[12] National Research Council, *Finding the Forest for the Trees: The Challenge of Combining Diverse Environmental Data – Selected Case Studies,* Academy Press, Washington, D.C., 1995.

[13] National Research Council, *Bits of Power: Issues in Global Access to Scientific Data*. National Academy Press, Washington, D.C., 1997.

[14] National Science Foundation, Maier, D., E. Landis, J. Cushing, A. Frondorf, A. Silberschatz, M. Frame, J. Schnase (eds). *Report on a NSF, USGS, NASA June 2000 Workshop on Biodiversity and Ecosystem Informatics*,http://bio.gsfc.nasa.gov, 2001.

[20] Nottrott, R., M. B. Jones, M. Schildhauer, Using XML-Structured Metadata to Automate Quality Assurance Processing for Ecological Data, *Proceedings of the Third IEEE Computer Society Metadata Conference*. IEEE. Bethesda, MD., 1999.

[21] Porter, J. H., D. L. Henshaw, and S. Stafford, Research Metadata in Long-Term Ecological Research (LTER). *IEEE Metadata Conference*, 1997.

[22] G. Spycher, J. B. Cushing, D. L. Henshaw, S. G. Stafford, N. Nadkarni, Solving Problems for Validation, Federation, and Migration of Ecological Databases. *EcoInforma*, 1996.

The Future Of Ecoinformatics In Long Term Ecological Research

James W. Brunt

Department of Biology, University of New Mexico

Albuquerque, NM 87131-1091

and

Peter McCartney

Center for Environmental Studies, Arizona State University

Tempe, AZ 85287-3111

and

Karen Baker

Scripps Institution of Oceanography, University of California San Diego,

La Jolla, CA 92093-0218

and

Susan G. Stafford

Department of Forest Sciences, Colorado State University

Ft. Collins, CO 80525

ABSTRACT

Emerging information technologies allow new exploration into tools for the management and use of information that solve problems for ecologists and create new and innovative lines of scientific inquiry. Collaborative, multi-disciplinary research programs to facilitate these new lines of inquiry have produced a need for scientific information systems that communicate data, information, and knowledge across spatial, disciplinary, and cultural boundaries.

INTRODUCTION

Increased need for ecologists to examine global change, bio-complexity, and sustainability is resulting in research and synthesis at larger spatial and temporal scales than traditionally addressed in ecological studies. The development of collaborative, multi-disciplinary research programs has produced a concomitant need for scientific information systems that communicate data, information, and knowledge across spatial, disciplinary, and cultural boundaries. The primary motivation for developing scientific information systems must be the new types of scientific inquiry that they make possible. Information systems science and related information infrastructure are leading to a paradigm shift in biology [1][2]. Thus far this has been most evident in the genomic community [2], where the creation of databases and associated tools have facilitated a tremendous increase in the understanding of the relationship between the genetic sequences and the actions of specific genes. Ecology is perched on the brink of a similar expansion, brought on through improvements in software tools and data communication. In long-term studies, retention and documentation of the data are the foundation upon which the success of the overall project succeeds or fails. Long-term studies also depend on information systems to facilitate sharing of data and to combine data for the purpose of integrated multidisciplinary projects. In addition, public decisions involving environmental policy and management frequently require data that are regional or national, but most ecological data is collected at smaller scales. Information systems make it possible to integrate diverse data resources in ways that support decision-making processes.

We recognize that knowledge (in the broad sense) is generated through an iterative process of acquiring data, transforming it into useful information, and drawing inferences that enable us to achieve understanding and informed make decisions. Information management is slowly undergoing an evolution through these three domains.

The Long Term Ecological Research (LTER) Network is a collaborative effort involving more than 1100 scientists and students investigating ecological processes over long temporal and broad spatial scales.

- The Network promotes synthesis and comparative research across sites and ecosystems and among other related national and international research programs.
- The National Science Foundation established the LTER program in 1980 to support research on long-term ecological phenomena in the United States.
- The 24 LTER Sites represent diverse ecosystems and research emphases.
- The Network Office coordinates communication, network publications, and research-planning activities.

As LTER moves into its third decade, ecoinformatics will continue to play a critical role in defining and facilitating this

expanding, new ecology. The third decade of the LTER has been designated as a *Decade of Synthesis* for which the scale and complexity of the information management tasks presents a number of challenges for organizing and coordinating the diverse skills and resources within the LTER Network of research sites.

The LTER program was designed from its inception to incorporate data management as an integral component of the research. Within the overall goal of facilitating research, data managers at the site-level spent significant portions of their effort in developing methods to handle documentation and the custodial aspects of codes, data formats, and consistency during the 80's. In the 90's we built upon these developments, utilizing the rapidly expanding technology of the Internet to address the design of information systems. As we move to the next millennium, our goals are now expanding to building an infrastructure for the next level, a parlaying of information management into knowledge management.

In addition to supporting LTER site and intersite research, the data management program within LTER provides information to a diverse set of end users who expect to have access to our publicly funded research data. As scientific inquiry becomes more multi-disciplinary, we are challenged to find solutions for making primary ecological information more directly useable by (and valuable to) non-specialists. We embrace the research goals as a guiding force behind our work, but we also recognize that general scientists, public policy makers, businesses, the legal profession, K-12 educators, and even the entertainment industry, all make use of information about the natural world. There can be serious implications if those users make uninformed decisions based on faulty, outdated, or incomplete information.

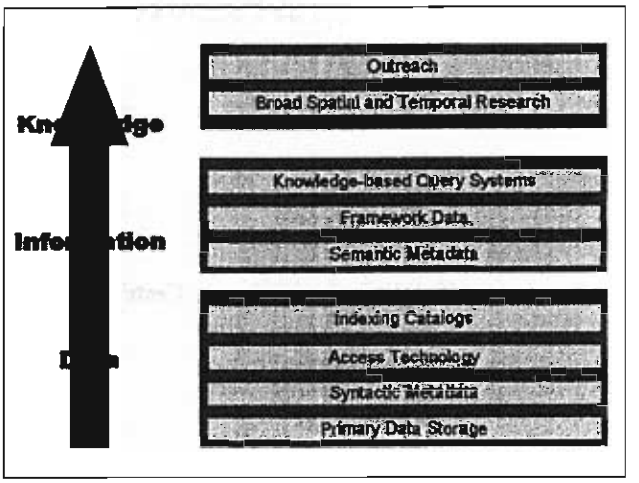
The purpose of this paper is to summarize the directions that information management, both within and outside the LTER program, need to look to respond to the changing needs of ecological science as we enter the next millennium.

METHODS FOR ADDRESSING ECOINFORMATICS CHALLENGES IN THE ECOLOGICAL SCIENCES

The challenges and expectations identified above call for the application of information technology beyond simple data storage and electronic publication to development of an active, globally integrated information network with the capacity to discover, access, interpret and process data facilely across the comparability and scaling barriers. Creation of this infrastructure requires investment of effort and resources into three broad areas that span the transformation of observations from data to information and to knowledge (Figure 1):

- Design a system of networked data storage to provide long-term management and accessibility of ecological data.
- Develop tools and procedures for facilitating the integration and synthesis of primary data.
- Promote and support research activities focusing on applications of archived data sources in broad, synthetic research.

Figure 1. Component Model of Ecoinformatics Infrastructure



DATA: Establishing Data Storage Network

We foresee a series of activities contributing to development of a global infrastructure for ecological data management and access. We envision this infrastructure developing through (1) the creation of data repositories that actively accumulate valuable data sets and ensure their long-term viability and accessibility, (2) development and adoption of standards for documenting databases and the research that produced them to enhance the usability of these data, and (3) development and adoption of procedures that will reverse the traditional attrition trends for data sets.

Long term management:

Data management plans seek to ensure data quality and availability. Quality control protocols operate from field sampling, lab analysis to data management. In the data flow, data quality is maintained through application of a variety of quality assurance/control procedures such as foreign key enforcement, validation triggers, and exploratory data analysis (EDA) screening.

Effectiveness of these methods is determined by the extent to which data management plans can be included in the actual research design for data collection. Availability of data can be threatened by both short-term factors such as power or device failure or long-term factors such as technology drift, media decay or format obsolescence. Solutions for these difficulties involve regular backups to removable media, and/or use of redundant subsystems. Long-term availability is dependent on maintenance of data in an online system with a plan for timely migration to new hardware and software. To successfully maintain data, it is necessary that an institution have the resources for and the commitment to managing and upgrading the necessary equipment and software, and to maintain connectivity. While it is relatively easy and inexpensive to install a data server and a web connection, it is quite another matter to keep that connection up over several years, let alone in perpetuity. Both site resources and commitment are required to form such a network of data repositories.

The most familiar type of primary data in ecological sciences is tabular data. Storage formats such as SQL and object-oriented

databases continue to evolve following open standards set by international consortia. New technologies are enabling far more sophisticated solutions for text data such as "grey" literature reports or records. With Extensible Markup Language (XML), it is possible to go beyond simple indexing of electronic documents to actually mapping their structure, enabling more rapid and accurate location of relevant information. For example, instead of searching the index for every occurrence of the words *bird* and *populations* in the hope of finding some survey data, a search engine could locate the section of a document tagged as <census results> to access the information requested. With XML rapidly being adopted as the language of the internet during this remarkable time of transition, it holds the potential to fill a variety roles in both the management and the exchange of information in the future.

Access technology

Most LTER sites now possess technology to provide access to data through a variety of methods. These range from simple downloads of static files to interactive query applications that support more sophisticated search and selection of information. Selected LTER sites have been active in exploring new approaches to data management. The Z39.50 was adopted as a client server search technology and is in common use among libraries and museums, as a platform-independent capability for searching multiple data sources on diverse hardware and software implementations. Further, the LTER Network Office, in partnership with the San Diego Supercomputer Center are considering the Storage Request Broker (SRB), software providing a single interface to hide differences between data storage systems. Many sites are active with geographic information system (GIS) so are poised for activity on map-based query interfaces which can provide a spatial framework for referencing data queries.

While these approaches are subject to continued innovation, a recognized limitation of some of these solutions is that they are inherently proprietary to the specific data content, storage and delivery system and thus are time-consuming to develop. A layer of open access technology needs to be draped over this network of data repositories that could facilitate the ability to conduct the most fundamental search and query operations from a single agent using a single protocol. Several technologies currently in development and limited implementation can provide multi-tier solutions:

- XML (eXtensible Markup Language) - a universal language for data exchange that is self-documenting,
- UDDI – Universal Data Discovery and Integration
- WSDL Web Services Description Language

Syntactic (Data-bound) Metadata

The term metadata refers to data that describe data. Metadata represent the key elements to transforming archived data sets into useable research resources. In this level of the model (Figure 1), we are concerned with information about the syntax of the data – information that describes each specific data set. This information is inextricably bound to the data set and is thus expected to be stored and managed in close conjunction with the data.

In a seminal paper about the survival of ecological data, Michener et al. [3] identify five levels of metadata description required to fully document an ecological dataset. These range from information about the research project that produced the data – names of investigators, sampling strategies, collection methods, etc. – to detailed attributes of the columns, data types and file formats of the data tables that were archived. The Federal Geographic Data Committee has published a standard for the content of spatial metadata that exhibits a similar hierarchical, albeit more exacting content structure. A Kansas University effort has developed a metadata standard for indexing and describing museum collections data; similar efforts to index electronic resources on the World Wide Web have been made by the Dublin Core (http://purl.org/dc).

There is a need for the development of widely accepted standards for ecological metadata that go beyond this simple beginning. For this to become a manageable task, these metadata need to be developed following a modular approach similar to other well know standards efforts such as the W^3 Consortium, FGDC and Dublin Core. Discrete working groups focusing on specific content domains within ecological research would produce modules that would be required only for datasets that contain certain kinds of information. Each product of these workshops would contribute toward a comprehensive standard that serves not to dictate research methods, but rather how to effectively document the structure and design behind one's methods and observations.

Indexing Catalogs

As the corpus of online data resources grows, the need for efficient indexing and searching far outstrips the capacity of static, unsophisticated aids such as html link pages and webcrawler-based search engines. Current efforts to develop online metadata catalogs such as the LTER Data Table of Contents database (URL: http://www.lternet.edu/DTOC) are building a valuable infrastructure for navigating the growing network of digital data. The design of these catalogs should be sufficiently open to support searches by search applications commonly in use across the internet such as Z39.50 and follow a development model based on other indexing efforts such as the National Spatial Data Infrastructures network of clearinghouses for geospatial data sets, the Council for the Preservation of Anthropological Records (http://archaeology.asu.edu/copar), and the Dublin Core.

Sustainability

For a system to be sustainable, a strategy is necessary for handling the incorporation of new data into the knowledge network. Despite its vital importance, few active research projects have the time or take the effort to produce metadata for their research data is often prohibitive; another barrier is that until recently no adequate guidelines have been developed[4][5]. Practices observed in other disciplines suggest several options. One is to encourage funding and permitting agencies to endorse the submission of research data into knowledge repositories and to adopt a set of standards for this process. Another is to work with professional societies to develop programs that create recognition for data archiving and documentation. One such partnership, created through cooperation between the Ecological Society of America and the San Diego Super Computer Center, developed a peer-review

The term metadata refers to data that describe data. Metadata

several options. One is to encourage funding and permitting

process for datasets and associated metadata, with successful submissions being published in the ESA journals. Finally, the cost and difficulty of creating metadata might be mitigated by development of freely distributable tools that automate the documentation process through reverse-engineering of data files and use of "wizard" forms that query the investigator for information similar to the way tax preparation software gathers financial backgrounds.

INFORMATION: Integration and synthesis of data

Information is modeled here as a bridge between primary data and knowledge. We recognize a series of technologies and activities that create the interface between data storage systems described above and the kinds of synthetic research questions we wish to accommodate within our broad infrastructure. One component is a set of standards for decomposing these questions into smaller elements that can be documented in standardized, machine-parsable form. Another is a very experimental approach involving the development of *expert systems* - sophisticated software tools that are capable of performing intelligent searches and processing of diverse datasets. The third, enabled in part by the second, is to identify a basic set of parameters relevant to the most broad inquires in biodiversity and produce synthetic framework data in the form of GIS covers and/or summary tables.

Semantic (Query-bound) Metadata

Comparable metadata necessary to perform translation and processing for scale matching needed to forge compatibility between two datasets based on their metadata descriptions. This latter body of metadata we refer to as semantic metadata because it concerns itself not with the organization of information but with its meaning. It is query-bound in that it documents our diverse units of inquiry, not just the more familiar syntactic metadata that documenting diverse units of observation.

This class of metadata might consist of calibration curves, thesauri that equate nominal categories, or machine-readable codes defining a particular set of processing steps required for certain types of data. This information might be included with the metadata of the particular datasets or be stored in a separate, central knowledge base to which regular additions are made. As a simplistic example, metadata associated with two different datasets might indicate that radiocarbon dates from one palynology core giving spore/pollen counts were reported in calendar years calibrated to a particular curve while those from another were reported as raw dates. Reference to a knowledge base containing the necessary calibration data and the appropriate processing steps would enable an expert system to retrieve the data and cast them in a compatible form by cross-calibration. As is the case for this example, the knowledge and software tools for performing many of these processing steps already exist. Development of an expert system in such cases will involve gathering and electronically encoding the various calibration curves, thesauri, classification rules, etc. used in existing databases and either developing or incorporating existing processing code that can act upon this information.

Develop tools and procedures for automated integration of data

Much of the existing technology for query and retrieval of data requires some degree of familiarity with the data structure and its meaning. However, few tools exist to facilitate the task of synthesizing data from diverse primary sources. We may expect that the not-too-distant future will bring sophisticated search engines and software tools for automating many data synthesis steps presently done by hand. These tools would be based on technologies such as expert systems and would be able to (1) respond to some relatively easy-to-use query language, (2) access both semantic metadata about the categories of information requested and syntactic metadata about the databases to be searched, and (3) perform a certain amount of query, evaluation, and processing of primary data prior to returning a result.

Within an overall strategic plan, we expect the initial products of this effort would be a set of loosely integrated software components that will accumulate and evolve as new information is brought into the system. It is untenable at this point to envision a single massive system with a single interface to respond to all queries against all known data sets. However, the tools exist today to develop some discrete components that can automate many of the routine and time consuming tasks associated with synthesizing diverse data sources.

Expert systems exist and are in common use in other fields. The key to developing these tools lies in two areas. The first involves extending our partnerships with expertise in sophisticated computer technology such as artificial intelligence, expert systems and neural networks. The expertise is not traditionally found within ecological sciences nor are the funding sources for these disciplines familiar ground for ecologists. The other area will involve more in-house effort. It will be necessary to adopt a language for encoding both syntactic and semantic metadata in machine-readable form. This form will likely be based on a language such as Resource Description Format (RDF), an XML based standard for encoding machine-readable metadata under development by the World Wide Web consortium. XML, like other SGML derived languages provides a rich syntax for expressing complex, structured information and has extensive support in the commercial industry.

Pilot projects such as the LTER Network Information System (NIS) [6][1] provide a means of mobilizing collaborations for combining research questions with technology and algorithms. These projects are typically small enough in scope not to require extraordinary resources for implementation, and just complex enough to demonstrate the scientific principles and the technological methods used. While simplified in detail, they approximate what a full-scale, full-complexity implementation would be like, even if the prototype turns out to be neither transportable nor scalable.

The demands this sort of approach places on the development of metadata are significant and justify continued investments in developing metadata standards and in streamlining the process by which metadata are generated. To be effective, existing efforts will need to be augmented with more rigorous procedures for encoding semantic metadata such as classification systems, measurement parameters, analytic procedures, etc. Current metadata implementations still rely

heavily on open text representations of information and thus will require further revamping to meet the rigors that this sort of advanced processing tools require.

KNOWLEDGE: Promote and support research collaborations integrating information at broad spatial or temporal scales

Funding strategies for information technology during the last two decades have focused on developing infrastructure with the notion that "if you build it, they will come". The new crop of initiatives shows much more concern with ensuring that our data products are of significant value to current and future research – that is, new proposals are expected to provide *application*, not just availability, of data. Like most current research efforts, construction of this broad information infrastructure must have a strong question-driven component to enable a process of feedback between end-users and data sources. We need to challenge the ecology community to provide recognition to archive and share their data so that we can create both incentives and guidelines for developing the infrastructure components outlined above.

To bring about these changes requires adopting new perspectives on the relationship between data management and research. The traditional approach is for ecological scientists to make use of selected technology (and computer science methods). Recently, however, we see evidence from the development of database technology, web technology, modeling and simulation techniques, that advances in computer science and technology can provide new methods that can influence how ecological science is conducted. These are new opportunities for scientific explorations that were not conceived by ecologists. Also these advances are not directly driven by the current practice of ecologists. That is, there are opportunities for computer science to suggest changes (advances) to the methods of ecologists. LTER partners recognize that challenging problems in ecology provide focus and impetus to developments in computer science. Through partnership a synergy is created from the strengths and contributions of everyone. When data management strategies are explicitly aligned with the long term goals of synthetic and broad regional research, feedback is ensured to better guide multi-scale database and framework research strategies.

ORGANIZATIONAL STRUCTURE

The development of a hierarchical, networked clearly lies beyond the scope of a single project or institution. The scale and complexity of the task presents a number of challenges for organizing and coordinating the diverse skills and resources within the community of scientists that will be addressing these goals.

We need to envision a vehicle for continued ecoinformatics activities. We see this taking place at several levels –

- Creating a center to coordinate activities, i.e support a consortium
- Participating as distributed laboratories of ecoinformatics
- Developing mechanisms for cross-fertilization of ideas and technology
- Meeting staffing needs by developing mechanisms for professional development

ECOINFORMATICS CONSORTIUM (EIC)

To capitalize on the strengths of individual partnering organizations, we are developing an Eco-Informatics Consortium (EIC) that acts as the steering committee for the larger community. The consortium serves to formalize the ad-hoc partnerships between groups of scientists currently working to address similar questions and provides a mechanism to capitalize on synergism, increase communication and coordination, and accomplish "collective" goals, while at the same time maintaining individual autonomy.

Administrative functions of the EIC would have some sort of physical location, possibly rotating between institutions. However, research activities will predominately be carried out within a virtual environment, supplemented by workshops, conferences or other gatherings. We envision the structure of this entity to be somewhat along the lines of the Dublin Core or the W3 Consortium. The mission of the EIC is to develop the vision for the larger body and mobilize different pieces of the consortium as needed. The EIC structure enables each member of the EIC to concentrate on areas where they have expertise, whether it be information management, computer science, or social issues, yet have these efforts connected to a larger community.

In terms of the larger body, we envision the EIC to be the hub, around which the consortium members are arrayed like spokes on a wheel. Each of these spokes would then act as a local hub for organizations at the subsidiary levels. For example the LTER network office would act as the hub and the LTER sites would be the spokes of this subsidiary wheel. These sites in turn would be hubs for the local community of organizations surrounding their home institution or their domain science.

Depending upon the circumstances, different members of the consortium could be linked to take advantage of specific opportunities as they arise. This will range from a small group of organizations at one of the auxiliary levels mobilizing to develop a software tool to the entire EIC mobilizing to compete for large federal IT grant opportunities. We can think of the functioning of the EIC spokes/hubs as different sets of lights that blink on and off to take advantage of opportunities at a range of levels. This flexibility and balance between small and nimble and large and powerful will be crucial to the success of the EIC.

Information flow through this model of the EIC is critical to the success of its efforts and we envision a bi-directional flow of information, such that solutions can percolate up through the subsidiary wheels to the EIC hub and then be redistributed to the other components of the EIC. This bi-directional flow will allow each member of the consortium to take advantage of all the data, information, and knowledge of the entire EIC.

Distributed Laboratories for Ecoinformatics

While the EIC is expected to exist primarily as an organizational vehicle, we anticipate that it will function to catalyze the creation of many shared infrastructure resources. The most obvious resource will be the network of data sets that can be shared among partners, enabling access to a wider range

of data while at the same time allowing institutions to concentrate their management efforts on those data sets for which they have direct responsibility. Recent work at San Diego Supercomputer Center demonstrates the potential for meta systems - virtual machines that can compile resources from a heterogeneous computing environment, accessing data, memory, processor cycles, and resources from distributed physical computers across a network; network environments for application sharing; and distributed super-computing. Another project of the EIC will be to create collaborative environments that facilitate communication through network technologies such as video-conferencing, white boarding, and application sharing. Finally, we envision the creation of training and educational programs in which the skills and knowledge are disseminated.

Mechanisms for cross-fertilization

The EIC provides a method for organizing and implementing a variety of mechanisms proven to be of value in communicating ideas, technology and innovation. One such mechanism is the organization of topic-focused workshops. NCEAS provides a successful model for how such workshops can be managed. Another is the creation of cross-institutional positions such as the LTER postdoctoral position at San Diego Super Computing Center. Short-term visiting researcher opportunities for personnel involved in information management would also serve to cross-fertilize ideas and solutions. Exchange programs could be developed between sites, between sites and the Network Office, or even between LTER and other partners within the EIC. The tradition of external invitees to the LTER data management meeting has benefited both the LTER program and its guests.

Infrastructure Changes within LTER data management

Expansion of the scale of information management from the local to the network level will carry implications for the organization of IM personnel. First, it is important that information management be well represented in management and executive decision-making. This may require IM staffing changes such as redefinition of existing responsibilities, training for and creation of higher-level positions for IM, or assignment of existing PI-level personnel to IM leadership roles. Second, opportunities need to be created to give information management personnel time via sabbaticals or similar mechanisms for infrastructure-related projects that may involve collaboration with EIC partners, extended travel, etc. Third, there will be a need for increased funding and opportunities for training as information managers struggle to design systems, update technology for intersite data exchange solutions while maintaining existing services and receiving annual dataset additions. Finally, it is likely that information management staff will need to be expanded to accommodate the broader range of activities, particularly in the areas of information and knowledge management, while at the same time maintaining current levels of support in the area of site data collection and management.

CONCLUSIONS

The LTER is a network with a community of scientists well focused on long-term ecological science and with a community of information managers attending to data management. While the complexity of data handling requirements and expectations has increased, strategies are being developed to create and to benefit from a synergy with technology and advances in computer science.

ACKNOWLEDGMENTS

The authors wish to express special thanks to the editors and reviewers for many helpful comments. This work was supported NSF Grants DEB-9980154, DEB-9634135, DEB-9714833,OPP-9632763, DBI-0111544, DBI-9983132, and EIA-0131958

REFERENCES

[1] Baker, KS, Benson B, Henshaw DL, Blodgett D, Porter J, Stafford, SG. 2000.Evolution of a Multi-Site Network Information System: the LTER Information Management Paradigm, BioScience 50(11): 963-978,2000.

[2] Robert J. Robbins 1996. BioInformatics: Essential Infrastructure for Global Biology. Journal of Computational Biology

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

[4] Jones, M.B., C.Berkley, J. Bojilova, and M. Schildhauer, 2001. Managing Scientific Metadata. IEEE Internet Computing Sep-Oct 2001

[5] Cook, R.B., R.J. Olson, P.Kanciruk, L.A.Hook, 2001. Best practices for preparing ecological data sets to share and archive. Bulletin of the Ecological Society of America 82(2)

[6] Brunt JW. 1999. The LTER network information system: a framework for ecological information management. Pages 435-440 in Aguirre-Bravo C, Franco CR eds North American Science Symposium: Toward a Unified Framework for Inventorying and Monitoring Forest Ecosystem Resources; 2-6 Nov 1998; Guadalajara,l Mexico. Fort Collins (CO): US Department of Agriculture, Forest Service, Rocky Mountain Research Station. Proceedings RMRS-P-12.

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.

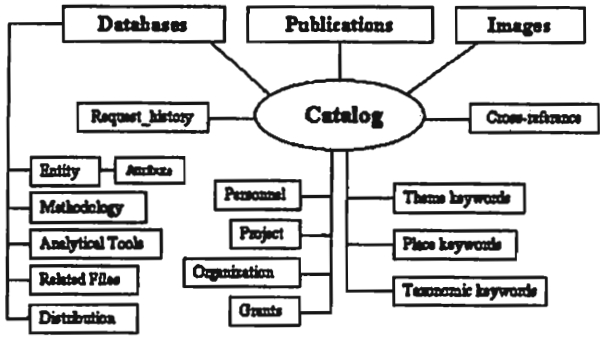


Figure 1. Simplified view of the normalized metadata structure.

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.

Using XML-encoded Metadata as a Basis for Advanced Information Systems for Ecological Research

Peter H. MCCARTNEY
Center for Environmental Studies
Arizona State University
Tempe, AZ 85282, USA

And

Matthew B. JONES
National Center for Ecological Analysis and Synthesis
University of California, Santa Barbara
Santa Barbara, CA, USA

1. ABSTRACT

Metadata provide information on the structure and meaning of data. It is one of the most basic components for building a scalable, networked infrastructure for accessing ecological data. Several partnering groups from ecology have collaborated to define a standardized format for metadata that is machine-parseable and extensible. This has enabled new projects focusing on the development of tools for managing metadata archives and for accessing and processing the datasets they describe. Ecological Metadata Language and its associated tools will have a significant impact on the integration and synthesis of ecological data at a global level.

2. THE ROLE OF METADATA IN ECOLOGICAL INFORMATICS

The goals of ecological informatics are to ensure the long-term availability of ecological data and to enhance the usability of those data in the pursuit of knowledge about our environment. The use of digital media to capture, store, and process increasingly larger volumes of data has contributed significantly to these goals, but this has in turn created new challenges for indexing, navigating and documenting this sudden wealth of information[1].

A critical tool for meeting this challenge is *metadata*. Metadata is the documentation that transforms data from a stream of numbers and characters into information. All of us who work with data have relied upon metadata such as column labels, data type declarations, etc., even if we didn't recognize those things by that name. Metadata provides information at many levels to support many phases of our interaction with ecological data. Information such as catalog identifiers, title, originator, etc. provides the base citation information for *identifying* a dataset. Search engines rely on keywords and coverage descriptors for spatial, temporal, or thematic domains to assist with the *discovery* of datasets. Information on the research context that produced the data assists in the *evaluation* of the dataset. Connection details, filenames, and access control information enable *acquisition* of a dataset. Finally, detailed descriptions of entities, attributes, and data quality enhance the *usability* of the dataset for analysis.

Members of the ecological research community have been compiling metadata as part of the data archive process for over a decade. Notable examples include the Long -Term Ecological Research network [2] and the Oak Ridge National Labs[3]. In 1997, following Michener's paper on ecologically relevant metadata [4], researchers at the National Center for Ecological Analysis and Synthesis (NCEAS) began implementing the first version of Ecological Metadata Language (EML), which was revised several times and culminated in EML version 1.4.1 [5]. As experience with this initial version of EML grew, it became apparent that it needed a revision to increase its usability and flexibility for the ecological community. The Knowledge Network for Biocomplexity (KNB) project thus began an effort to revise the EML specification to produce a second version that was

initial version of EML grew, it became apparent

even more broadly useful. Simultaneously, in summer 1999, the LTER information management committee evaluated the status of metadata within the LTER network in light of a series of long term goals for the future of informatics in ecology (see Brunt et al this volume). The committee found that (1) there was a need for standardization in both content and presentation format, and (2) that metadata needed to be presented in a machine-parseable form to support advanced development of automated data search and processing tools [6]. As a result, a metadata committee was formed to work with the two independently funded projects (Knowledge Network for Biocomplexity [7] and Arizona State University's Networking our Research Legacy project [8]) that had begun the process of revising Ecological Metadata Language (EML).

3. ECOLOGICAL METADATA LANGUAGE

Development of EML has followed several guiding principles. (1) It should be encoded in a machine-parseable format, with strong industry support and independence from particular platforms or software. (2) Extensive prior work in metadata standards both within and outside ecology should be used as a basis to enhance compatibility and reduce redundancy. (3) The standard should serve to integrate, rather than dictate, individual site solutions for creating, storing and managing metadata.

eXtensible Markup Language (XML) was selected for the encoding format XML is an SGML-based text syntax (UNICODE) for marking up data and documents. It bears similarities with HTML, but is designed for tagging the content of a document with a means for validating that content against a formal schema. Tools for parsing XML documents are available for all modern development languages and XML documents are easily transformed into other formats for display through the related eXtensible Stylesheet Language (XSL) specification. The XML Schema specification is itself an XML file and provides a powerful medium for designing and sharing content models through the use of commercial design tools or custom XSL style sheets.

A significant amount of prior research was reviewed in designing EML. Within the ecological community, the seminal paper by Michener et al. [4] had established guidelines for metadata content that was reflected in the text and HTML formats designed by various individual LTER sites. NCEAS encoded the content model developed by Michener et al. in XML in what was first released as the EML 1.0 specification. Outside ecology, extensive work on geospatial metadata standards by the Federal Geographic Data Commission (FGDC)[9] and the International Standards Organization (ISO)[10] resulted in comprehensive content models released as text and Universal Modeling language (UML) specifications respectively. The National Biological Information Infrastructure (NBII) extended the FGDC standard to accommodate biological datasets [11]. The resulting NBII standard adopted substantial portions of the original EML version 1.0 specification. Other standards such as the Dublin Core Element Set [12] for internet resources, the Global Change Master Directory DIF standard, and the Mercury metadata standard used by Oak Ridge National Labs [13] were also reviewed.

Considerable diversity existed across the 24 LTER network sites in terms of the content and format of metadata, and in the manner in which metadata catalogs were integrated into other aspects of site management. The goal in creating EML was to define a common standard and format that could be generated easily from existing metadata without burdening sites with significant alteration of their existing system.

EML 2.0 Design

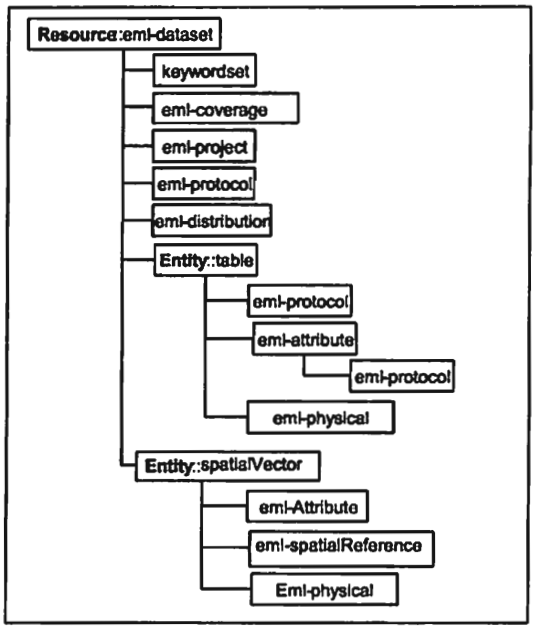
The resulting draft specification for EML 2.0 is a complete revision of the original EML 1.0. Detailed information on its development and downloads of draft specifications are available online[14]. EML 2.0 has several significant design features. (1) It is modular, with separate schemas defining sets of descriptors that relate to a specific category of information. (2) It uses XML Schema complex types to enable an object-oriented approach in which abstract classes are defined and then extended to create specific variants. Using this approach, EML defines several information resource types including "dataset" and "literature" (and potentially many more) that each inherit a common set of elements that correspond to the basic identification and discovery elements found in most metadata standards. (3) It is extensible by linking multiple modules within a package. Features derived from XML and form Resource Description Framework (RDF) allow subject-object relations to be defined between metadata document without modification of existing module schemas. (4) EML modules are organized to separate the description of the *logical* content of an information resource from that of its *physical*

instance. This feature automatically abstracts the details of physical formatting from users, allowing them to focus on the information itself and simplifies the maintenance of metadata, as disk formats or storage locations change through time.

Overview of content models for EML dataset
The super class for all EML documents is Resource. This set of elements defines those identifier and discovery elements that are common to any information resource and is based closely on the Dublin Core metadata standard. Resource is never directly instantiated – it is extended by several schemas including eml-dataset, eml-literature, and eml-software. Still others may be defined such as eml-model, or eml-collection.

Eml-dataset introduces several elements for describing a dataset and serves as the association point for a series of modules used in defining certain types of data or properties of data. Eml-project provides information on the research context that produced the data. A dataset is associated with one or more entities, each of which is described with a module that is extended from a basic Entity class. These may include tables, gislayers, images, grids, views, or stored procedures. Depending on the type of entity, other modules (such as attributes, constraints, spatial reference, spatial organization, data-quality) may be associated. The information provided in the entity and associated modules focuses on the logical information of the data. A related module, eml-physical, provides the descriptions of the actual digital instance of that entity (such as file information, connection information, column parsing instructions, etc). Changes in the format or location of a file can be made without altering the logical description represented in the entity section.

Other modules (such as protocol or responsible party) may be associated with several different modules whenever a particular class of information is appropriate. These modules define a consistent structure for specific kinds of information that could potentially apply in many different contexts within a metadata document. An EML metadata package would consist of one document such as eml-dataset or eml-literature, plus any other associated modules, and optionally the data objects as well. Resolving the linkages between modules specified by the triple statements would yield a nested tree of documents that can be easily traversed to locate any given element of information (Figure 1).



4. APPLICATIONS

The most significant feature of any metadata standard is the advances it enables for data discovery, access, and analysis. The projects responsible for developing EML have been working simultaneously on several software products that will facilitate access and use of environmental data.

Metadata creation
One of the most limiting obstacles to building networked data archives is getting past the learning curve and time burden of filling in metadata descriptions. EML is a fairly complex set of elements numbering in the 100's, many of which are not applicable for any given dataset. Two similar products are being developed to provide a simpler interface that would encourage scientists to prepare metadata without needing to either learn an entire management system or hire a data manager. *Morpho* is a Java-based metadata management tool developed by the KNB project. Building on an earlier XML editor developed by NCEAS[15], Morpho combines a user friendly forms environment for editing EML documents with a management client for submitting, maintaining and searching metadata packages on a networked storage system. Extensive configuration enables Morpho to accommodate changes or extensions to the EML schemas without requiring modification of the program code. Morpho includes a reverse engineering module for interactively walking users through the documentation of ASCII data files by parsing the file and allowing the user to view,

change, or provide more detail on the results. ASU is developing a related project called Xylographa which will be a web-based application consisting of three main components 1) a collection of reverse engineering modules running as either web services or Java applets (currently a relational database module is completed), 2) an import utility for parsing and importing other metadata formats into EML via XSLT style sheets or Java servlets, and 3) an interview wizard that walks a user through the metadata entry process in a step by step manner that provides navigation guides and access to contextual help. A metaphor for the design of Xylographa is modern tax software, such as *Taxcut* and *Turbotax* that use an interview mode and automated retrieval of information from related documents.

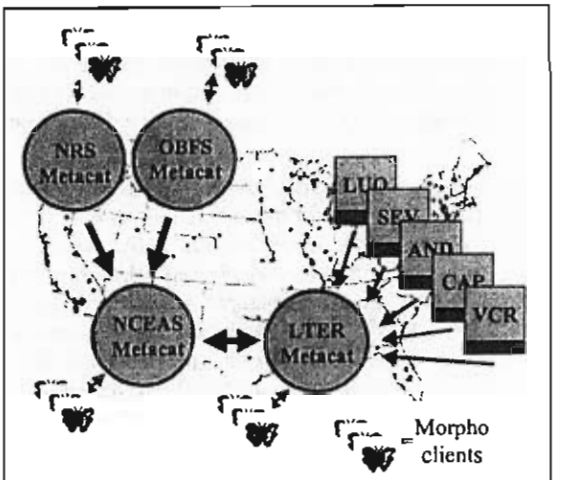


Figure 2. National network of Metacat servers.

Metadata management

XML's flexibility and extensibility pose new challenges for metadata storage and query. Standardization of format allows search expressions to be constructed using a universal syntax, sparing users to log on to specific archive web pages or learn a specific catalog's query syntax.

The KNB project has developed a native XML database storage system, *Metacat*, that decomposes the XML document object model into a linked list of node descriptions that are stored as individual records in a relational database [16, 17]. The *Metacat* servlet receives queries expressed in Xpath syntax and transforms them into SQL statements to locate matching nodes. The relational pointers are then resolved back up to the document's root so that Metacat can return the entire document. Metacat forms the basis for what KNB hopes will be a national network of replicated metadata servers that can be queried by Morpho or other client applications using an XML messaging format to encode Xpath search expressions and subsequent XML responses. For sites or individual researchers without an existing metadata management system, the Metacat/Morpho combination is an ideal solution as it is a complete system that is already configured to participate in a national network (Figure 2).

Xanthoria is a project developed by ASU to provide a threaded client-server search system for querying multiple, heterogeneous XML data catalogs. The goal in developing Xanthoria was to build an easily configured solution to provide an EML interface to existing SQL-based metadata catalogs. Xanthoria works very similar to Z39.50, a text search system used in many library networks, but is based on XML and Xpath. Xanthoria services connect to several types of storage systems including SQL databases, XML file folders, Xindice XML database, and Metacat. In the case of the SQL connector, a user-configurable Java bean performs the SQL to XML translation. In all connectors, differences in content schema are handled by user-supplied XSLT stylsheets that translate the native schema to EML. Each connector runs as a web service, listening for Simple Object Access Protocol (SOAP) requests from a client application (Figure 3). The structure of these requests is an extension of the XML messaging format used by Morpho to communicate with Metacat. The query application communicates with the targets, and collates and paginates the results for the user. The query form for generating the search expression is generated from the schema itself and can thus accommodate search on any XML target for which an XML Schema file has been provided. It uses an external configuration file identifying available targets and the schemas

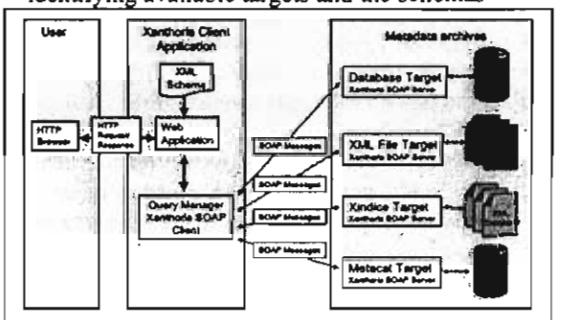


Figure 3. Xanthoria Web Service for XML Metadata.

supported. This configuration file also provides a map to the schema hierarchy so that the client is aware that a given target can support queries based on its own schema as well as on a more generic schema from which it was extended. For example, the client will send queries based on EML-Resource to both dataset and literature targets, but queries based on EML-Dataset will be sent only to targets for that schema.

Processing and Analysis

Impacts that more directly affect researchers are to be made by other current projects that seek to make use of machine-parsable metadata to enable more automated processing and analysis of data. At ASU, a collection of web-based data access and processing tools are being created to provide users with a richer array of exploratory and download tools than is currently available in most data clearinghouse systems. Basic GIS operations such as resample, reproject, or clip can be requested before data are downloaded. Tabular data will be queryable via a JDBC connection using a wizard that helps users construct SQL queries. Exploratory Data Analysis (EDA) functions (such as charts, plots, and cross-tabluations) will be available as an enhanced data browsing package.

ASU is also working on applications that build upon this basic data access infrastructure to target specific user groups. One example is a biodiversity applicaton for the Ecology Explorers educational program at CAP LTER. In this application, a series of XML configuration files will be used to provide a map between the EML descriptions of several datasets and some fundamental parameters that K-12 users will explore through this guided application. Users will be able to analyze parameters such as species richness using a choice of input data such as birds, arthropods, vegetation without needing to understand how to extract that particular query from the different source databases.

The KNB project is working on a project called *Monarch* that provides an exploratory data analysis and modeling environment for data described by EML metadata. Monarch uses an XML configuration system to describe analyses and models that are implemented in commonly used analytic tools (such as SAS, Matlab, etc). Monarch uses the information from an EML dataset description to generate the appropriate command scripts for a particular analysis and then executes the analysis using a plug-in architecture for the target execution environment. As a result, any data that is accessible and described in an EML format can be automatically analyzed over the web using these powerful statistical packages, which dramatically speeds up the process of understanding and interpreting data in synthetic and collaborative analyses. Monarch is expected to be a very useful technology for network developers to provide distributed access to common data processing and analysis functions.

5. FUTURE DIRECTIONS

The applications described above illustrate some of the initial efforts to draw upon the power of standardized, machine-readable metadata. Within them, several common themes point to future directions for informatics development.

One apparent goal is the development of integrated, networked applications that provide users with access to the full range of analytic functions without the need to install or learn specialized statistical or GIS software. This not only benefits many researchers, it has a profound effect on our ability to make ecological data available to a broader community including educational users, policy makers, and the general public. Standardized metadata, combined with online network access to data, will enable many applications to be constructed for the same data sources, each targeting a specific kind of audience.

Another clear trend is the abstraction from the physical details of data organization, encouraging the user to express their analytic requests in a syntax that is much closer to the logical content of the data. File formats and storage solutions are constantly evolving. One of the functions of metadata should be to provide the linkage between physical storage and information in a manner that frees the user from tracking changes or details within the physical component. Future research aims at higher levels of abstraction still. While EML provides a consistent *syntax* for addressing datasets, it does little at the present to overcome the *semantic* differences between datasets. New goals for metadata research will turn to ontology-based solutions for linking the EML descriptions of data to inquiry-based concepts that come closer still to the parameters by which we define ecological knowledge.

6. CONCLUSIONS

Standardized metadata is a significant step forward in ecological informatics. It provides the means for cataloging the growing base of data archives and

for addressing these data through a common syntax. This in turn is leading to the development of much more versatile applications that enable users to contribute to, navigate, and make use of, networked archives of ecological data.

REFERENCES

[1] Nature. 1999. It's sink or swim as a tidal wave of data approaches. Nature 399:517-520.

[2] Long Term Ecological Research. http://www.lternet.edu.

[3] Olson, R., and R. A. McCord. 1998. Data archival. Pp. 53-58 in W. K. Michener, J. H. Porter, and S. G. Stafford, eds., Data and information management in the ecological sciences: A resource guide. Long-Term Ecological Research Network Office, Albuquerque, NM.

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

[5] Nottrott, R., M.B. Jones, and M. Schildhauer, 1999. Using XML-structured metadata to automate quality assurance processing for ecological data, Proceedings of the Third IEEE Computer Society Metadata Conference. Bethesda, MD, April 6-7, 1999.

[6] McCartney 2000. Report of the Long-Term Ecological Research Metadata Committee Meeting, February 2000, NET Office, Albuquerque. [http://caplter.asu.edu/data/metadata/report2k.doc]

[7] KNB, 1999. The Knowledge Network for Biocomplexity. [http://knb.ecoinformatics.org/]

[8] NRL, 2000. Networking our Research Legacy. http://ces.asu.edu/bdi

[9] FGDC 1998. Content Standard for Digital Geospatial Metadata. Federal Geographic Data Committee.

[10] International Standards Organization, 1999. CD 19115, Geographic information – Metadata. Norwegian Technology Standards Institution, Oslo, Norway.

[11] Frondorf, A., M.B. Jones, and S. Stitt, 1999. Linking the FGDC geospatial metadata content standard to the biological/ecological sciences, Proceedings of the Third IEEE Computer Society Metadata Conference. Bethesda, MD. April 6-7, 1999

[12] Dublin Core Metadata Initiative. [http://dublincore.org/].

[13] Olson, R., L. Voorhees, J. Field, and M. Gentry. 1996. Packaging and distributing ecological data from multisite studies. Pp. 93-102 in Proceedings of the Eco-Informa Workshop, Global Networks for Environmental Information, 4-7 November 1996, Lake Buena Vista, FL. Environmental Research Institute of Michigan, Ann Arbor.

[14] Ecoinformatics.org [http://www.ecoinformatics.org]

[15] Nottrott, R., M.B. Jones, and M. Schildhauer, 1999. Using XML-structured metadata to automate quality assurance processing for ecological data, Proceedings of the Third IEEE Computer Society Metadata Conference. Bethesda, MD, April 6-7, 1999.

[16] Berkley, C., M.B. Jones, J. Bojilova, and D. Higgins, 2001. Metacat: a Schema-Independent XML Database System, 13th International Conference on Scientific and Statistical Database Management, IEEE Computer Society.

[17] Jones, M.B., C. Berkley, J. Bojilova, and M. Schildhauer, 2001. Managing Scientific Metadata, IEEE Internet Computing 5(5): 59-68.

Common Information Management Framework: in Practice

Eda C. Meléndez-Colom
Luquillo Long-Term Ecological Research Program, Institute for Tropical Ecosystem Studies,
Natural Science Faculty, University of Puerto Rico
San Juan, PR 00936-3682

and

Karen S. Baker
Palmer Long-Term Ecological Research Program, Scripps Institution of Oceanography
University of California, San Diego
La Jolla, CA 92093-0218

ABSTRACT

A common goal of information management systems (IMS) is to share information among its users and originators. These systems are usually implemented by project managers and sponsors. The design and implementation of an IMS in a research organization, prompted by directives from sponsors and mediated by the vision of information managers, reflects the activities performed by a research team (science investigators, students, and field and laboratory technicians). The Long-Term Ecological Research program is presented as an example of a data sharing community whose distinctive goal, in addition to sharing, is to preserve information for future generations. A Common Information Management Framework (CIMF) is the product of the LTER community interactions and the information tools to achieve their goals. At the same time, it provides an integrated platform to promote, facilitate, and guide the members of an LTER community in the management of the information they generate.

Keywords: management framework, community, information management, system design, LTER information system, data sharing and preservation

1. INTRODUCTION

To share and preserve information are the principal goals of developing common management systems within a spectrum of communities (IRRI, 2001; Water and Sewer Board of Montgomery, Alabama, 2001). Three system components can be identified: drivers that motivate the conception (design, definition, and implementation), infrastructure components that define data management, and a community that defines and uses this framework.

The Long-Term Ecological Research (LTER) program provides an example of a research community sharing data. The LTER consists of twenty-four independent research sites, each studying a specific ecosystem, linked into a network of sites supported by a Network Office. Research activities begin with field data collection and include data management. Depending on the type of physical environment where the studies are performed, the LTER sites collect data that range from plankton, fungal, and vertically integrated acoustic biomass data to climate, water and soil chemistry, animal, vegetation and population, and disturbance data such as hurricanes, fires, and lightning strike data. The requirement that each site designate an information manager ensures initiation of site data management, critical for long-term research but often untended for short-term projects.

An LTER information management (IM) system must balance responsiveness to immediate local research needs with a plan for long-term data storage. Within a research community, the needs are to facilitate short-term and long-term science, to enable analysis and synthesis as well as to capture and archive information for future generations (Bowser, 1986; Michener and Brunt, 2000).

The role of an LTER information manager is to design and develop a digital framework that supports and reflects the community's shifting research activities. We extend the term *common management framework* originally used for describing the centralized set of data, computational tools, and schemas for organizing information at an LTER site [11] to the term *Common Information Management Framework* (CIMF) which includes the drivers and the community that interact in forming the framework. This paper presents the components of a CIMF identified while working with the team of information managers within the LTER community.

definition, and implementation), infrastructure components

interact in forming the framework. This paper presents the

2. ENVIRONMENT FOR A FRAMEWORK

System Drivers

The stimulus to share data (and ideas) in a broad context is, in many instances, conceived and implemented top-down by the higher administrative or intellectual hierarchy of an organization. In a local scenario, centralizing information in a Local Area Network (LAN) addresses the technical component of an organization's communication issues. When the need to become more efficient in the data management services provided to the community of users develops, a common framework provides a mechanism to scale from accessing and manipulating data to sharing information. In the case of the LTER, approaches to IM change as the research team's dedication to share data shifts from simply complying with sponsor directives to trying to integrate the different components of their research work. To act as a research team and to do synthesis, the scientific community members must be willing to share.

The LTER program sites adopt a holistic approach to local CIMF while addressing a broad scope of research activities [9] and sharing an articulated common IM vision [1]. The timeframe for data release depends on the complexity of the processes to gather, submit, and quality control the data. The LTER community responded to the sponsor's directive [15] to share data by preparing a data policy outlining data availability within two years of collection. In 1990 the LTER Information Managers Committee had reported several potential positive, negative and legal aspects of data sharing and recognized the value of making data accessible to others

Table 1. Data Access Policy

- *Identify existing community data set types according to their accesibility*
- *Agree on what data will be made available online*
- *Define data availability timeframe*
- *Define compliance guidelines for core and non-core data*
- *Provide justification for data sets not posted online*
- *Create written policy for secondary use of public data*
- *Document data use and data policy agreement*
- *Post a citation format, a disclaimer, and encouragement for ethical behavior to protect the data provider*

in order to advance science [14]. This document formed the initial guidelines for site information management policies. Table 1 summarizes some important aspects of data policy.

Infrastructure

Information management systems provide an infrastructure to serve the common interest of a research site's scientific community with tools to achieve synthesis and cross-site activities. When centralized, data generated in the field and laboratories along with its metadata are accessible to all participants (Strebel et al, 1994; Ingersoll et al, 1997; Brunt et al, 1998; Baker et al, 2000). Table 2 lists the infrastructure elements important in defining a complete CIMF.

Table 2. Infrastructure

- *A personnel directory to describe project participants*
- *A bibliography to list site related publications*
- *Metadata forms for data submissions*
- *A catalog of data sets consisting of data and metadata*
- *A description of projects for which datasets are generated*
- *In-house accessibility to data set owners while ensuring data integrity*
- *Data rights so owners have the option of exclusive access to their data*
- *Mechanisms giving file system access to local and off-site investigators*
- *Applications providing tools to extract, manipulate, and analyze data*
- *Back up system procedures for system security and disaster recovery*

Within this framework, the diversity of content and data formats must be captured by standardized data documentation (ISO, 2002; FGDC, 2002). The primary objective is that data be accompanied by metadata (data documentation) so that users are provided with the research context within which the data were created, the methodology used, and the specifications for each measurement [13]. Table 3 shows the relationship of the CIMF infrastructure elements, as defined and documented in the metadata, to the principal research components of the site. Each site adopts metadata standards to reflect its range of research activities. A set of minimum standards developed by the Network's Information Management Committee is available to all sites [16].

Community

The research community members can be divided into three main groups: a research team, an information management team, and an administration team. The administration team holds all the institutional administrative personnel who give support to the other two teams of the community. The research team (scientist, students, and field and lab technicians) has two distinct roles in a CIMF: the driving force for the ongoing revision of the infrastructure elements and the users whose needs are to be met. In an ideal situation the information manager works on infrastructure, computational, and dissemination tasks with the assistance of a staff of information specialists performing system and web design/ development/ maintenance, as well as data entry and programming. The information manager fills multiple roles within this scenario. Given the breadth of the responsibilities

Table 3. Infrastructure Organization

- *A cluster of two or more data sets could be defined by grouping related datasets within a broader research context*
- *When a cluster is defined, it can be described by answering the question of why the data sets are collected, why are they useful to the originator (the owner and creator of the data sets), and the scientific questions to which they are pertinent*
- *A title and the description of each cluster could be included in the metadata for each of the data sets involved. The description can point (link) to another metadata file containing the description*
- *The list of these clusters and the stand alone datasets provide a front end interface for the users visiting the data web page*

that must undertaken by the information management team, it is critical that the role of the information manager be well defined (Table 4).

Table 4. Role of Information Manager

- *Participate actively as member of local data manager committee overseeing information management priorities, policies, and compliance*
- *Participate in the design of data and metadata*
- *Agree upon the timeframe for data submission*
- *Establish the format or software with which data will be passed*
- *Identify who will do data processing, review, and quality assurance*
- *Define protocols for the role of field personnel in dataset creation*
- *Agree upon individual responsible for dataset document*
- *Prepare documentation standards*
- *Designate web page manager*

Although a CIMF coordinates a research team's data, the challenge of addressing human communication barriers still exists. A strong scientific background helps the information manager to better understand research needs and to maintain effective communications across a research group, but the role of the research team leader is critical in addressing the human communications dilemmas. Having a holistic view of the site's research work, the research leader can assist the information manager in identifying the critical components of the CIMF infrastructure that must be developed, enhanced or changed to better serve the continuously changing needs of the scientific community. This direct communication with the research leader of the community helps the information manager to better understand the vision and goals of the site's research. This interaction is critical specially when the information manager is not one of the community's researchers. In any situation, it is necessary for the information manager to have this vision for the design and implementation of a more efficient CIMF infrastructure. Also, the research leader could be the key element to resolve differences that might arise in regard to data management and sharing policy issues.

Another community challenge that the research and information management teams meet is to share data. The recognition of the need for a coherent information system as part of the organizational infrastructure increases with a better understanding of information flow. This understanding is blocked in an environment in which sharing jeopardizes a company's competitiveness in the market [10]. Resistance to sharing based on the potential for data misuse and misinterpretation is a common experience in other communities [17]. The fact that program sponsors require data sharing addresses this problem technically but not functionally. Given the contemporary consideration of intellectual property rights as well as the potential for data misuse and misinterpretation, data sharing requires not only attention from the information manager but the attention and insight of the research team leader as well.

3. TECHNICAL IMPLEMENTATION

Technical decisions impact data handling at all levels: entry, storage, conversion, processing, extraction, merging, and delivery. As an example, format is a central issue for data storage in an information system design. The decision to move data storage from a basic technical implementation to a more advanced approach adds complexity (raising significant support issues) while improving functionality. There are a variety of options for moving from flat ASCII-based files to

Table 5. Technical Elements

- *Establish where data archives will reside*
- *Create mechanisms to ensure security of the data*
- *Consider integration of personnel and bibliographic information*
- *Decide on storage format for metadata and for data*
- *Produce a data catalog that can be accessed from other information databases within the system*
- *Decide upon the role and design of the web page*
- *Create a virtual server address to give network identity*
- *Create mechanisms to facilitate the entry, update, and extraction of information such as data, metadata, publications and personnel*

relational databases today. The main purpose is to improve interface with the data and to provide an online dynamic rather than a static web interface (Table 5). Research groups often consider public domain, commercial software packages because of their low cost. Project choices are influenced by local computational infrastructure including: availability of platform support for Unix or NT systems, existing database licenses, existing file and web servers, and local expertise in interface languages like Lite, Perl or PHP. The choices involve an interplay of software elements including RDMS, web interface, and web server components (Table 6).

Table 6. Common RDMS/Web Implementation Options

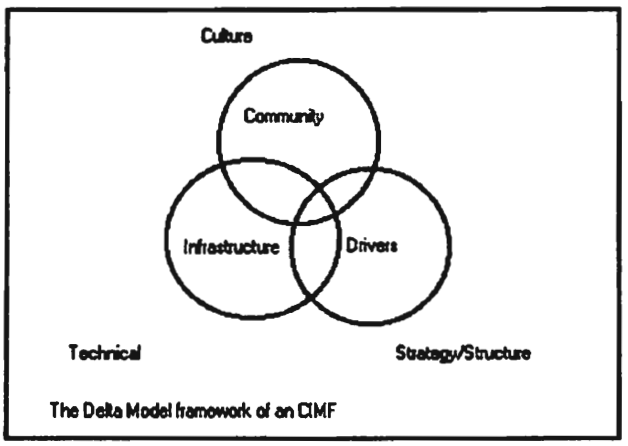
Platform	Web Server	Web Interface	DBMS
Unix	Apache	Lite	MiniSQL
Unix	Apache	Perl	MiniSQL, MySQL, Oracle
Unix	Apache	PHP	MiniSQL, MySQL, Oracle
Unix	Apache	Perl/DBI	Oracle, *others*
Unix	Apache/Tomcat	JSP (Java Servlet)	MiniSQL, MySQL, Oracle, Access, SQL Server
NT	Apache/Tomcat	JSP (Java Servlet)	MySQL, Oracle, Access, SQL Servert
NT	IIS	ASP	Access, SQL Server
NT	IIS	Perl/CGI	Access, SQL Server
NT	IIS	PHP	MySQL, Oracle
NT	IIS/Apache	Perl/CGI	Oracle

4. CONCLUSIONS

A common information management framework follows the Delta Model framework for Information Technology (IT) (Figure 1). This model has been used to induced organizational changes in which the components of technology (Infrastructure), strategy/structure (drivers), and culture (community) are considered critical for a holistic view when introducing IT to the community [2]. In a scientific community, a CIMF defines a research community and an infrastructure designed to facilitate the sharing and preservation of data for present and future generations (the goals of an LTER scientific community). A CIMF emerges in time as a result of the need of this community to make synthesis and cross-site studies. At the same time, the CIMF infrastructure promotes good science. By providing the needed technical (infrastructure) elements (centralized and integrated data set catalog, metadata, personnel data base, bibliography, research descriptions) it helps the research team and the rest of the scientific community to get a holistic view of their research and further promote synthesis and cross-sites studies. The teams' members (research, information management, and administration personnel) must aim to function in harmony with respect and mutual understanding. Even if this may be an idyllic situation, when the need to share information originates, people, especially scientists, work out their personal differences and do the work. The CIMF provides these teams with a set of common terms that facilitates technical discussions at the team level at the site, between the research and information management teams, and at the cross-site level, among the information managers.

Advanced techniques can provide more robust approaches to data query and retrieval for subsequent analysis and synthesis in a CIMF. These techniques can have a wide variety of technical implementations. Maintainability and sustainability of the CIMF software are other issues that must be addressed, in consideration of personnel and site infrastructure, as part of a total system assessment. A CIMF can be a evaluated by how well it serves its community in accomplishing its research goals. The research success of the LTER community, which manages an enormous amount of

Figure 1. The Delta Model framework of an CIM



site data in order to create relevant scientific discoveries [9], points to the LTER success at the information management level.

To be effective, a CIMF must be closely integrated with all components of an organization. It requires a commitment of resources and time not only from information managers but from the entire community.

5. ACKNOWLEDGEMENT

This work was supported by National Science Foundation (NSF) DEB-LTER 0080538, USDA Forest Service, and the University of Puerto Rico for the Luquillo LTER site, and NSF Grants OPP-96-32763, DBI-01-11544, and EIA-01-31958 for the Palmer site. We recognize the strength and uniqueness of the Long-Term Ecological Research program vision. We acknowledge the insights of the Luquillo and Palmer LTER site research communities and of the LTER information manager community past and present. Specific contributions by Donald Henshaw from the Andrews LTER site to form the common management framework concepts are acknowledged.

6. REFERENCES

[1] Baker, K.S., B.J.Benson, D.L.Henshaw, D.Blodgett, J.H.Porter, S.G.Stafford. 2000. Evolution of a Multisite Network Information System: The LTER Information Management Paradigm. BioScience 50: 963-978.

[2] Boekhoff, H. 1999. The Delta Model: A framework for the effective implementation of Information Technology to enable organizational change. World Multi Conference on Systematics Cybernetic and Informatics., Vol 1: 252 - 259

[3] Bowser, C.J. 1986. Historical data sets: lessons from the past, lessons for the future. In Pages 155-179, W.K.Michener , editor. Research data management in the ecological sciences. , University of South Carolina Press, Columbia, South Carolina, USA

[4] Brunt, J.W. 1998. The LTER network information system: a framework for ecological information management. In Pages 435-440, Aguirre-Bravo, Celedonio; Franco, Carlos Rodriguez, editors. 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.

[5] Federal Geographic Data Committee (FGDC). 1994. Content standards for digital geospatial metadata. Washington DC. http://www.fgdc.gov

[6] Ingersoll, R.C., T.R.Seastedt, M. Hartman, 1997. A model information management system for ecological research. BioScience 47: 310-316.

[7] International Organization for Standardization (ISO). 2002. http://www.iso.ch

[8] International Rice Research Institute (IRRI)'s Genetic Resources Center. 2001. Common Information Management System. http://www.irri.org/GRC/Datamanagement/IRGCIS.htm

[9] Kaiser, J. 2001, An experiment for all seasons. Science 293: 624-627.

[10] Lyyntinen, K. and D.Robey, 1999. Learning Failure in Information Systems Development, Information Systems Journal 9:85-101.

[11] Meléndez-Colom, E.C. 2001. A Definition of a Common Management Framework. http://luq.lternet.edu/publications/reports/informationmanagement/commonframe-version2.htm

[12] Michener, W. and J. Brunt, 2000. Ecological Data: Design, Management and Processing (Methods in Ecology). Blackwell Science.

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

[14] Michener, W., Stafford S.G., and Nottrot, R., editors. Proprietary Issues. 1990. Pages 16 - 17 in Proceedings of the 1990 Data Management Workshop, Snowbird, Utah

[15] National Science Foundation (NSF). 2000. Long-Term Rcological Research (LTER): Special Competition for cross-site research. Program Announcement. NSF 00-61. http://www.nsf.gov/cgi-bin/getpub?nsf0061

[16] Porter, J. H., D.L. Henshaw, and S.G. Staford. 1997. Research Metadata in Long-Term Ecological Research (LTER). http://www.computer.org/proceedings/meta97/papers/jporter/jporter.html

[17] Rocheleau, B. 1997. Governmental information system problems and failures: A preliminary review. Northern Illinois University, Division of Public Administration, Public Administration and Management: An Interactive Journal. ISSN 087-0091. http://www.hbg.psu.edu/Faculty/jxr11/roche.html.

[18] Strebel, D., B.Meeson, A.Nelson, 1994. Scientific Information Systems: A Conceptual Framework. Pages 59-85 in W.Michener, J.Brunt, S.Stafford, editors. Environmental Information Management and Analysis: Ecosystem to Global Scales. Longdon: Taylor and Francis.

[19] Water and Sewer Board of Montgomery, Alabama. 2001. Total Utility Management System: a Common Information Systems. http://pasture.ecn.purdue.edu/~aggrass/esri95/

Ecological Informatics: a Long-Term Ecological Research Perspective

William K. MICHENER & James W. BRUNT
LTER Network Office
University of New Mexico
Department of Biology
Albuquerque, NM, 87131-1091, United States

Kristin L. VANDERBILT
Sevilleta LTER Program
Department of Biology
University of New Mexico
Albuquerque, NM 87131-1091, United States

ABSTRACT

Scientists within the Long-Term Ecological Research (LTER) Network have provided leadership in ecological informatics since the inception of LTER in 1980. The success of LTER, where research projects span wide temporal and spatial scales, depends on the quality and longevity of the data collected. Scientists have devised data collection, data entry, data access, QA/QC and archiving strategies for ensuring that high quality data are appropriately managed to meet the needs of a broad user base for decades to come. The LTER cross-site Network Information System (NIS) is being developed to foster data sharing and collaboration among sites.

Keywords: LTER, ecological informatics, information management, metadata

1. INTRODUCTION

Ecological informatics is a *broad interdisciplinary science that incorporates conceptual and practical tools for the understanding, generation, processing, and dissemination of ecological data and information.* It encompasses activities that are central to the management of ecological data and information including: (1) the project or experimental design phase; (2) data design; (3) data acquisition and data management; (4) quality assurance and quality control; (5) metadata implementation; (6) data archival; (7) data access and dissemination; and (8) facilitation of data analysis [1].

Ecological informatics plays a prominent role in the United States' Long-Term Ecological Research (LTER) Program. The LTER program was conceived to elucidate multi-decadal population to ecosystem-scale phenomena [2]. Successfully addressing these phenomena requires multidisciplinary, broad-scale, and long-term approaches and perspectives, as well as the availability of data from many hierarchical scales of aggregation (e.g., population, community, and ecosystem).

LTER is supported by the National Science Foundation and includes 24 sites (ranging from urban watershed to tropical rainforest to marine and dry valley sites in Antarctica) plus a Network Office that coordinates intersite communication and other activities. During the early development of LTER in the 1980's, attention to management of LTER data and information was almost entirely based on initiative and perceived needs at individual sites. As the LTER Program matured, the number of LTER sites expanded, the size and complexity of site-specific databases increased, and synthesis and integration increased in importance. Consequently, there has been greater emphasis on coordinating site activities and developing a network information system.

Objectives of this paper are to examine ecological informatics in the LTER Network emphasizing site activities, personnel, policies, and the development of a network information system (NIS) that provides some coordination for the 24 sites. In addition, we outline some of the future plans for LTER ecological informatics.

2. CURRENT STATE OF ECOLOGICAL INFORMATICS IN THE UNITED STATES LTER PROGRAM

Site

In this section, we describe data acquisition, data entry, metadata, data archival, and data access and dissemination. Next, we discuss the types of personnel that are associated with information management, costs of informatics activities as a percentage of project budgets, and LTER data policies. Lessons learned over the past two decades of LTER Program development will be emphasized.

Data design and data management: Experiences from the LTER Program indicate that the experimental design through the analysis phase benefit

from close collaboration among scientists, statisticians, and information management personnel throughout the duration of the project. In designing databases and the structure of data sets, we conceptualize and implement a logical structure within and among data sets that can facilitate data acquisition, entry, storage, retrieval and manipulation. When creating a data set, it is useful to follow the six recommendations that are presented below [3]. These recommendations reflect experiences from managing data at individual LTER sites, as well as several of the large data archives at Oak Ridge National Laboratory. In following these guidelines, the data will be better organized, more usable, and persistent for longer periods.

1) Assign descriptive file names -- File names should be unique and reflect the file contents.

2) Use consistent and stable file formats -- ASCII file formats, or other generic formats, should be used rather than proprietary formats that may become obsolete in the future. Data should be consistently formatted, ensuring that the number and order of columns is the same throughout the data file. Within the ASCII file, fields may be delimited by commas, tabs, pipes (|), spaces, or semicolons, preferably in that order.

3) Define the parameters -- Use commonly accepted parameter names that describe what the parameter is, and denote the parameter name consistently throughout the data file.

4) Use consistent data organization -- Files should contain sets of similar measurements taken for one study, using the same methods and instruments. One large data file spanning sites and time is easier to manage than several smaller files defined by, for example, month or site.

5) Assign descriptive data set titles -- Data set titles should ideally describe the type of data, time period, location, and instruments used. Titles should be restricted to 80 characters, and should be similar to names of data files. The title "Net Nitrogen Mineralization in Grasslands at the Sevilleta LTER, New Mexico, 1999-2001" is, for example, preferable to the title "N Mineralization."

6) Provide documentation (metadata) -- Record why the data were collected, when they were collected and how they were collected. Describe the structure of the data file, and note any changes that have been made to the data.

Data acquisition and management: Acquisition of high-quality data depends on a number of factors. Data quality is inextricably linked to the knowledge and skill levels of the personnel that collect the data. Thus, the time invested in training field and laboratory personnel can pay handsome dividends in data quality. Instrument accuracy (proximity of measurement to "truth") and precision (variation of measurement within a sampling distribution) also influence data quality. Accuracy and precision are a function of the quality of the instrument, instrument maintenance and operation, and independent verification of results.

The way in which data are acquired also affects data quality by influencing the amount of human error introduced into measurements. Properly designed data sheets are inexpensive, easy-to-use, and provide a long-term hard copy of data, but may be less efficient in the field than other methods. Tape recorder data collection eliminates initial transcription errors, but problems including battery and tape maintenance, wind noise, and dust and rain in equipment may make this method undesirable. Field entry into hand-hand computers reduces data entry errors because data are entered only once, but problems include battery life and heat, cold, dust and rain that may damage computers.

For large data collection efforts like LTER, there are often many benefits associated with entering data into a commercial database management system (e.g., facilitation of data entry, sorting, security, etc.). Some of the more common DBMS types include: (1) file-system based DBMSs, which utilize files and directories to organize information; (2) relational DBMSs, which store data in tables that can be linked by key fields; (3) object-oriented DBMSs, which store data in objects that include methods for accessing and manipulating the data; and (4) hybrid DBMSs, which use a combination of relational and object-oriented schema [4].

Quality Assurance and Quality Control (QA/QC): QA/QC refers to strategies that are designed to prevent the introduction of errors, or data contamination, into a data set. Specifically, quality control mechanisms are applied during the data acquisition process to help identify data entry errors or malfunctioning instrumentation. Quality assurance mechanisms are applied after the data have been entered into a computer to identify potential outliers.
Application of quality control measures during data acquisition and data entry can greatly reduce data contamination. Simply double-checking data sheets as they are completed to confirm that all fields have been entered and that codes and measurements were entered correctly can greatly reduce errors. Enforcing standards for formats, codes and measurement units helps ensure that data are entered consistently. Illegal data filters in data entry programs that flag data not meeting variable constraints (e.g., a legal range of values) permit data entry personnel to correct typing errors as they occur or to document data points that may be incorrect due to measurement or instrumental error. Double keying of data by independent data entry technicians followed by

they were collected. Describe the structure of the data correctly can greatly reduce errors. Enforcing standards

computer verification is an ideal way to prevent data contamination.

Quality assurance measures include checking for unreasonable patterns in data, performing and reviewing statistical summaries, and assessing overall data quality. There are numerous graphical methods and statistical tests for detecting unusually extreme values of a variable (i.e. "outliers") [5]. Outliers may or may not represent data contamination and an explanation should be sought for extreme values. Statistical summaries of data can be compared to summaries from previous years to determine if central tendency or variability within the data has changed markedly. Finally, data validation through review by qualified scientists also increases confidence in data quality.

Data documentation (metadata): Metadata are defined as "data about data" or, more appropriately, "higher level information that describe the content, quality, structure, and accessibility of a specific data set" [6]. Comprehensive metadata are critical for slowing "information entropy" (Figure 1) which is defined as the normal degradation in information content associated with data and metadata over time [6]. For example, specific details are generally "lost" first, followed by more general details. Accidents, as well as retirement, career change, or death of key personnel can accelerate the rate of information loss.

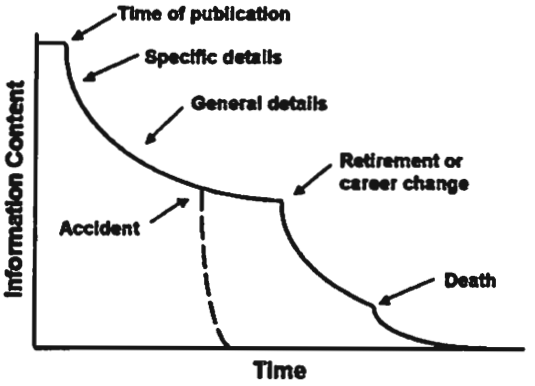


Figure 1. "Information entropy," the loss of information content of data and metadata over time [6]

Numerous metadata standards have been developed or are under development. For example, the Dublin Core, Global Change Master Directory, and others focus on providing a limited number of descriptors that primarily support data discovery. The International Standards Organization (ISO) is currently preparing a more comprehensive metadata standard for release to the international community. The ISO standard will be appropriate for many types of biological data, particularly those that have a large geospatial component.

Each site within the LTER Network is responsible for its own metadata management system, which has led to a high level of heterogeneity in site metadata content, format and storage. This heterogeneity, which ranges from ASCII text residing in a flat file to more comprehensive DBMSs, makes the development of software tools for cross-site data sharing extremely difficult. To enhance opportunities for data discovery and synthesis among LTER sites, the LTER Network has recently adopted the Ecological Metadata Language (EML) as its metadata standard. EML is based on earlier efforts to describe a standard for non-spatial ecological metadata [6]. EML is a modular and extensible means of documenting ecological data through a series of XML document types (http://knb.ecoinformatics.org). Each EML module describes one logical part of the comprehensive metadata that should be included with all ecological data sets. A user-friendly EML management tool, Morpho, has been developed which permits users to enter, edit, query, and retrieve EML documents (http://knb.ecoinformatics.org). EML documents generated by LTER sites may be stored in a centralized Metacat (http://knb.ecoinformatics.org), an XML database, to enhance data searching capabilities. Implementation of EML by the LTER Network will facilitate cross-site data synthesis as tools for sharing, integrating, and analyzing data are developed that can work together seamlessly since they rely on standard EML input.

Data archives: A data archive is a collection of data sets, usually electronic, stored in such a way that a variety of users can locate, acquire, understand and use the data [7]. The goal of ecological archives is to foster broader ecological objectives, such as regional and multidisciplinary analyses, through data sharing. A data archive must not only preserve the data, but also provide complete metadata to guide the use of the data, offer search mechanisms to allow archive users to readily identify data sets of interest, and provide a means of delivering the data to the user.

Several formal archives exist that house data related to a particular research theme. Examples of such archives include NASA's DAACs (Distributed Active Archive Centers) where multidisciplinary data of interest to global change researchers and policy-makers are stored. A web-based search engine is often used for browsing and viewing the data before ordering it via an electronic interface. Data may be made available on a variety of media.

LTER project databases provide some of the functionality of archives. For instance, the Sevilleta LTER stores data locally and makes data accessible via the Internet. To prevent data loss, incremental tape backups are made

daily of all Sevilleta data; that is, any changes to data sets are recorded on backup tapes each night. Monthly, the Sevilleta database is completely backed up on tape. A complete set of a month's backup tapes is stored off-site. Sets of tapes are recycled every three months. Some GIS and remote sensing data sets are stored on CD, due to their size.

Data access and dissemination: Almost all LTER data are easily accessed through the LTER web site (http://www.lternet.edu/data/). More than 2,000 data sets can be accessed from this central location. Although most data sets are managed at individual sites using a relational database management system, publicly available data are typically in comma-delimited ASCII text format. Data and metadata may be accessed independently or as a bundled package, depending on the practices at individual LTER sites. Other data, including climate and hydrological data, may be acquitted from a centralized server that harvests records from multiple sites. These data may be acquired in a variety of formats (e.g., daily values to monthly averages). In addition, value-added programs (e.g., graphical analyses) are available to facilitate interpretation.

Data are also available through individual sites. For example, data may be acquired directly from the Sevilleta's web site (http://sevilleta.unm.edu/). Anyone downloading Sevilleta's data is asked to email the Sevilleta information manager with details of how the data were used; such policies encouraging attribution of data to the site and data set authors are commonplace in the LTER Network.

Data discovery and dissemination are facilitated through an online data catalog at the LTER web site. Many LTER data sets are also cataloged in the Global Change Master Directory and NASA DAACs. Furthermore, most sites have begun to invest heavily in developing data discovery, dynamic querying, and analysis tool s that facilitate research.

People and the Cost of Information Management

Most LTER sites employ a full-time information manager who is responsible for the design and implementation of the information management system. This individual frequently has interest and training in both the ecological and computer sciences (e.g., programming, DBMS). Other personnel requirements vary considerably from one site to another and partially depend upon the degree of institutional support available for system administration and other ancillary activities. Consequently, other personnel may include a system administrator, programmer, data entry specialist, and GIS specialists.

The cost of managing data at individual LTER sites is estimated to range from 10 to 20% of the annual budget, although this figure may actually range up to 40% or more if all technician time devoted to data management and GIS activities are included. The lower figure of 10% may best represent newer sites, where there are less data to manage, or to established sites, where considerable data management support may also come from other sources. Budget items primarily include salaries, communication costs (e.g., internet service providers), supplies, and minor equipment. Major pieces of equipment (e.g., servers, RAID units) are generally not included in annual budgets, but are often acquired through specific equipment grants and other sources.

Data Policies

LTER sites are dedicated to having all long-term data sets and key short-term data on-line and available to the scientific community and general public in a timely fashion (see http://www.lternet.edu/data/netpolicy.html). It is also generally recognized that researchers have an obligation to publish LTER data and that LTER investigators must have a reasonable opportunity for first use of data they collected. The LTER data access policy attempts to address these competing demands.

In return for providing access to LTER data, many LTER sites require that the LTER program and the individual investigators receive credit for their efforts. For example the Konza Prairie LTER site (KNZ) includes the following statement in their access policy:

> "Finally we ask all publications, reports and proposals who use any data from KNZ acknowledge/cite the KNZ program using the following statement: "Data for XXX was supported by the NSF Long Term Ecological Research Program at Konza Prairie Biological Station"; where XXX is the list of data sets used in the publications, reports or proposals."

The LTER data access policy has resulted in many significant changes, including an increase in the number of cross-site and multi-site analysis and synthesis activities, better data and science via enhanced attention to the quality of data and metadata, greater reliance on web-based databases to facilitate data accessibility, and growing recognition that data and metadata are valuable products of the research enterprise (in addition to publications).

3. THE LTER NIS INFORMATION SYSTEM

The Long-Term Ecological Research (LTER) Network Information System (NIS) is a cooperative, federated database system (Baker et al. 2000) supporting more than 1200 scientists and students that are investigating complex ecological phenomena at the 24 sites. The LTER NIS was established with the overall goal of implementing a transparent shared virtual environment

that is high-performance, distributed, and secure. The LTER NIS builds and expands upon successful grass-roots efforts that are initiated at one or more of the 24 LTER research sites. Currently, the LTER NIS consists of operational modules that are located at the LTER Network Office at the University of New Mexico as well as prototype modules that are being developed at individual sites. These modules are designed to facilitate intersite communication and technology transfer, and support research. Modules include a personnel directory, electronic mailing lists, bibliography, data catalog, site description directory, climate database, and aboveground primary productivity database [8].

Modularity and extensibility have been critical to the success of the LTER NIS. Modules may be developed at individual or cooperating sites, and often undergo extensive revision as they are implemented across the network. The "LTER cycle" is one whereby new software and information management approaches are experimented with at a few sites. In some cases, the experiment may be perceived as a failure and the software or information management approaches are abandoned. Conversely, the experiment may be viewed as successful and the software or information management approaches spread to other sites where they continue to be used and evaluated. The process is never static; new software and approaches continue to be evaluated and only those approaches that are robust and successfully scale to the full network are broadly adopted (John Porter, pers. communication;[8]).

Several important lessons have been learned from LTER NIS efforts. First, meeting standardized goals with a variety of site-specific solutions has built strength into the LTER Network. Second, leaving data at sites where they can best be managed, while making them accessible via a common interface represents a viable solution. Third, without an adequate information system and the requisite personnel, the significant effort required to manage large and complex data sets can present a substantial barrier to intersite research and synthesis. Finally, the successful evolution of the LTER NIS can be related to four key design elements: (1) partnerships and a central focus that are well defined; (2) explicitly stated individual site responsibilities; (3) modularity in software design; and (4) development of independent prototypes that undergo rigorous testing for scalability and sustainability prior to broad acceptance [8].

4. FUTURE OF ECOLOGICAL INFORMATICS IN LTER

Future efforts in LTER ecological informatics will focus on forming and expanding partnerships to facilitate data discovery, methods standardization, data sharing and integration, and data archival. There are currently many barriers to these activities. Discovery of ecological data remains a challenging endeavor, particularly for data that are not "registered" in a data directory and have not been cited in scientific peer-reviewed publications. Numerous opportunities will be explored for directly linking LTER data to national and international data cataloging efforts.

Non-standardized methods can inhibit data integration; conversely, adoption of standard methods supports intersite comparisons, facilitates documentation of the methods for publications and metadata, and can reduce project costs [9]. Collection of data in a consistent manner is vital. LTER scientists have published books like Standard Soil Methods for Long-Term Ecological Research [10] in an effort to encourage consistent data collection among sites. Similar efforts to standardize net primary production and other data types are underway.

Other barriers to data sharing and integration include inadequate and incomplete metadata [6], as well as concerns over publication rights, credit and other "rewards" for database development [11]. Many valuable ecological data sets are lost because researchers have few incentives to preserve their data once results have been published. An exception is the Ecological Society of America's Ecological Archives, which publishes peer-reviewed and cited data papers that consist of data sets and associated metadata deemed to be of significant interest to ecologists, thereby rewarding scientists for investing time in data archival. It is anticipated that significant future effort will be devoted to publishing LTER data in outlets like Ecological Archives.

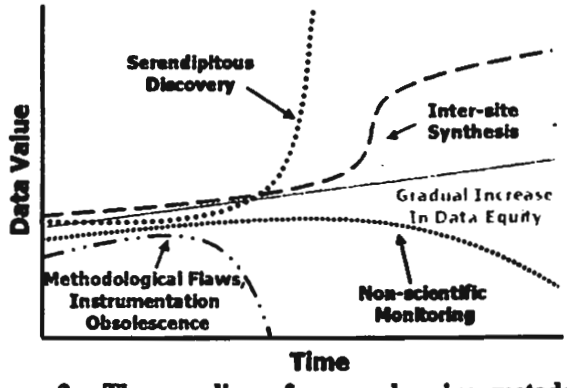


Figure 2. The coupling of comprehensive metadata with data generally facilitates the gradual increase in value of a data set over time.

Data sharing and collaboration can be further enhanced through the development of data distribution and archive centers, as well as information analysis centers (e.g., see [12]). Coupling comprehensive metadata with data in a data archive generally enhances the value of a data set over time (Figure 2). In some cases, well-documented

data will lead to an important serendipitous discovery whereby the data rapidly accrue new value. Similarly, utilization of documented data for an intersite synthesis project (e.g., broad-scale comparison) can increase the value of a particular data set for both the short term and long term. In contrast, however, well-documented data can lose value. For instance, discovery of methodological flaws or obsolescence of instruments may render a data set worthless. Likewise, monitoring without a clear scientific objective can lead to a data set that decreases in value over time. The next decade of LTER will likely see much more effort focused on partnering with environmental data archives and facilitating data integration and synthesis efforts through workshops, new technologies, and other mechanisms.

Rapid increases in environmental data holdings, as well as increasing demand for the data and information that are required for informed environmental decision-making, will require additional advances in data mining and knowledge discovery tools [13]. Particular emphasis in LTER will likely focus on value-added data products that can be more rapidly understood and used by decision-makers, resource managers, and educators. One example of progress in this area is the Ecology Explorers program at the Central Arizona Phoenix LTER site that directly involves K-12 students in the research process, including information management activities (see http://caplter.asu.edu/explorers/index.htm).

New challenges will confront LTER as new data types are increasingly integrated with ecological data. For example, increased collaboration with social scientists will require that information managers address the attendant data confidentiality issues of socioeconomic data. In addition, increased miniaturization of environmental sensors coupled with developments in wireless communication have enormous implications for LTER and will result in exponential increases in the volumes of data that sites manage. Such massive data streams will require new automated approaches for processing data streams and visualizing the resulting information.

Acknowledgments

LTER is supported by the National Science Foundation.

5. References

[1] Brunt, J.W. 2000. Data management principles, implementation and administration. Pages 25-47 in: *Ecological Data: Design, Management and Processing* (W.K. Michener and J.W. Brunt, eds.), Blackwell Science, Oxford, Great Britain.

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

[3] Cook, R.B., R.J. Olson, P. Kanciruk, and L.A. Hook. 2000. Best practices for preparing ecological data sets to share and archive. *Bulletin of the Ecological Society of America* 82:138-141.

[4] Porter, J.H. 2000. Scientific databases. Pages 48-69 in: *Ecological Data: Design, Management and Processing*. (W.K. Michener and J.W. Brunt, eds.), Blackwell Science, Oxford, Great Britain.

[5] Edwards, D. 2000. Data quality assurance. Pages 70-91 in: *Ecological Data: Design, Management and Processing* (W.K. Michener and J.W. Brunt, eds.), Blackwell Science, Oxford, Great Britain.

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

[7] Olson, R.J. and K.A. McCord. 2000. Archiving Ecological Data and Information. Pages 117-141 in: *Ecological Data:Design, Management and Processing,* (W.K. Michener and J.W. Brunt, eds.), Blackwell Science, Oxford, Great Britain.

[8] Baker, K.S., B.J. Benson, D.L Henshaw, D. Blodgett, J.H. Porter, and S.G. Stafford. 2000. Evolution of a multisite network information system: the LTER information management paradigm. *BioScience* 50:963-978

[9] Michener, W.K. 2000. Research design: translating ideas to data. Pages 1-24 in: *Ecological Data: Design, Management and Processing* (W.K. Michener and J.W. Brunt, eds.), Blackwell Science, Oxford, Great Britain.

[10] Robertson, G.P., D.C. Coleman, C.S. Bledsoe, and P. Sollins (eds.). 1999. *Standard Soil Methods for Long-Term Ecological Research.* Oxford University Press, New York.

[11] Porter, J.H. and J.T Callahan. 1994. Circumventing a dilemma: historical approaches to data sharing in ecological research. Pages 193-203 in: *Environmental Information Management and Analysis: Ecosystem to Global Scales* (W.K. Michener, J.W. Brunt and S.G. Stafford, eds.), Taylor and Francis, Ltd., London, England.

[12] Scurlock, J.M.O., R.J. Olson, R.A. McCord and W.K. Michener. 2002. Environmental data banks: archiving ecological data and information. Pages248-259 in: *Encyclopedia of Global Environmental Change* (E. Munn, ed), Vol. 2: The Earth system: biological and ecological dimensions of global environmental change, Ecosystems Section (H. Mooney and J. Canadell, eds.). John Wiley, Chichester.

[13] Michener, W.K. 2000. Ecological knowledge and future data challenges. Pages 162-174 in *Ecological Data: Design, Management and Processing* (W.K. Michener and J.W. Brunt, eds.), Blackwell Science, Oxford, Great Britain.

[1] Brunt, J.W. 2000. Data management principles,

[13] Michener, W.K. 2000. Ecological knowledge and

Integrating Ecological Data: Tools and Techniques

John H. PORTER
University of Virginia, Charlottesville, VA 22903 USA

and

Kenneth W. RAMSEY, Jr.
New Mexico State University, Las Cruces, NM 88003 USA

ABSTRACT

Integration of data is critical to achieving new levels of understanding of ecological systems and processes. Typically, data integration is achieved only through a painstaking manual process that rules out large-scale integration. We believe that many of the techniques related to uncertain reasoning (fuzzy logic, Baysian networks, and evolutionary algorithms) and data mining might be usefully applied to ecological data integration. Here we present two case studies. One characterizes a traditional approach to integration. The second focuses on using software system integration to integrate geospatial and research data, along with providing data discovery services. We discuss those case studies where advanced techniques might prove useful and where modifications are needed to support scientific research.

Keywords: Data Integration, GIS, Uncertain Reasoning, Data Mining, Information Management Systems, Ecoinformatics

INTRODUCTION

Advancement of ecological science is increasingly dependent upon our ability to integrate data from diverse sources. The understanding of ecological processes at the spatial scales of the landscape, region and globe and at the temporal scales of the decade, century and the millennium require data that span these scales. Such data go beyond the collection abilities of any individual investigator or single research project and thus require data integration [1].

Although need to integrate diverse ecological and environmental data is not new, the opportunity to do so is. Traditionally, ecologists have not shared data, nor have they had adequate incentives to do so [2]. Traditionally, data has been collected, analyzed and publications prepared by a single individual or small group of researchers, typically a professor and associated graduate students. As discussed by Strebel et al. [3] and Michener et al. [4], over time most of this data have been lost through a slow process of "data decay" as, in the absence of metadata, our ability to locate or interpret data has been diminished or lost. However, in the last decade there have been sociological and technological developments that have led to increased sharing of data. These include the role of the Internet and World-Wide Web in lowering the costs of sharing data [5], the implementation of information management policies by research projects that define the responsibilities of data providers and data users [2], the recognition by funding organizations such as the National Science Foundation that data products, as well as publications, are valuable results from research projects and even the development of "data journals" such as the Ecological Society of America's new journal "Ecological Archives."

With ecological data becoming more readily available, the critical issue becomes not how we can get it, but rather what we can do with it. In this, ecology is not alone. As was stated in a recent *Science* article: "Computer technology has facilitated the collection of data so well that now, in a growing number of fields, the availability of data is no longer (or soon will not be) the limiting factor for addressing fundamental scientific questions. Paradoxically, the new limitation is computer technology: Only with the help of computer science can we make sense of the masses of data that computers have enabled us to collect, and share and discuss the data with colleagues around the globe. The challenge now is to design aids to help us comprehend data so complex or interconnected that we cannot organize, integrate, or understand it alone" [6]. Here we present two case studies: an example of project-specific data integrationand an example of system-based integration of spatial and thematic data and discuss research directions for developing techniques to address large-scale integration needs.

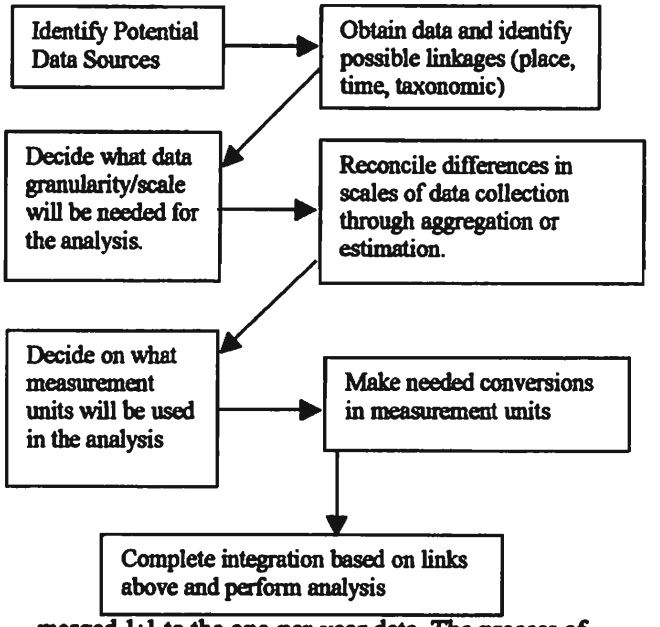
Scientific data systems are required to deal with data that is both more complex and incorporates more (necessary) inconsistencies than traditional business databases [7-9]. However, the development of "data warehouse" and "data mart" systems in business [10], coupled with the development of standards for ecological metadata [4] diminish these differences in important ways and lead to opportunities for cross-fertilization of disciplines.

CASE STUDY: PROJECT-SPECIFIC INTEGRATION

Project-specific integration of ecological data remains

primarily a manual process (Figure 1). Once a scientific hypothesis is defined, an investigator or group of researchers will identify possible datasets, either datasets that they maintain themselves, can obtain from published literature or (more recently) can download from an on-line database. Once obtained, decisions are made about parameters that will be used to link data and about the scale and level of aggregation. For example, if some data are collected on an hourly basis while other data are collected only once per year, the hourly data must be aggregated to provide an annual value that can then be

Figure 1: The Data Integration Process



merged 1:1 to the one-per-year data. The process of integration demands painstaking concentration on maintaining data quality as misleading results can occur if errors are introduced at any point in the analytical process.

The identification of linkages is a key step in data integration. For ecological data, the most important linkages tend to be temporal and spatial. Every observation was made at a point in space at a particular time, although the availability and representational form this information takes may be highly variable. Taxonomic linkages are also possible, although evolving taxonomic standards often makes this difficult because species names are not invariant over time [11].

A study underway at University of Virginia investigates the relationship between specific meteorological and climatological factors and primary production of vegetation as part of the Virginia Coast Reserve Long-term Ecological Research (VCR/LTER) project [12]. Here data on climate and productivity is being integrated based on indices of time and space. After discussion, the investigators determined that, although hourly and daily meteorological measurements were available, monthly and annual aggregations of data were the most likely to yield results, as most of the productivity data are at those time scales. The ongoing integrative analysis depends on individual investigators using a suite of traditional software tools, such as statistical packages and spreadsheets, individually preparing temporally-indexed data structures (Table 1). Although software is used, the process remains primarily manual, with each decision being made by experts familiar with a specific portion of the data and the final integration being performed by a group of scientists during the course of intensive analysis sessions.

Table 1: Software used for project-specific integration

Software	Use
WWW browsers (Internet Explorer, Netscape)	Data discovery and download
Spreadsheet Software (Excel, Quatro)	Data entry and display, some limited graphing. May also be used for some final analyses using data imported from statistical packages
Statistical Packages (SAS, SPSS)	Data merging, analysis and graphics
Teleconferencing (NetMeeting, Polycom and other H.323 compatible products)	Communication between collaborating researchers. Especially important is T.120 application sharing capabilities.

CASE STUDY: SYSTEM-BASED INTEGRATION

System-based integration depends on developing standardized data sources and software systems that support their integration. As in the previous case study, maintaining data quality is a primary issue. This demands standards for content and format be enforced at the level of the individual databases to be integrated. It also demands special attention to the fields in the databases that ultimately enable the linkages. Erroneous data in these fields can result in, at best missing data and at worst inclusion of inappropriate data in analyses.

One example of a system-based integration project is underway at New Mexico State University (Figure 2). The Jornada Basin LTER (JRN) and the U.S. Department of Agriculture (USDA) Agricultural Research Service (ARS) Jornada Experimental Range (JER) are currently developing three separate databases that require integration to facilitate information management and improve discovery of and access to JRN and JER research datasets and associated metadata. The databases include a Research database for storing research datasets and documentation, a centralized Geographic Information System (GIS) database for storing GIS and Remote

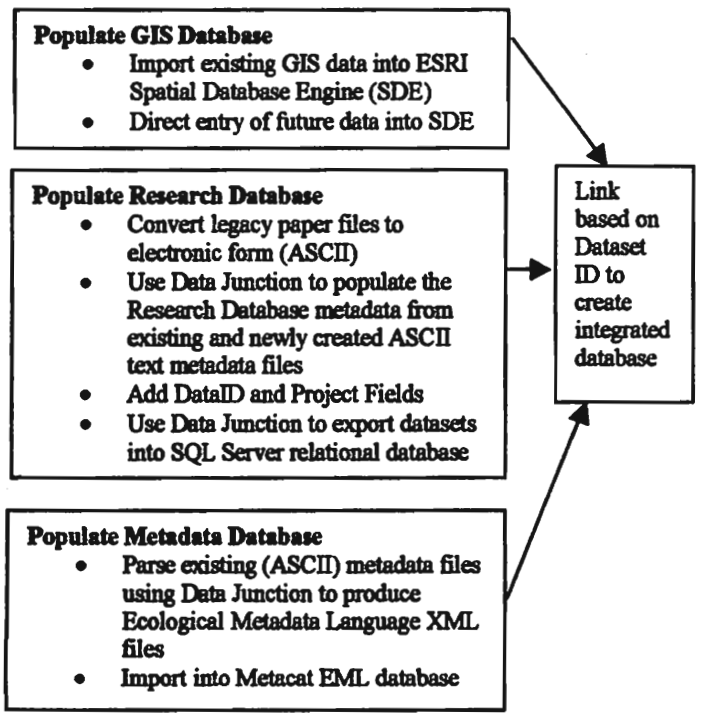
the relationship between specific meteorological and developing three separate databases that require

Sensing data, and a metadata search engine (Metacat) database for querying metadata. These databases are being integrated to ensure that data and metadata within these applications are synchronized and accurate as well as to decrease redundant data storage as much as possible. Through integration of these databases, applications developed to support the automated Jornada Basin Information Management System (IMS) will be much more powerful and useful to information managers, researchers, and policy makers.

Populating the Research Database

All JRN research data and associated metadata are, and will be, archived as ASCII text files. These text files will be modified to permit the Research Database to be integrated with the GIS and Metacat databases. Modifications include adding project and dataset ID fields and ensuring that all data and metadata files conform to the standardized file formats. JER data and documentation will be converted to standard formatted ASCII text files as JER adopts the JRN IMS for archiving and managing research datasets and documentation. Some JER datasets date back to 1912 and many will need to be transcribed into electronic format.

Figure 2: Developing an Integrated GIS and Research Database



All archived JRN and JER text files will be parsed into a Research Database running Microsoft SQL 2000 Server. The entire process of parsing the text files to create or update the Research database will be scripted and automated using Data Junction Enterprise Integration Studio 7.51 (DJ) (Table 2).

Table 2: Software used in system-based integration

Software	Use
Data Junction Enterprise Integration Studio http://www.datajunction.com	Parsing of text documents into structured forms
ESRI ArcGIS http://www.esri.com	Input, retrieve, and manage spatial and remote sensing data and associated metadata stored in the Geographic Information System (GIS)
ESRI Spatial Database Engine (SDE) http://www.esri.com	Interface between relational database, GIS software and distributed applications
Microsoft SQL 2000 Server http://www.microsoft.com	Relational database used to support the Research Database and SDE
Metacat http://www.ecoinformatics.org	Customized metadatabase for use with EML-compliant metadata

Populating the GIS Database

The GIS database will provide centralized storage and access to spatial data such as remote sensing data covering the Jornada Basin or GPS data collected by JRN and JER researchers and technicians. Previously, all spatial data and metadata have been stored on the GIS Specialist's computer. Sharing spatial data and documentation in the past has been somewhat difficult. Now, by using newer spatial database software, ESRI ArcSDE 8.1 (SDE), it is much easier to share spatial information. The centralized GIS will be an integral part of the IMS; allowing spatially referenced querying of research data, remote sensing data, and associated metadata.

Existing remote sensing data imagery, coverages, and shapefiles will be imported into SDE using tools included in ESRI ArcGIS 8.1 software. Future GIS data and metadata collected will be stored and maintained in SDE.

Populating the Metacat Database

The Metacat database provides redundant storage of dataset documentation to allow the metadata to be queried using web-based applications [13]. Metacat uses the Ecological Markup Language (EML) to store metadata and for communication with the Metacat server using the XML protocol [14]. EML provides a standard exchange format for exchanging ecological metadata. EML will help reduce and spread the costs of application development by allowing development to be spread amongst the ecological research community. As new Metacat enhancements or EML based applications are released, they could be readily added to the IMS with minor modifications. The Jornada Basin Metacat server

will be linked to a LTER Network Office Metacat super node in the future. Metacat will provide another method for researchers to discover JRN and JER datasets; improving visibility and usefulness of JRN and JER scientific datasets.

Data Junction (DJ) will be used to create and populate EML compliant XML files, which subsequently can be imported into the Metacat database. DJ will also be used to automate this process in order to keep metadata stored in Metacat synchronized with metadata from the archival project and dataset documentation ASCII text files.

Application Development: The Need for Integrated Databases

The IMS will provide a web-based dynamic, interactive mapping and querying application that will be a powerful tool for JRN and JER information managers, researchers, and visitors. Some uses of the application include facilitation of research site selection and approval as well as automation of storage and retrieval of restricted and unrestricted spatial and non-spatial data and metadata into and from the IMS. Such an approach assists in land management decisions and eco-health evaluations, as well as monitoring of spatial and temporal vegetation changes. In order to achieve this level of functionality, the Research, GIS, and Metacat databases must be integrated, or linked.

Challenges: There are several limiting factors, or challenges, that had to be addressed in order to integrate the Research, GIS, and Metacat data to support the IMS, as well as develop the IMS. These challenges include adopting EML and Metacat, centralizing the GIS, and populating the Research database.

By adopting standards such as EML and utilizing Metacat software, it is hoped that the IMS can be enhanced in the future by using EML-based applications developed by other ecological research organizations. By developing the IMS to utilize EML, our development efforts can subsequently be used and enhanced by other organizations. Using tools such as DJ and XML Spy IDE Suite 4.1 for data conversion and extending the EML schema can greatly reduce development time and costs.

Linkages: Dataset IDs will be the common link that relates the Research, GIS, and Metacat databases. Other common thematic, temporal, and keyword fields will be created in the databases to allow queries of spatial and non-spatial data and associated metadata. If needed, pivot or lookup tables would be created in the databases using the common ID, thematic, temporal, and keyword fields. Stored views will be used to simplify development efforts where possible.

Steps for Integrating Research, GIS, and Metacat Data

The following list contains the steps required to integrate the Research, GIS, and Metacat data prior to the development of the IMS application:

1. Assign and add dataset IDs to dataset data and documentation files.
2. Assign project IDs and add dataset and project IDs to project documentation files.
3. Parse project and dataset data and documentation files to populate IMS database.
4. Perform QA/QC on parsed IMS database and correct any errors found in the archived ASCII text files and/or parsing scripts.
5. Parse project and dataset data and documentation files again if any errors were found during QA/QC of the IMS database.
6. Import or load existing spatial and remote sensing data and metadata into the GIS database.
7. Perform QA/QC on GIS data and metadata stored in the GIS database and correct any errors directly within the GIS database.
8. Add dataset IDs to research site location layers and features attribute tables stored in the GIS database.
9. Parse project and dataset documentation ASCII files and related IMS and GIS database tables to create EML XML DTD or schema files.
10. Perform QA/QC on parsed EML XML files and correct any errors found in the archived ASCII text files and/or parsing scripts.
11. Parse project and dataset data and documentation ASCII files and related GIS database tables again if any errors were found during QA/QC of the EML XML files.
12. Import EML XML files into the Metacat server database.
13. Import remote sensing images into the GIS database.

Upon completion of the sequence of steps, the IMS and GIS Internet and Intranet applications can then be developed and implemented.

The JRN and JER have completed the planning and evaluation stages of this IMS project. Data integration is currently underway. The project is designed to be modular to allow for prioritizing and planning the project development cycle.

RESEARCH AREAS FOR DATA INTEGRATION

The difficulty of a data integration project is directly proportional to the scale of the integration process. As the case studies above show, integration in specialized projects and systems can be accomplished using conventional software tools and information systems. However, for global-scale projects, which must integrate data from a huge number of data sources, these methods are not practical. The most common approach is to focus

proportional to the scale of the integration process. As the

on a few, large, very standardized, data sources. However, this approach excludes the vast majority of data sets, which are not standardized.

The challenges posed by dealing with large amounts of heterogeneous data are increasingly being confronted by developing techniques of data warehousing and data mining [15, 16]. Chen [10] calls for the application of data mining techniques to be used in conjunction with techniques for uncertain reasoning, such as fuzzy logic, genetic algorithms, Bayesian networks and rough set theory. Below is a brief outline of how a few of these techniques might be used in integrating ecological data.

Fuzzy Logic: Fuzzy logic allows probabilistic statements to be made about the true state of a variable. For example, for land cover classifications derived from remotely sensed data, you might conclude that there is a 75% chance that an area is forest and a 25% chance that the area is a shrubland. Traditional forms of analysis demand that we go with our best guess. However, with fuzzy logic, analyses can also consider additional guesses, each with its associated probability.

For ecological research, where qualitative determinations for land cover, habitat, community type and even taxonomic identity [11] are often suspect, (especially when integrating data from diverse sources) data integration incorporating fuzzy logic offers opportunities to incorporate a larger amount of information into analyses. Its use is not widespread in ecology, although it has been used in ecological applications in the context of remote sensing [17], ecological decision support [18] and modeling and prediction [19-21]. In our case studies, above, fuzzy logic could be applied to geographical locations that have varying degrees of specificity, and to land cover designations.

Evolutionary Algorithms: Evolutionary algorithms use a process of highly iterative trial and error to derive functional relationships and estimate parameters. Although primary uses have been primarily for developing search strategies and modeling, use of evolutionary algorithms holds promise for "harmonizing" data sources where the functional relationships between two ways of measuring an environmental variable are unclear. In our case studies, these techniques could be applied to harmonizing measurements taken at different scales or using different methodologies.

Data Mining Techniques: Traditional database approaches have had great difficulty dealing with heterogeneous information sources. However, techniques used in the rapidly evolving field of data mining can help to surmount these difficulties [15, 16]. Clustering, classification, and association rules have obvious uses in the data discovery process. However, they can also be used for at least partial automation of data integration by helping to identify similar variables in different datasets. Some of these techniques are widely used in ecology, although typical uses are more oriented towards data analysis than data integration.

Meta-Analysis: In addition to the techniques listed by Chen [10], meta analysis provides tools for a different approach to integration. Ecological meta-analysis integrates results from previously published studies to attack broader questions and to strengthen individual conclusions [22]. The effect sizes observed from a variety of studies, each using different data sources and methods can be statistically combined to reach new conclusions.

CONCLUSIONS

Use of advanced data mining and techniques for dealing with uncertainties would be a powerful approach for the facilitation of ecological synthesis, but such developments need to be coupled with specific enhancements that ensure use by the scientific community. As previously noted, scientists place a high value on data quality. Unlike some disciplines where a final product is evaluated on its intrinsic merits (regardless of origin), scientific research products are evaluated based primarily on the methods and data used to produce them. Complex, integrated datasets pose problems because a full explanation of data sources may be impossible and reviewers and readers need assurance that results are real, and not artifacts of the integration process.

Enhanced analysis systems are required to support use of integrated data. These systems need to make datasets auditable, so that each datum used in an analysis can be traced back to its original source and transformations reproduced. Reproducibility is critical to developing trust in a scientific product.

Second, tools need to be developed that facilitate sensitivity analyses, wherein specific data sources can be added or subtracted from an analysis. This allows researchers to determine whether particular data sources have undue influence on the final result, or whether their conclusions are robust with respect to changes in data sources.

Finally, visualization techniques can clarify the roles of individual data sources. Animated graphs which highlight specific data sources make it possible to review a large number of data sources in a short period of time. This approach assures that patterns are discernable within, as well as between, data sources.

Meeting the challenges inherent in large-scale data integration is the subject of ongoing research in both the computer science and ecological communities.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grants No. DEB-0080381 and DEB-0080412. William Michener, James Laundre and Karen Baker provided constructive comments on the manuscript.

REFERENCES

[1] National Research Council, Bits of Power: Issues in Global Access to Scientific Data. National Academy Press, Washington, D.C., 1997.

[2] J.H. Porter and J.T. Callahan, Circumventing a Dilemma: Historical Approaches to Data Sharing in Ecological Research. in W.K. Michener, S. Stafford and J.W. Brunt eds. Environmental Information Management, Taylor and Francis, Bristol, PA, 1994, 193-203.

[3] D.E. Strebel, B.W. Meeson and A.K. Nelson, Scientific Information Systems: A Conceptual Framework. in W.K. Michener, S. Stafford and J.W. Brunt eds. Environmental Information Management, Taylor and Francis, Bristol, PA, 1994, 59-85.

[4] W.K. Michener, J.W. Brunt, J.J. Helly, T.B. Kirchner and S.G. Stafford, "Non-Geospatial Metadata for the Ecological Sciences". Ecological Applications, Vol. 7, No. 1, 1997, pp. 330-342.

[5] B.R. Schatz and J.B. Hardin, "NCSA Mosaic and the World Wide Web: Global Hypermedia Protocols for the Internet". Science, Vol. 265, 1994, pp. 895-901.

[6] B. Hanson and R. Coontz, "A Computer Science Odyssey". Science, Vol. 293, No. 5537, 2001, pp. 2021.

[7] J. Pfaltz, Differences between Commercial and Scientific Data. in Scientific Database Management, a Report to the National Science Foundation, 1990.

[8] R.J. Robbins, "An Information Infrastructure for the Human Genome Project". IEEE Engineering in Medicine and Biology, Vol. 14, No. 6, 1995, pp. 746-759.

[9] J.H. Porter, Scientific Databases. in W.K. Michener and J. Brunt. eds. Ecological Data: Design, Processing and Management, Blackwell Science Ltd., London, UK, 2000.

[10] Z. Chen, Data Mining and Uncertain Reasoning: An Integrated Approach. John Wiley and Sons, New York, 2001.

[11] D. Maier, E. Landis, J. Cushing, A. Frondorf, A. Silberschatz, M. Frame and J.L. Schnase. Biodiversity and Ecosystem Informatics: Report of an NSF, USGS, NASA Workshop on Biodiversity and Ecosystem Informatics Held at NASA Goddard Space Flight Center, June 22 - 23, 2000., NASA, Beltsville, MD, 2001, 36.

[12] B.P. Hayden, R.D. Dueser, J.T. Callahan and H.H. Shugart, "Long-Term Research at the Virginia Coast Reserve: Modeling a Highly Dynamic Environment". Bioscience, Vol. 41, 1991, pp. 310-318.

[13] M.B. Jones, C. Berkley, J. Bojilova and M. Schildhauer, "Managing Scientific Metadata". IEEE Internet Computing, Vol. 5, No. 5, 2001, pp. 59-68.

[14] R. Nottrott, M.B. Jones and M.P. Schildhauer, Using XML-Structured Metadata to Automate Quality Assurance Processing for Ecological Data. in Proceedings of the Third IEEE Computer Society Metadata Conference, (Bethesda, MD., 1999), IEEE.

[15] M.J.A. Berry and G.S. Linoff, Mastering Data Mining: The Art and Science of Customer Relationship Management. John Wiley & Sons, Inc., New York, 2000.

[16] M.A. Bramer, Knowledge Discovery and Data Mining. The Institution of Electrical Engineers, Herts, UK, 1999.

[17] G. Metternicht, "Assessing Temporal and Spatial Changes of Salinity Using Fuzzy Logic, Remote Sensing and GIS. Foundations of an Expert System". Ecological Modelling, Vol. 144, No. 2-3, 2001, pp. 163-179.

[18] D.M. Stoms, J.M. McDonald and F.W. Davis, "Fuzzy Assessment of Land Suitability for Scientific Research Reserves". Environmental Management, Vol. 29, No. 4, 2002, pp. 545-558.

[19] A. Pistocchi, L. Luzi and P. Napolitano, "The Use of Predictive Modeling Techniques for Optimal Exploitation of Spatial Databases: A Case Study in Landslide Hazard Mapping with Expert System-Like Methods". Environmental Geology, Vol. 41, No. 7, 2002, pp. 765-775.

[20] L.O. Odhiambo, R.E. Yoder, D.C. Yoder and J.W. Hines, "Optimization of Fuzzy Evapotranspiration Model through Neural Training with Input-Output Examples". Transactions of the ASAE, Vol. 44, No. 6, 2001, pp. 1625-1633.

[21] V. Gomez and A. Casanovas, "Fuzzy Logic and Meteorological Variables: A Case Study of Solar Irradiance". Fuzzy Sets and Systems, Vol. 126, No. 1, 2002, pp. 121-128.

[22] C.W. Osenberg, O. Sarnelle, S.D. Cooper and R.D. Holt, "Resolving Ecological Questions through Meta-Analysis: Goals, Metrics, and Models". Ecology, Vol. 80, No. 4, 1999, pp. 1105–1117.

Efforts to Link Ecological Metadata with Bacterial Gene Sequences at the Sapelo Island Microbial Observatory

Wade M. SHELDON
Department of Marine Sciences
University of Georgia
Athens, Georgia 30602, USA

and

Mary Ann MORAN
Department of Marine Sciences
University of Georgia
Athens, Georgia 30602, USA

and

James T. HOLLIBAUGH
Department of Marine Sciences
University of Georgia
Athens, Georgia 30602, USA

ABSTRACT

The existence of public databases for archiving genetic sequence data, such as GenBank and the Ribosomal Database Project, coupled with the availability of standardized sequence alignment and comparison tools has led to rapid advances in the field of bacterial genetics and systematics. Many microbial ecologists now routinely submit gene sequences obtained from environmental isolates, clones, and bands excised from electrophoretic gels to public sequence databases. As the amount of environmental sequence data in these systems has increased, ecologists have begun using sequence databases for broader classes of studies, such as biogeography and community ecology. Unfortunately, the general lack of documentation and data quality control standards has resulted in many sequences being entered without appropriate metadata, effectively orphaning records from their ecological context information and making comparisons impossible.

In order to address the shortcomings of public sequence databases, an independent 16S rRNA sequence database was recently developed at the Sapelo Island Microbial Observatory (SIMO) in Georgia, USA. The database was created to store complete information from all SIMO research activities using a hierarchical structure designed to reflect the actual flow of information from sample collection through final publication. By incorporating key fields from external databases, such as GenBank, the SIMO database is able to serve both as an independent research tool for SIMO scientists and as a reference source for SIMO data stored in other databases.

Keywords: Database, Metadata, 16S rRNA, Bacteria, Genomics, GenBank, Microbial Observatory

1. INTRODUCTION

The creation of public repositories for archiving genetic sequence data, such as GenBank and the Ribosomal Database Project, coupled with the availability of standardized sequence alignment and comparison tools has led to rapid advances in the field of bacterial genetics. In most research laboratories, comparison of 16S rRNA gene sequences has replaced traditional culture and phenotype-based methods for bacterial classification (Giovannoni and Rappe, 2000; Gonzalez et al., 2000). Most major microbiology journals, such as Microbiology and Applied and Environmental Microbiology, now require GenBank accession numbers as a pre-requisite for all sequences published in manuscripts. Not surprisingly, public

sequence databases have been growing at exponential rates in recent years. As of February 2002, there were approximately 15,465,000 sequence records in GenBank alone (NCBI, 2002).

Many microbial ecologists now routinely submit gene sequences obtained from environmental isolates and even bands excised directly from electrophoretic gels to public sequence databases for archival and analysis. As the amount of environmental sequence data has grown, these investigators have also begun to apply these databases to new classes of ecological studies, such as microbial community ecology and biogeography. Unfortunately, these efforts are often hindered by the lack of robust documentation standards for environmental sequence data. In the absence of standards, most sequences are entered without appropriate ecological and methodological metadata, effectively orphaning records from their research context information and making comparisons impossible.

The lack of complete metadata is particularly problematic for sequences obtained by direct amplification of DNA obtained from environmental samples (i.e. environmental sequences), because unlike sequences from bacterial isolates there is usually no reference material available for independent verification. In addition, recent reports of significant data quality control problems (Karp et al., 2001) and lapses in sequence format enforcement (Karp, 2001) for GenBank records further underscore the need for improving metadata standards for genetic sequences used for ecological research.

2. SIMO 16S rRNA DATABASE

A workshop was held in August 2000 during the LTER All-Scientists meeting to facilitate inter-site comparisons of microbial community composition and biogeography (Hollibaugh and Priscu, 2000). The workshop resulted in a draft proposal for the creation of an integrated environmental sequence database as an extension or companion to the existing public database resources to address the issues listed above. While this proposal has yet to be funded or implemented, many of the principles outlined were recently incorporated in a new 16S rRNA genetic sequence database developed at the Sapelo Island Microbial Observatory (SIMO) in Georgia, USA.

The SIMO database was designed to store complete information from all SIMO research activities in a hierarchical structure modeled after actual laboratory workflow patterns. Each successive step in the data hierarchy references the preceding steps, so that the full research context of all data is maintained from sample collection through analysis and final publication. Figure 1 illustrates the conceptual design of the database, and Table 1 lists the database table entities used to implement the model with commercial relational database management software (Microsoft SQL Server™). The full entity-relationship diagram is also available on the SIMO World Wide Web site (http://simo.marsci.uga.edu/public_db/).

The 'Samples' table represents the top of the database hierarchy, and also functions as the primary link between sequence data, spatial (i.e. geographic) and environmental information, and ancillary analyses and data. The subsidiary data tables, such as 'Source', 'Sequence', and 'SeqComparison', store both research data and laboratory management details, including storage amounts and locations, user-assigned codes or aliases, and processing notes. The inclusion of these laboratory attributes allows the database to serve an additional role as a laboratory information management system.

Metadata are primarily associated with individual data records by references to fields in related lookup tables (e.g. study sites, macro-environments, microenvironments, standard methodology, personnel records, storage locations). This design simplifies data entry (see below), minimizes redundancy, and encourages research standardization. In addition, the inclusion of metadata fields as foreign keys in data tables facilitates fine-grained queries and data sub-setting by metadata category, which would be far less effective for free-text fields without controlled vocabularies.

In addition to SIMO data, key fields from external databases are also stored in the database (e.g. GenBank accession numbers, Georgia Coastal Ecosystems LTER Project sampling site codes, bibliographic citations). These links provide access to additional supporting information and dramatically increase the scope of the SIMO database.

3. DATA ENTRY AND INTEGRITY

Both the database model and user interface are designed to maximize the quality of data and metadata stored in the system. Declarative referential integrity constraints are defined for all table relationships to prevent orphaned records and duplicate entries. Strict content control of key organizing terms is maintained by limiting data entry to lists of values that reference fields in lookup tables that are maintained exclusively by SIMO administrators and database managers. When list-based entries are not possible, input masks and validation functions are used to ensure that data entered by users is appropriate for the corresponding field.

A web-based data entry application allows users to enter data in a series of discrete steps (e.g. sample information, clone and isolate information, sequence data, phylogenetic information), which supports the discontinuous nature of molecular genetics research and allows delegation of data entry tasks to multiple students and technicians. The web application automatically generates appropriate hyperlinks to support both sequential data entry (e.g. samples through phylogenetic information) and multiple data entry into a single table (e.g. sequences from one or more clones) in any combination. The rigid application of referential integrity and hierarchical design of the database ensures that data cannot be added unless supporting information is already in place.

A separate administrative application was also developed using Microsoft™ Access 2000. This application allows authorized SIMO staff and database managers to update metadata and lookup tables, restore records deleted by mistake, and control the content of web application selection lists by setting values in management fields included in database tables. A granular, role-based security scheme is used to control access to all database entities, which prevents unauthorized personnel from using the administrative application to modify the database.

4. SIMO METADATA

Various metadata standards and data formats have been developed to facilitate exchange of genetic sequence data among researchers and various computer programs. Four of the most popular formats are as follows:

- FASTA file format[1]
- GenBank flat-file format [2]
- Biopolymer Markup Language (BIOML)[3]
- Bioinformatic Sequence Markup Language (BSML)[4]

These standards describe the formatting of the sequence data and provide varying levels of support for basic research metadata (e.g. researcher contact information, comments on methodology and analysis, and sequence annotation), but none provide the level of detail required to support ecological research and syntheses. In particular, support for specific research origin descriptors (e.g. site characteristics, sampling design, detailed methodology) and supplemental descriptors (quality control measures, reference materials, publication history) are totally lacking (see Michener, 2000, for a discussion of ecological metadata requirements).

In the absence of an established ecological metadata standard for documenting bacterial gene sequences, metadata information is currently displayed in summary format along with sequence details on the SIMO web site. Dynamic hyperlinks are created to provide access to the corresponding sampling site information, contact information for the responsible researcher, and ancillary data available in the Georgia Coastal Ecosystems LTER Project database.

Sequence data is also provided in the terse FASTA file format, with the SIMO unique identifier and web site URL listed in each sequence comment line. We are currently evaluating the suitability of emerging standards for ecological metadata, such as EML (Ecological Metadata Language; http://ecoinformatics.org/), for providing complete metadata for SIMO sequences in a more autonomous and computer-parseable format suitable for both data archival and dissemination.

5. FUTURE DIRECTIONS

The initial implementation and population of the SIMO database has been very successful, but a

[1] National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/BLAST/fasta.html)
[2] National Center for Biotechnology Information (http://www.ncbi.nih.gov/Genbank/)
[3] Proteometrics, LLC (http://www.bioml.com/BIOML/)
[4] LabBook, Inc (http://www.labbook.com/products/xmlbsml.asp)

number of potential barriers remain to realizing its full potential. We initially planned to provide SIMO unique identifiers and database URL pointers with all data submitted to GenBank to support bi-directional searching based on paired database keys. Unfortunately, our attempts to date have met with mixed success. NCBI GenBank personnel frequently remove SIMO pointer information from submissions made using BankIt (i.e. GenBank web form submission tool) based on poorly qualified criteria, such as concern over long-term URL stability. In contrast, sequences submitted as batches using the stand-alone Sequin program often retain their pointer information, indicating they receive less individual scrutiny.

To work around this limitation, efforts are now underway to list SIMO as a non-bibliographic sequence data provider in LinkOut, a referral system paired with GenBank as part of Entrez (http://www.ncbi.nlm.nih.gov/Entrez/). LinkOut is specifically designed to provide hyperlinks to ancillary information for sequences stored in GenBank, making it an ideal candidate for managing external SIMO data links. Unfortunately few microbial ecologists use the Entrez system because of its traditional association with medical journals (e.g. PubMed) rather than scientific research databases.

A combined approach that includes listing of SIMO sequences in LinkOut and a researcher education campaign will probably be required to accomplish a functional linkage between SIMO sequences stored in GenBank and the SIMO 16S rRNA database.

6. CONCLUSION

The vast amount of information available in public gene sequence databases represents an important resource for microbial ecologists studying biogeography, community structure, and biodiversity. Analytical tools and techniques that are currently available to mine these databases allow researchers to integrate data from across the world and ask truly large-scale questions that could never be addressed by a single researcher or laboratory. Without access to high quality ecological metadata, though, the scope of these studies will be severely limited.

The SIMO 16S rRNA database provides a working model of a system for coupling genetic information with ecological metadata. The query pages on the SIMO web site best illustrate the potential power of this model, providing researchers the ability to search for sequences by various phylogenetic characteristics, ecological characteristics, or an intersection matrix of combined characteristics (http://simo.marsci.uga.edu/public_db/). A national database incorporating key features of the SIMO database could dramatically enhance the potential for large-scale studies on bacterial biogeography and community structure.

7. ACKNOWLEDGEMENTS

We wish to thank Scott Federhen of NCBI for information on the Entrez LinkOut program, and Karen Baker and two anonymous reviewers for their editorial comments that helped improve this manuscript. This material is based upon work supported by the National Science Foundation under Cooperative Agreements #MCB-0084164 and #OCE-9982133.

8. LITERATURE CITED

Giovannoni, S. and Rappe, M., 2000. Evolution, Diversity, and Molecular Ecology of Marine Prokaryotes. In: D.L. Kirchman (Editor), Microbial Ecology of the Oceans. Wiley-Liss, New York, pp. 47-84.

Gonzalez, J.M., Simo, R., Massana, R., Covert, J.S., Casamayor, E.O., Pedros-Alio, C. and Moran, M.A., 2000. Bacterial community structure associated with a dimethylsulfoniopropionate-producing North Atlantic algal bloom. Applied and Environmental Microbiology, 66(10): 4237-4246.

Hollibaugh, J.T. and Priscu, J.S., 2000. Microbial Biogeography: Cross-site comparison of aquatic systems, LTER All-Scientists Meeting, Snowbird, Utah, USA.

Karp, P.D., 2001. Many GenBank entries for complete microbial genomes violate the GenBank standard. Comparative and Functional Genomics, 2(1): 25-27.

Karp, P.D., Paley, S. and Zhu, J., 2001. Database verificiation studies of SWISS-PROT and GenBank. Bioinformatics, 17(6): 526-532.

Michener, W.K., 2000. Metadata. In: W.K. Michener and J.W. Brunt (Editors), Ecological Data - Design, Management and Processing. Methods in Ecology. Blackwell Science Ltd., London, pp. 92-116.

NCBI, 2002. GenBank Overview. National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html).

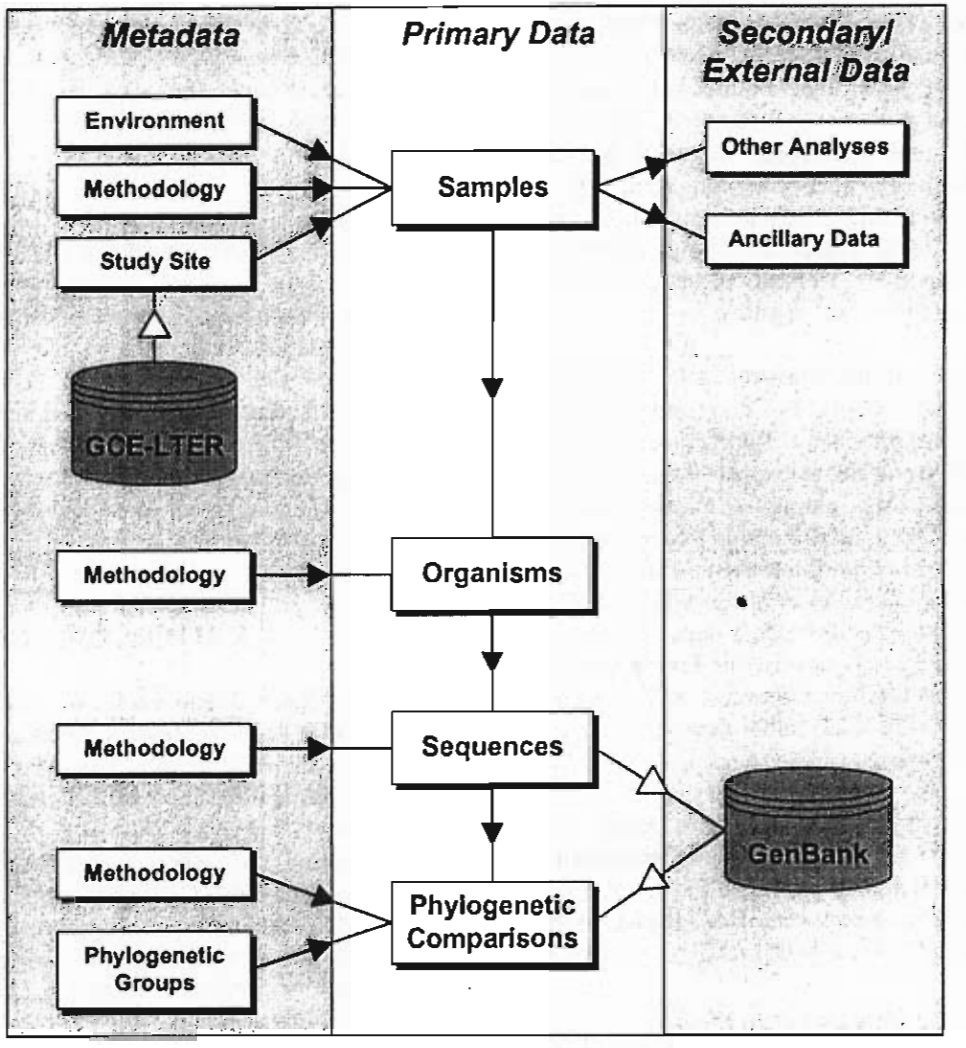


figure 1. Conceptual diagram of the Sapelo Island Microbial Observatory 16S rRNA database illustrating the relationships and directions of information flow between metadata, primary data, and secondary data entities (note that some related metadata tables are combined for clarity). Solid arrows represent internal relationships with referential integrity constraints, hollow arrows represent external relationships based on shared keys, and cylinders represent external databases. GenBank is the NIH genetic sequence database, and GCE-LTER is the Georgia Coastal Ecosystems LTER Project metadata database.

Table 1. Relationships among the primary tables comprising the SIMO 16S rRNA sequence database. The full entity-relationship diagram is available on the WWW at http://simo.marsci.uga.edu/public_db/.

Data Tables	Descriptions	Related Tables	Related Table Descriptions
Samples	Information about physical samples collected for DNA extraction or bacterial isolation	Site	Sampling site details, including geography
		Environment	Macro-environment designation and description
		MicroEnv	Microenvironment designation and description
		Zone	Marsh zone designation and description
		Sampling	Sampling methodology description
		Storage	Storage method and location details
		People	Contact information for the responsible party
Source	Information about DNA sources (bacterial isolates and vectors into which extracted and PCR-amplified DNA was cloned)	**Samples**	Sample from which the sequence source was derived
		SourceType	Category of sequence source (lookup table)
		SourceTechnique	Methodology used for isolation or cloning and maintaining the organism
		People	Contact information for the responsible party
		Storage	Storage method and location details
Sequence	Information about nucleotide sequence data obtained by DNA analysis of 16S rRNA genes	**Source**	Source from which the DNA sequence was obtained
		SeqAnalysis	Sequence analysis methodology and instrumentation
		Primers	Description and vendor information for DNA primers used in the analysis
		People	Contact information for the responsible party
SeqComparison	Information about phylogenetic comparisons between SIMO sequences and other sequences published in GenBank, made using sequence alignment and analysis programs such as BLASTN and FASTA	**Sequence**	Sequence used for comparisons
		Programs	Information about the program used for the comparison
		BactDivision	Bacterial division (taxonomic lookup table)
		People	Contact information for the responsible party
Phylogeny	Information about the most closely related organisms identified by sequence comparisons	**SeqComparison**	Comparison reference

Designing Web Database Applications for Ecological Research

Dan J. SMITH
Barbara J. BENSON*
and David F. BALSIGER

Center for Limnology, University of Wisconsin-Madison
Madison, Wisconsin 53706, USA
* corresponding author

ABSTRACT

Many sites conducting ecological research must routinely manage a diverse suite of datasets and make them accessible to researchers. This paper presents an approach to creating an ecological data query system that dynamically creates predefined dataset query interfaces for managed datasets. The query interfaces are created from stored dataset metadata and query creation metadata and include only those field selections and filtering options identified as relevant for the specified dataset. Using only stored metadata for query interface creation provides an extensible and scalable metadata framework, creating a standardized, yet robust, system that can handle diverse datasets. This metadata framework also provides opportunities for developing centralized querying of remote datasets and multi-site data exchange through the use of predefined dataset query interfaces.

Keywords: Ecoinformatics, Metadata, Dynamic Database Access, Ecological Metadata Language (EML), XML, Query System, Web Database Interface.

1. INTRODUCTION

The World Wide Web has created a ubiquitous environment for access to scientific information. At the North Temperate Lakes Long Term Ecological Research (NTL-LTER) site [1][2], the information management staff has developed a Java application [3] that provides a web interface to a relational database housing a diverse collection of long-term physical, chemical and biological limnology datasets. A query engine driven by metadata tables stored in the same database generates the user interface.

Researchers value the ability to easily extract only the data relevant to a particular research question and have that data in a format readily transported to their favorite analysis packages. This dynamic query application permits the selection of dataset attributes of interest and allows the construction of conditional filters on dataset attributes to accurately specify the data that are relevant to the researcher. The user can select from a variety of output formats as well as request sorting of the data by designated fields.

The Long Term Ecological Research (LTER) [4] program is a network of 24 sites funded by the National Science Foundation to study long-term phenomena in ecosystems. The LTER sites are committed to making data publicly available. Many sites meet this commitment by serving text data files from their websites. This approach limits the ability of a user to select only the data of interest. Other sites create a static template for each data set to permit dynamic database queries. Our approach was to generate a web form for dynamic queries "on-the-fly" for all data in the database using a standard description of the data that was also stored in the database. This design not only provides dynamic access to the database but also is very extensible because of the ease of adding the description of a new dataset to the database.

The standard description of the data was based on an emerging metadata standard for ecological data, Ecological Metadata Language (EML) [5][6]. In this paper we describe some extensions that were made to the metadata standard to support query features such as filtering. We also describe the architecture of the software and discuss its features and opportunities for future development that would support multi-site interoperability and data exchange. The extension to multi-site data sharing will be made possible by the acceptance of a common metadata standard.

2. QUERY SYSTEM DESIGN

Several desired features guided the creation of this query system. The system needed to meet the following criteria:

- deliver consistent and current data on demand
- handle diverse table layouts for different datasets
- export EML metadata for each managed dataset
- allow retrieval of only data relevant to the user
- provide the data in various output formats including XML
- maintain data usage and access statistics
- include access control on datasets and users
- scale well to multiple sites to allow data interchanges

The prior system at NTL-LTER, publishing datasets as static data files on the Web, did not meet any of these goals. A system capable of meeting these goals must have access to a stored description of the dataset and present the user with various options for data retrieval based on that description.

The role of the query program (Figure 1) is to act as a conduit between the user and the datasets stored in the database. It guides the user to a particular dataset, presents information about that dataset, translates user choices into a query, and returns a recordset adhering to the conditions of the user request.

The heart of our query prototype is the interface that is dynamically generated for each dataset. This interface has four main tasks:

- providing the acceptable query specification choices to the user (the search framework) for each dataset
- accepting and verifying user input into an input framework
- creating a query to be executed by the query engine
- receiving and formatting the results of an executed query

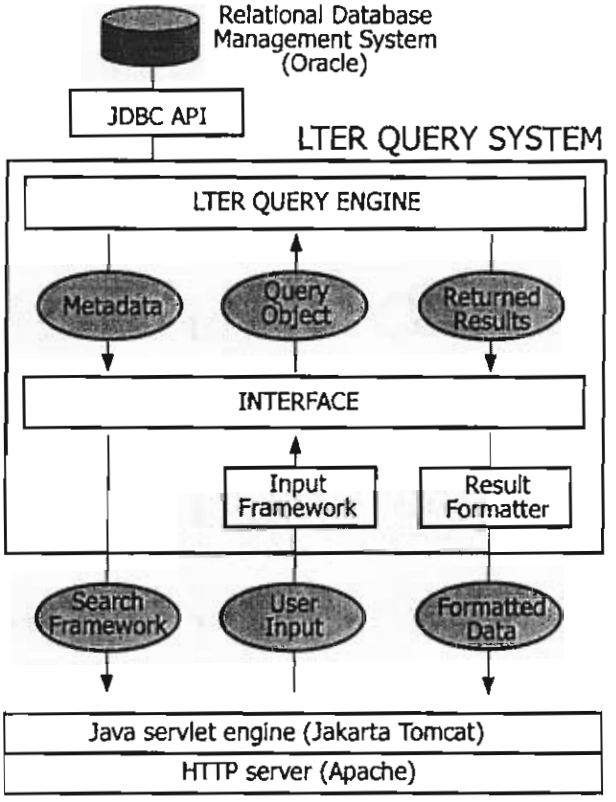


Figure 1. Schematic diagram showing the objects and components of the dynamic query application.

This system was designed to be fully extensible and to easily accommodate new datasets with no additional programming or software engineering. In other words, the same software must be able to handle any type of dataset and its associated filters, conditions, and sorting based on stored, known information. In order to accomplish this functionality, information about the datasets must conform to a standard representation. Thus, the first task in implementing the query prototype was to create a standard description of the various datasets contained in the database.

3. METADATA

The datasets managed by the NTL-LTER are diverse in the number of fields and field types. In order for the query system to handle these different datasets, a common descriptive language is required. This language allows the query system to understand the table structures of all the datasets in order to query them appropriately and successfully. The language effectively encapsulates the diverse datasets into uniformly accessible entities.

Metadata Content Standard

To perform this encapsulation, our implementation uses a beta release of the Ecological Metadata Language (EML) version 2 as our data definition language. EML, an emerging standard in ecological metadata, is being developed by the National Center for Ecological Analysis and Synthesis in collaboration with the Long Term Ecological Research program and the San Diego Supercomputer Center. The content standards in EML are based on standards for ecological metadata developed by the ecological community (Michener et al. 1997 [7]) and are implemented in Extensible Markup Language (XML) [8]. EML consists of a set of modules, each of which specifies one logical part of the complete metadata.

Table 1. Description of EML modules used by the query system.

EML module	Description
EML-project	describes the research context in which the dataset was created
EML-dataset	contains general information about the dataset; includes abstract, contacts, coverage, etc.
EML-attribute	describes attributes (variables) in a dataset; includes type, range, definitions of coded values, etc.
EML-party	describes a responsible party (person or organization); contains detailed contract information for the party

We used the project, dataset, attribute, and party modules (Table 1) from EML as part of the content standard for our metadata. Some additional information that has traditionally been part of the NTL-LTER metadata is not included in EML. These fields were retained in the NTL-LTER metadata content.

The dataset module of EML was reproduced and extended to form the center of the entity-relationship model. Entities in our model reference this dataset object through the use of a unique dataset ID. This linkage provides easy access to all the information the program uses to perform queries on specific datasets.

To describe the data in each dataset, we included the attribute module of EML. Linked to specific datasets, the attribute entity describes the type of data stored in the dataset as well as information about measurement units, precision, enumerated domains, and data quality for each dataset attribute. To store information about people involved with each dataset, we constructed a party entity based on the EML party module. This entity is linked to a role entity that associates individual investigators and contact people with a given dataset. This type of information is required to export our traditional metadata included with our datasets in EML form.

Extended Metadata

The EML-based dataset descriptions do not include instructions to the query system as to how the metadata and the datasets described therein should be handled. Filtering and display information is still needed in order to present the user with appropriate selection criteria. Accommodating this additional handling information required an extension to the stored metadata. Including this additional information in the extended metadata enabled the development of a deterministic query

system, fully database driven, which is capable of interpreting and creating the query interface in a standard way for all datasets. This well-defined behavior is needed in multi-site scenarios.

The query interface provides the user with a selection of fields to be included from a dataset, and in addition, offers the user a selection of data filters appropriate to that dataset. Fields within a dataset that might be useful and appropriate as data filters for that dataset are specified within the extended metadata. In addition, the extended metadata stores information as to which of the five available types of filters will be used in each case. Storing this information as extended metadata allows the system to accommodate new filters on additional attributes easily and facilitates the quick integration of new datasets and their associated query interfaces into the program.

Table 2. Description of metadata extensions used by the query system to enable dynamic query interfaces.

Metadata Extensions	Description
Dataset Extensions	data access restrictions based on specified dates
Attribute Extensions	restrictions on availability of fields for querying and suppression of null records for sparse datasets
Filtering Extensions	module created to store filter definitions (Figure 3)

To accommodate the extended metadata, we expanded our previous entities along with creating several new entities (Table 2), which are not part of the current version of EML. One important new entity contains the query display information that controls the data filtering available in a specific query interface. To accommodate a variety of filter types, this entity matches a specific type of filter with a particular dataset attribute, allowing for fine-tuned filtering descriptions that are specific to individual attributes and datasets. There are currently five types of filters possible for a query:

- enumerated codes displayed as a list (e.g. selecting a subset of lakes)
- date range
- year range
- range on attribute values
- multiple selections from a list of values; displayed as two boxes: the original list of codes and the selected list (e.g., selecting species names)

Additional information was needed to provide access restrictions on certain data by date as well as information about handling sparse datasets. The access restrictions were included in the previously created dataset entity since they apply to all records in the dataset. Information about display and result suppression for individual attributes was placed in the attribute entity and is applicable within each dataset that includes that attribute.

4. PROGRAM COMPONENTS

The query program creates a search framework that provides the user with query choices, an input framework that accepts and verifies the input from the user, a database query based on user input, and formatted results returned to the user.

Search Framework and Filter Creation

The task of the search framework (Figure 2) generated by the query interface is to provide the possible appropriate query condition choices to the user. The search framework creation begins with the gathering of information about the data contained in the dataset of interest. This information is stored in the standardized metadata in the database. The data description for the chosen dataset is queried from the database backend and delivered to the interface.

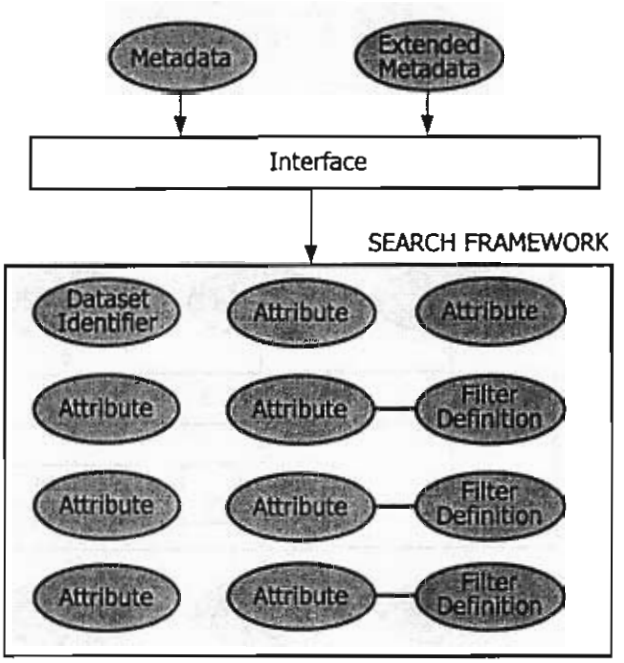


Figure 2. Schematic diagram of the search framework and its contained objects. The program interface gathers metadata for a particular dataset to create a search framework with the appropriate attribute definitions and associated filters.

Queries on datasets include pre-specified filters (Figure 3) to allow the user to limit retrieval to only relevant data. The "filter type" specifies which type of filter has been defined and how the filter will be rendered to the user. It also specifies if the defined values object or dates object will be used for this filter. The defined values are enumerated codes derived from EML attribute metadata, database tables, or stored queries. Defined values objects are used in enumerated code or multiple selection queries. The dates object specifies the acceptable range of dates for temporal queries. Filter title and filtering phrases are used to describe to the user the function of the filter and the necessary input from the user.

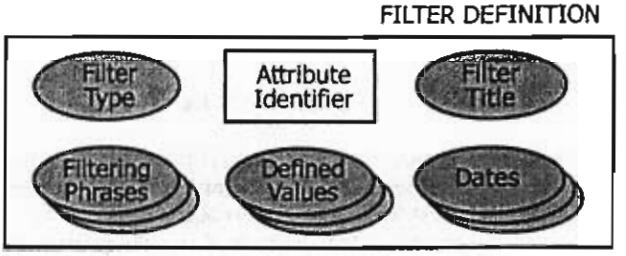


Figure 3. Diagram showing the filter object created by the interface and placed into the search framework.

After obtaining the dataset and attribute descriptions, the interface has structural knowledge of the specific fields present in the dataset, along with the description of the dataset attributes. The interface does not yet contain information as to which types of filtering criteria should be attached to those fields. The interface again queries the database backend and requests the extended metadata that contains the filtering conditions that will be presented to the user for this dataset.

These filter definitions are extracted from the extended metadata and attached to a specific attribute. After filtering information has been added, the search framework is completed and the user is presented with a choice of field selections and filters for the dataset. In the current implementation, this framework is transformed into a query form that is presented to the user as a web browser form. In a multi-site scenario, this search framework would be encoded and presented to a remote site in an XML representation, or alternatively could be derived by the remote site from downloaded metadata retrieved in XML.

User Input

The user input to the interface contains various types of information ranging from the fields selected for output to filters on ranges and enumerated codes. To manage all this incoming information, the interface must create an input framework for the given dataset. This user-input framework tells the interface what type of input data to expect in order to match the incoming information about the user-specified query conditions to the specified dataset.

The user-input framework is constructed in much the same way as the previous search framework using metadata and extended metadata from the database backend. In addition, the system retrieves any information about restrictions placed on access to the data in this dataset and incorporates this into the framework. This user-input framework can be thought of as an inverse of the earlier constructed search framework. Whereas the search framework specifies to the user what type of input is acceptable for a given dataset, the user-input framework specifies to the query interface what type of data it might be receiving. The user input is then matched against this framework to create an object that represents the user's query.

The user input in a multi-site system would be encoded by the central query system into an XML encapsulated query object. This encapsulation of the query object is possible because of the self-descriptive nature of XML data. The query object would be transported directly to the appropriate remote query engine using the SOAP [9] protocol for object messaging.

Query Creation and Execution

After the user input has been parsed and matched against the input framework, a query object exists in the interface. The query object specifies which data fields are to be retrieved, along with various selection criteria for the fields. In addition, any system-imposed selection criteria such as access restrictions or null suppression on sparse datasets are added to the query object as well.

The query object is systematically parsed and translated into a standard SQL command. This newly created SQL query is passed to the query engine where the query is run against the data source containing the dataset. The tuples returned by the query are then handed back to the interface for output.

Output of Query Results

The formatting of the output is the final task of the interface. The result output routine takes the set of tuples returned by the query and transforms them into a file or stream in the user-specified format. The choice of output format was presented earlier to the user in the search framework, collected as user input, and stored in the query object.

Because the data are returned from the database simply as a set of tuples, any output format is easily assembled. The interface processes the returned recordset and converts it into the chosen format, adding data description and formatting to the file. The output formats included in the current implementation are XML, web browser display, Microsoft Excel, and comma-delimited text. The web browser display conveniently breaks the result set into smaller screens (pages) and includes easy navigation between pages. The metadata for the dataset can optionally be included in the result file, and due to the design of the system, the metadata included in the XML output format has the added benefit of conforming to the EML standard.

5. EXTENSIBILITY

This data query system is driven entirely by tables residing in a database backend. This design provides extensibility for the creation of new interfaces, as can be seen in the ease of adding additional datasets, modifying query interfaces, and in the organization of datasets into hierarchies.

One approach for implementing dynamic database access is to create a unique or customized data query program for each individual dataset based on it's own distinctive layout. Our prototype, instead, utilizes a single program that can access any number of datasets as long as the system has access to their metadata descriptions. Extending the query interface for a new dataset table simply requires placing the metadata for the dataset into the database. This universal program approach allows for any number of additional datasets to be added to the query engine without any additional program development. In addition, when datasets evolve or are modified, the associated metadata can be easily updated and immediately reflected by the search program.

This system also allows updating of handling and filtering options without any changes to the program. The advantage here is that it eliminates the need for creating or updating multiple programs when datasets or query interfaces change.

With a large number of datasets, organization is essential for allowing researchers quick and straightforward access to the relevant data. We group the datasets by category. By adding a category entity to the extended metadata, an extensible system was created for classifying individual datasets. The entry document for the data catalog on the web is generated dynamically from the information stored in the category entity in the database.

A project level entity was also added to encapsulate the various categories. Unrelated projects can share the same query system while their data are kept separate and consistent. Different user interfaces are also possible using templates and cascading style sheets on a per project basis. Our prototype implementation is currently querying and serving NTL-LTER data, Biocomplexity Project data, and Microbial Observatory data.

6. MULTI-SITE INTEROPERABILITY

An important goal is to design information systems that allow users to find relevant data quickly. When a researcher is looking for data that are hosted at a separate site, the data acquisition process can become slow. Allowing researchers to run queries across multiple sites in search of data from a central location would streamline this process.

Based on our evaluation of this prototype, we believe there are three main requirements for achieving consistent query interfaces for datasets hosted across multiple LTER sites. This prototype fulfills several of these initial requirements for multi-site searches. The first requirement is a standard descriptive language for the data. For ecological research, EML appears to be the emerging data definition language of choice. When EML is finalized, data descriptions for datasets from multiple sites will be available to any system that conforms to this standard.

The second requirement is a standard format for data representation. In our prototype, we have implemented four distinct data representations for data output. XML is the most extensible and allows for the inclusion of EML metadata within the actual dataset data. For these two reasons, our preferred approach, which is implemented in this prototype, is to serialize the returned records in XML and include the EML description of the data. One problem with XML data retrieval is that it is not as widely accepted as comma delimited text files. For this reason, our prototype allows researchers to choose from several formats for the results of their queries. We expect that XML will gain more widespread usage as more applications become XML compliant.

The third requirement is a descriptive language that allows the query systems to define filters for datasets. Filters could be created dynamically from the EML data representations, but we feel that this approach is not optimal. It is difficult to add filters for a dataset that would be relevant to the researcher without knowing more information than is presented by EML descriptions of attributes. For complex datasets, a large number of different filters could be possible, leading to the user being overwhelmed with unneeded criteria selections. For these reasons, a standardized language to describe relevant filters is needed. Our prototype uses the extended metadata to implement the description of possible filters. Multiple sites will be able to interpret the extended metadata in a deterministic way, enabling the creating of a standard query interface. While this is a start toward meeting this third requirement, the extended metadata that we have created is not a standard nor is it implemented in any other system.

Upon the development of a standardized language for implementing selection and filtering descriptions, a central site will be able to generate query interfaces, as defined by the data managers at the remote sites, for remote datasets. With the ability to create a query interface for a remote dataset, coupled with the ability to serialize query results, multi-site or centralized querying of remote LTER datasets will be attainable.

7. CONCLUSIONS

The NTL-LTER query system resulted from a need to create a system that was more robust than the previous system of downloadable static data files. This was accomplished by shifting from static data files to data delivered directly from the backend data warehouse. This query system has achieved many of our intended goals and has created opportunities for future expansion and development. One significant area that this system does not address is the querying and retrieving of spatial data.

At NTL-LTER, spatial data are not stored in the relational database. The current dynamic database prototype provides a link to a separate system for accessing spatial data which are stored in a variety of vector and raster formats. The type of query interface optimal for spatial data would extend beyond that implemented in our prototype to include a flexible array of spatial query and mapping tools. This approach might incorporate web-based geographic information system (GIS) software, such as ArcIMS from Environmental Systems Research Institute (ESRI). The goal of such a system would be to permit the user to query and retrieve spatial data using both spatial and non-spatial search criteria. In addition, the results of spatial data queries should be available both as map graphics in the user's web browser and in a standard vector or raster format suitable for later analysis. The beta version of EML 2.0, upon which the NTL-LTER query system is based, does not contain a module for spatial data although standards exist for spatial metadata [10][11].

The challenge in the creation of this system was defining how query interfaces for datasets would be created. Our solution to this problem uses a metadata framework to describe the datasets and query interfaces. All query-related information needed in the query creation process is based on and interpreted from stored metadata. A key feature of the design was to extend EML to include metadata that describe display and filtering information. This extended information enables the creation of unique predetermined query interfaces for each dataset. The reliance on stored metadata allows for changes to and additions of datasets and query interfaces without additional software engineering costs. The metadata definitions used in the program also provide the needed information to export EML upon request.

XML documents conforming to standard Document Type Descriptors, such as those for EML, will soon be a widely used exchange medium for data and metadata for ecological research. These common data definition formats will allow for rapid and standardized information exchange and inter-site operability within ecological research networks. Since the query program described here is based on these standards, it is staged to deal with these future multi-site scenarios.

8. ACKNOWLEDGEMENTS

The authors gratefully acknowledge Matt Jones for comments on our EML implementation and the reviewers of this paper for their insightful critiques. This paper is based on research supported by a grant from the National Science Foundation (DEB-9632853).

9. REFERENCES

[1] Benson, B. J. 1996. The North Temperate Lakes LTER research information management system, Proceedings of Eco-Informa '96: Global Networks for Environmental Information, volume 11, Environmental Research Institute of Michigan, Lake Buena Vista, Florida, 4-7 November 1996, pp. 719-724 .

[2] North Temperate Lakes LTER http://lter.limnology.wisc.edu/

[3] NTL-LTER query engine http://lterquery.limnology.wisc.edu/

[4] Long-Term Ecological Research program http://lternet.edu

[5] Nottrott, R., M. B. Jones, and M. Schildhauer. 1999. "Using XML-structured metadata to automate quality assurance processing for ecological data." Proceedings of the Third IEEE Computer Society Metadata Conference. Bethesda, MD. April 6-7, 1999.

[6] Knowledge Network for Biocomplexity http://knb.ecoinformatics.org/software/eml/

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

[8] T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler, "Extensible Markup Language (XML) 1.0 (Second Edition)," W3C Recommendation 6 October 2000 http://www.w3.org/TR/REC-xml

[9] "SOAP Version 1.2 Part 0: Primer" W3C Working Draft 17 December 2001 http://www.w3.org/TR/soap12-part0/

[10] Federal Geographic Data Committee. FGDC-STD-001-1998. Content standard for digital geospatial metadata (revised June 1998). Federal Geographic Data Committee. Washington, D.C.

[11] Federal Geographic Data Committee http://www.fgdc.gov/metadata/contstan.html

Through the Looking Glass: What do we see, What have we learned, What can we share? Information Management at the Shortgrass Steppe Long Term Ecological Research Site

Susan G. Stafford, Nicole E. Kaplan, and Christopher W. Bennett
Department of Forest Sciences, Colorado State University, Fort Collins, Colorado 80523-1470

Abstract

This paper documents the development of a successful information management system at a Long Term Ecological Research (LTER) site that has a rich history of data collection and management. Over sixty years of data from three separate projects are incorporated into the Shortgrass Steppe (SGS) LTER information management system and databases. People with different strengths and expertise ranging from clerical administrator, programmer, to ecologist, have filled the role of Information Manager (IM) at the SGS-LTER. Today the information management needs of the SGS are provided by a team of IMs with various levels of expertise in a wide variety of domains from information technology administration to education and outreach. It is critical for IMs at any long-term research site to understand how information and data were managed in the past and what recent changes have been added to the system, in order to effectively implement a management plan for the future. We are able to evaluate the effectiveness of different approaches to information management and have a commitment to share our successes with the information management community.

Keywords: Information management, data sets, metadata, databases, long term ecological research, grasslands

Introduction

A functional information management system must be well organized, yet nimble enough to support ecological research efforts that change with advances in technology (Stafford 1986a, b). Data collection methods at the SGS-LTER have evolved from paper field forms and notebooks to automated data loggers and downloadable satellite images. As a result of technological advances in scientific equipment, the stream of incoming information to manage and data to archive has grown in volume and complexity. The community of IMs across the LTER Network and other organizations is able to share what they have learned, as well as continue to develop tools to better manage and synthesize more data and information. As a site, we are able to look back to see how research projects, data collection and management have changed over sixty years.

History of research programs and issues

Since 1939, research has been conducted on the Central Plains Experimental Range (CPER), a 15,500 acre site located in the northeast corner of the Pawnee National Grasslands, 13 kilometers northeast of Nunn, Colorado. The CPER research site was established in response to the impact of drought and overgrazing across the Great Plains during the era of the Dust Bowl (Figure 1). Researchers from the United States Forest Service studied ways to improve management practices and promote rangeland sustainability across the fragile landscape. Early research projects focused on understanding native plants, recovery on abandoned plowed fields, and techniques for measuring plants' responses to grazing by cattle. Today, cattle from the local grazing association still graze the research site and grazing is considered a "treatment" for some long-term experiments.

In 1968, the CPER began working with the Colorado State University's Natural Resources Ecology Lab (NREL) on studies funded by the National Science Foundation including the International Biological Program (IBP) and later the LTER program (Figure 1). During the IBP, research projects were started to gain an understanding of grassland ecosystems. Over 70 scientists collaborated to obtain information from field studies that examined ecosystem interactions and grassland productivity. The IBP pioneered the ecosystem analysis approach, which brought together scientists from diverse disciplines and organizational structures (Van Dyne memo 1970). Some of the goals of IBP focused on understanding the productivity and the economic importance of grassland ecosystems. Grasslands such as the CPER and Pawnee National Grasslands, were studied within the context of human use and resource management.

The LTER program has less of a utilitarian approach to understanding and studying the shortgrass steppe compared to the IBP (personal communication, W.K. Lauenroth). LTER scientists focus on five core areas of research and monitoring that are mandated by the National Science Foundation for the LTER program. Callahan (1984) organized these foci as patterns and controls on primary production, organic matter accumulation, inorganic inputs and transport, disturbances, and populations. Since 1982, research projects at the SGS-LTER have been designed to answer questions about how factors that regulate structure and function and coupling of biotic and abiotic components in the shortgrass steppe ecosystem vary over space and time, and determine vulnerability to changes (Figure 1). Field data are collected from long-term plots on the SGS-LTER, across regional transects spanning the Great Plains, and in satellite images.

As a network, LTER researchers seek and conduct long-term continuous measurements and analyses of ecological patterns and processes at different spatial and temporal scales. They also emphasize the integration and synthesis of results within and among specific sites and the generalization of results over broader spatial scales and for a broader audience. Together IMs and researchers in the Network effectively formulate questions, coordinate data collection, manage and access electronic archival information, and exchange complex data sets.

SGS-LTER IMs maintain data sets from a variety of

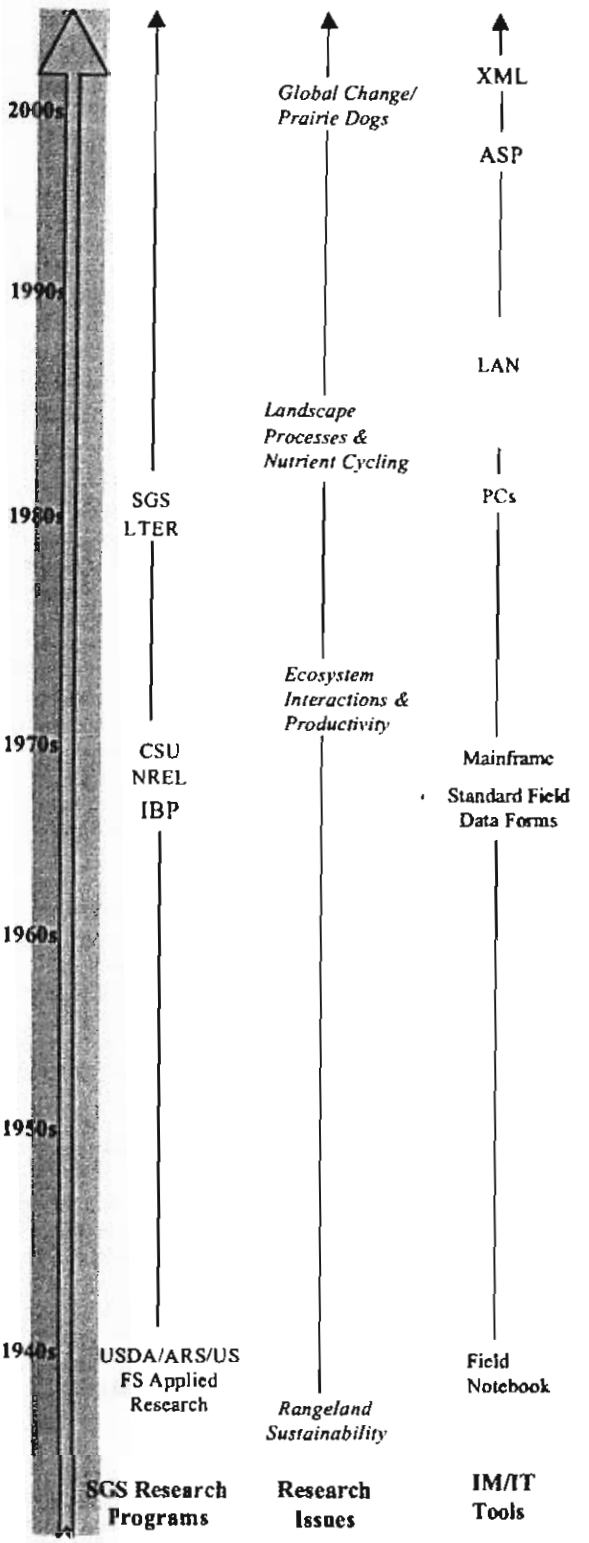


Figure 1. Development of research programs, research issues, and information management and technology tools from 1939 to present.

studies dating back to 1939 (Figure 2). Over one hundred short-term and long-term experiments have been or are being performed. Data sets collected earlier in time helped CPER researchers gain an understanding of what drives production in the SGS ecosystem. Additional data sets were collected since the 1960s during the IBP to gain a more complete understanding of the ecosystem. The LTER now manages many legacy data sets, and data sets generated by cross-site efforts between multiple LTER as well as other community sites. Research projects such as net primary production or meteorological monitoring are conducted at all LTER sites simultaneously and produce data and metadata that are synthesized and published for the entire LTER Network and greater scientific community (Henshaw et al. 1998). SGS-LTER IMs give high priority to working with researchers and IMs from other sites to manage information in a way that will organize and maintain clear and accurate data and metadata for use in synthesis projects, publications, and by the broader ecological community in the future.

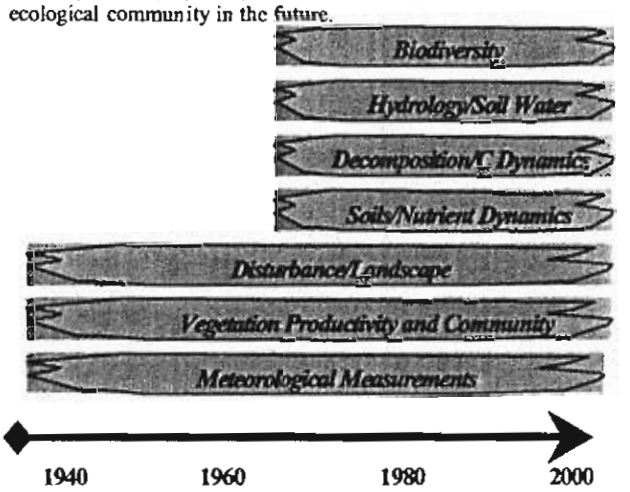


Figure 2. Data sets have been collected from research conducted on the CPER, and during the IBP and LTER programs. Diverse categories of data sets continue to be collected.

History of Information Management and Technology From CPER to LTER

The SGS-LTER information management system serves a community of educators, students, researchers, land managers, and public policy makers and manages data and other information from over sixty years of research. Data from early studies conducted during the initiation of the CPER were collected in field notebooks and on paper. Much of the data still remain in original format and research methods, results, and other information have been published by the United States Department of Agriculture and Agricultural Research Service (ARS) in technical reports.

An infrastructure for the IBP was available through facilities at the new Natural Resource Ecology Laboratory (NREL) at Colorado State University (CSU) and at the time,

state-of-the-art IBP buildings at the CPER site. The NREL clerical office staff supported researchers' efforts to integrate and synthesize data and information from seven satellite field sites in different grassland regions of North America (Pacific Northwest Bunchgrass, High Mountain Grassland, Mixed Prairie, Desert Grassland, Tallgrass Prairie, Shortgrass Prairie, and Annual Grasslands). Initiated in the planning phase of the IBP project, data management, publication, and chemical and statistical analyses were identified as needed along with support services and administration of the project. Two years into the project, it was reported by Van Dyne (1970) that researchers recognized, "...more and more opportunity to undertake integrative synthetic activities which require utilizing data from different trophic levels, from different sites, or combinations of these." IBP researchers then created a policy to share data and complete cooperative analysis and review (Van Dyne 1971). IBP scientists were expected to share data, keep data accessible to be used again, and disseminate information through publishing a series of technical reports and peer reviewed articles.

However, there were difficulties. In 1970, Van Dyne and the IBP collaborators realized the new challenges of working with a large group of scientists from various disciplines. Project information and data needed to be managed within an organizational framework designed to produce broader knowledge and optimize resource management in various grassland ecosystems. Problems arose while meeting the challenge of managing all of the data that IBP produced. As IBP continued, a significant amount of data and information were generated, synthesized, and then published in various reports and publications and archived on Fortran card decks. This system made it very difficult to locate information. In 1972 a keyword index was generated and one volume of abstracts was produced in an effort to collate and categorize the information in one place. It also was difficult to get senior researchers to endorse data sharing, to maintain field equipment, and to invest in adequate support staff to meet the needs of the project. To solve these problems, program directors sought to keep researchers stimulated and dedicated to the project by providing more services to researchers, including data and sample analyses. Involvement in education and outreach activities at the university and resource management agencies, was possible because of site support, as well as the central storage of information.

At the time, available technology limited storage of IBP data and information to hard-copies of field data collections forms, Fortran card decks used for analysis on a mainframe computer at NREL, an extensive series of technical reports, and a closet full of over two dozen 7-track tapes. In total, more than 80 data formats, 111 theses and dissertations and over 200 publications were generated. The IBP project generated such large amounts of data and information between 1969 and 1978 that the SGS-LTER was able to use background information to plan new research strategies (Chaffee 1984).

In 1980, the National Science Foundation selected six sites to establish a Network of LTER sites. The SGS was in the first cohort of LTER sites. From the onset of the LTER Network (Callahan 1984), data management and accessibility were identified as major components of the LTER mission with the LTER IM mission explicitly stating "our goal is to promote ecological science by fostering the synergy of information systems and scientific research." (Baker et al. 2000). The first major undertaking of SGS-LTER information management was to finish documenting and indexing the hoards of IBP data that went un-catalogued after funding ran out. The IBP 7-track tapes were converted to 9-track tapes and more current data were stored on-line on a 10-megabyte removable disk pack that was indexed by project and study site. Metadata were stored where possible. Data managers also recognized that it was almost impossible for secondary users to navigate this data management system to location information of interest (Chaffee 1984). They looked to the future, when user-friendly, expandable, and visible tools would be developed.

By the late 1980s, the SGS-LTER sought to develop a truly integrated computer environment using a LAN (Local Area Network). Researchers built a regional model and a site-level Geographic Information System (GIS) database with data they collected over different spatial and temporal scales. There was an interest in applying new computer technology to ecological research through simulation modeling, spatial analysis using GIS, and remote sensing (Burke 1992, Kirchner 1989). SGS-LTER acquired the hardware and software to build a computing network to support these efforts and began to think about managing data and information from site, regional, and network-wide levels. Computer programmers developed software for models (e.g. Century) and analysis that could be transferred easily to other types of computers such as the IBM PC and Sun Workstation, rather than a mainframe.

IMs and researchers sought to publish data and information by means of hosting field trips to interest groups; promoting newspaper, television, and radio coverage of their research efforts; and publishing in the CSU magazine. Scientific and technical information was disseminated throughout the ecological community through synthesis activities, publishing in scientific journals, and working with an interdisciplinary group of researchers on simulation models and regional analysis.

The 1990s brought a new decade of technological advances, ecological issues, and synthesis research. SGS-LTER researchers and scientists from other LTER sites were generating and manipulating more complex data in models that predicted environmental factors from soil water movement to the impact of global climate change (2002-2008 SGS-LTER Proposal). The abilities of the SGS-LTER information management system were enhanced by hardware and software capable of mass storage, management of spatial and point data sets, increased computing power and speed, and the digitizer, which created digital information on the fly.

In the earlier days of the SGS-LTER computer network, the dominant work platform consisted of a network of Sun unix workstations. While Personal Computers (PCs) played an increasing role over time for each individuals' work, as of early 1999 the SGS-LTER unix file services network had grown to about 20 separate physical drives that were attached to many different physical computers and networked into a model via NFS

("network file system"--the networking part of the unix workstation operating system). SGS-LTER expanded their storage capacity by buying individual drives as money became available, without any deliberate planning to centralize or organize the storage of information. Often, space for projects would be allocated on the "homestead" model, i.e., the drive with the largest available free space would be used for the next new project. PCs added complexity to the data management system by further decentralizing the location of data. This resulted in a network of 20 to 30 drives used for managing various data sets and any metadata associated with SGS-LTER projects. Another problem, consistent throughout each research program through time, was the lack of adequate project documentation, i.e. metadata. Some metadata were recorded for GIS data sets, but as an operating principle the process was essentially lacking, making it difficult to re-use old data. This has been recognized as contrary to the SGS-LTER mission, and so it has been addressed in the new data storage model.

We are able to meet the current and future needs of SGS-LTER well into the foreseeable future with a new storage technology called RAID (Redundant Array of Inexpensive Disks). The product is a single box that contains 5 or more large-capacity commodity disk drives made to operate as a single physical drive. The new data storage model organizes the total storage system into hierarchical components by identifying key attributes for categorization. In addition, some physical optimization is gained by defining an expected pattern of frequent versus infrequent file modification and access. The system is organized in a manner that allows flexibility for developmental data, while recognizing the need for growth and stability of finished data products.

After data are entered and processed through a single point of entry on our new data storage network and assured for quality, the information is transferred to the SGS-LTER Microsoft Access RDBMS (relational database management system) and may be downloaded from our web site (sgs.cnr.colostate.edu). Corresponding attribute definition tables and metadata are included in the database and also are accessible through our website to download in ASCII or XML format.

History of the people from CPER to LTER

In 1939, researchers at the CPER site did not think about archiving, sharing, or synthesizing their data sets or metadata with other scientists or institutions. The ARS who after 1953 administered the CPER, keeps most of the original data sheets in safe storage. Staff at CPER and SGS-LTER are currently transferring CPER data from paper to electronic format. Administrators at IBP realized a need for a data management policies. They organized published information in indexes and attempted to use whatever limited technology was available to convert information from data sheets to a digital format. Programmers during the first SGS-LTER project (LTER I, 1982-1986) continued efforts to convert information from IBP to a more stable format.

In the beginning of LTER I, the SGS-LTER data management system focused on three issues: (1) reduce software development costs by creating utilities that can serve several types of data; (2) centralize processing and screening of data in a timely manner; and (3) safeguard data documentation to avoid loss or corruption. By LTER II (1986-1990), researchers and IMs recognized the need to archive and communicate research information from project initiation. Support staff became more involved with researchers during data and sample collection, verification, lab processing or data entry, QA/QC, archival, and publication (Brunt 2000). SGS-LTER researchers set goals and policies for the IM personnel. Researchers and IMs formed teams to assure data quality, documentation, and timeliness. LTER II began to use data collected in the field to develop models. Programmers again served as IMs and helped researchers to synthesize data and develop models.

After the first two funding cycles, LTER III (1990-1996), IV (1996-2002), and V (2002-2008) placed new demands on SGS-LTER IMs (Figure 1). Our IM system has evolved since the late 1980s, when we saw major technological advances such as faster and larger capacity PCs and the internet. Through the early 1990s, during LTER III, IM personnel consisted of two programmers with backgrounds in biological sciences. Just prior to LTER IV, IM staff became involved in the field and laboratories, and at all steps from project initiation to publication and archival information (Figure 3).

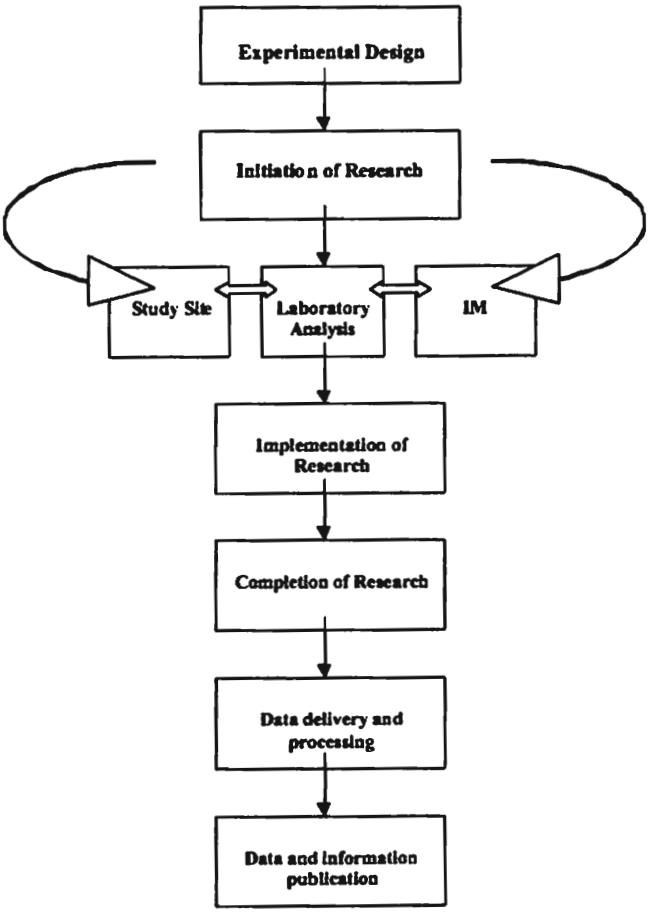


Figure 3. Systematic approach to involvement of staff, information managers and researchers in all steps from project design to data and information publication.

As technological tools developed, our approach changed. Originally, we needed programmers because there were no off-the-shelf products to meet our needs. More recently, the focus of the IM personnel changed to the development and construction of hierarchical relational databases for legacy data from CPER and IBP and new data sets collected throughout the Great Plains region. The development of the World Wide Web created a new venue to publish data and empowered every LTER site to reach a broader audience with ecological information. IMs envisioned a RDBMS capable of providing an organized long-term storage area for data and allowing the public and researchers to access data, metadata, and other information through a dynamic link to the SGS-LTER web site (1996 SGS-LTER proposal). These technological advances made it necessary for IMs to not only understand ecological sciences and current research projects, but also to be proficient in information technology, computer network and database administration, programming, and web site development. By building a team of IMs, we were able to garner expertise in several areas and work together with researchers to set new goals and meet new challenges (Figure 4).

This diagram illustrates the important functions in which the IM Team sees an opportunity to be involved. Components central to the IM system include non-spatial and spatial data and metadata management. Other pieces to the puzzle are part of greater efforts within the University, LTER Network, regional school districts, and the legislative branches of the government. As a team, we are able to cooperate with other ecological research and education and outreach organizations, agencies, and institutions, to manage, archive, and distribute data and information. The SGS-LTER project and LTER Network also benefits from retaining IMs that stay stimulated and valued in their jobs. We are able to attract part-time workers that want the opportunity to focus on their areas of interest and expertise and contribute to the work of the team to the benefit of the overall project.

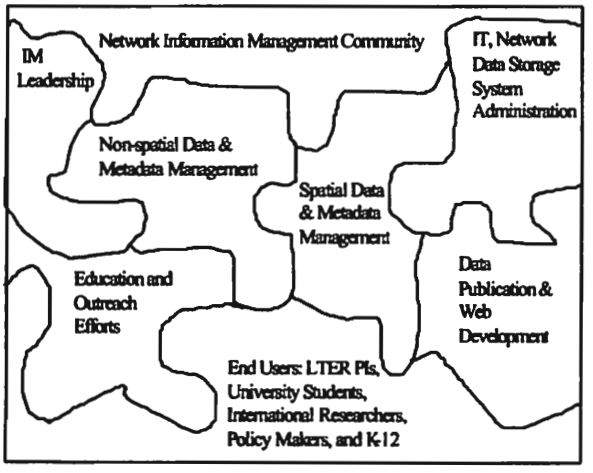


Figure 4. Information Management Team: current and proposed personnel components at SGS-LTER

Through the development of our team approach to IM, we have learned to balance the project's needs for cutting edge IT knowledge and development with efficient, more routine daily IM tasks, while maintaining a strong connection between the IM Team and the researchers. The community of IMs across the Network have to juggle a set of "healthy tensions" on a regular basis. These include balancing a top-down versus bottom-up approach; a site-centric versus network-centric focus; adapting software and other tools that are more appropriate for commercial, production-driven inquiries versus scientific, research-driven inquiries; needing standardization, but working in a network composed of a diversity of approaches; and overcoming the old-fashioned penchant for holding data in a centralized enterprise system versus a more modern enlightened approach of sharing data and metadata in a decentralized enterprise system. We are committed to sharing our experiences and history and becoming a mentor to new LTER sites as they plan their IM activities and organization.

Future Directions for Information Management at SGS-LTER and CPER

The SGS-LTER information management Team has identified areas in which we can become a leader in IM including managing models and their metadata, providing training and mentoring for other sites, and involving IM in our education and outreach programs. Managing metadata for models, controlling model versions, and documenting the range of validity, implicit assumptions, ecological context, and related long-term studies and data sets are challenging goals to meet. The IM Team would like to develop efficient ways to store, document, access and publish this type of ecological modeling information. We will collaborate with other modeling-intensive LTER sites and share results with the LTER Network and greater scientific community.

We also included an outreach and education coordinator on our IM Team. This is a timely opportunity to incorporate IM into Schoolyard LTER efforts and to impart a sense of the importance of IM to the next generation of scientists. We are able to work with existing education centers with ties to SGS-LTER, such as the University of Northern Colorado, Math and Science Teaching Center and the Center for Learning and Teaching in the West. IM will be a part of distance learning programs, professional development programs for teachers, and we will use the web as a tool for data delivery.

Our success as an LTER site depends on building partnerships with other LTER sites, as well as agencies, organizations, and institutions outside the LTER Network. We are currently enlisting the participation of ARS scientists and IMs in SGS-LTER research and information management activities. Researchers and IMs at SGS-LTER, CSU, and ARS are building tools to facilitate the collection and archival of project-level information and metadata in a RDBMS (Figure 5). This tool has proven to be useful and flexible enough to serve particular needs of managers at the SGS-LTER site, as well as the broader scientific community. People across the Network, academicians, agency personnel and policy makers can access this information for administrative, research, teaching, or synthesis research purposes with the use of active server pages (ASP) on our web site (Figure 1). Together with ARS staff, we were able to document project metadata in an accessible database that dates back to 1976. With this information, ARS site managers have the power to keep track of scientific experiments, treatments, land

use changes, and cattle stocking rates across the CPER site and to manage the land on which we conduct our research.

The IM Team also is excited about our active involvement with developing a "content standard" for metadata within the LTER community. This new tool, called Ecological Metadata Language (EML), will simplify data access (http://caplter.asu.edu/data/metadata/workshop012002.htm). The hierarchical design of our new data storage system and the new RDBMS containing metadata and project level information is similar to the design of the EML schema. We hope to be able to transfer metadata from our system to EML and take advantage of a remote server (e.g. Metacat) to reach a broader ecological community.

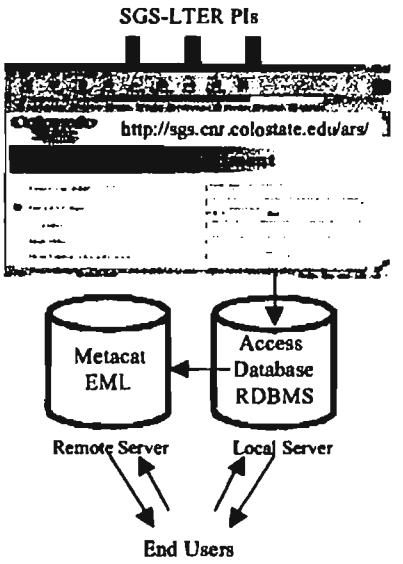


Figure 5. Web tool for researchers to contribute metadata directly to the SGS-LTER RDBMS.

Conclusions

The "Decade of Synthesis and Standardization" of metadata continues to be a pressing issue at both the site and Network level (Stafford, personal communication and Baker et al. 2000). The mission for LTER in 2002 has been reaffirmed. The central, organizing intellectual aim of the LTER program is to understand long-term patterns and processes of ecological

metadata continues to be a pressing issue at both the site and Network level (Stafford, personal communication and Baker et al. 2000). The mission for LTER in 2002 has been reaffirmed. The central, organizing intellectual aim of the LTER program is to understand long-term patterns and processes of ecological systems at multiple spatial scales. To accomplish our mission we must gain a better understanding of the ecosystems which we study, by conducting cross-site, comparative research and synthesizing long-term data sets.

It is essential that LTER sites maintain a leadership role in the dissemination of information and continue to create well-designed, documented databases that are accessible on the World Wide Web. Our legacy of long-term data sets, as well as our understanding of IT implementation in information management is valued as an LTER asset. The Network has the expertise and the information to serve as a gateway for educators, policy makers, and greater scientific community to reach data and information to gain knowledge about a diverse set of ecosystems around the world. The site and Network's commitment to information management means that IM professionals will continue to participate in training and development exercises, and to share our experiences with a broader scientific community through outreach activities, such as this symposium (SCI2002 www.iiis.org/sci2002).

Acknowledgements

This material is based upon work supported by the National Science Foundation under Cooperative Agreement #DEB-9632852.

References

Baker, K.S., B.J. Benson, D.L. Henshaw, D. Blodgett, J.H. Porter, and S.G. Stafford. (2000) Evolution of a multi-site network information system: the LTER information management paradigm. BioScience. 50 (11):963-978.

Burke, I.C. and W.K. Lauenroth. (1992) Enhancement of data management capabilities: a supplement to "BSR-9011659 LTER".

Brunt, J.W. (2000) Data management principles, implementation, and administration. Pp. 25-47. *In* W.K. Michener and J.W. Brunt. Ecological data design, management and processing. Blackwell Science Ltd., Oxford, UK.

Callahan, J.T. (1984) Long-term ecological research. BioScience. 34:363-367.

Chaffee, S. (1984) LTER Data Management reports: CPER Site report.

Kirchner, T.B. (1989) TIME-ZERO: the integrated modeling environment. Ecol. Modelling. 47:33-52.

Henshaw, D.L., M. Stubbs, B.J. Benson, K.S. Baker, D. Blodgett, and J.H. Porter. (1998) Climate database project: A strategy for improving information access across research sites. Pp. 123-127. *In* Michener, W., J.H. Porter, S.G. Stafford. Data and information management in the ecological sciences. A Resource Guide. Albuquerque, N.M.: LTER Network Office, University of New Mexico.

Stafford, S.G, P.B. Alaback, K.L. Waddell, and R.L. Slagle. (1986a) Data management procedures in ecological research. Pp. 93-113. *In* W.K. Michener, editor. Research data management in the ecological sciences. (Belle W. Baruch Library in Marine Science Number 16). University of South Carolina Press.

Stafford, S.G, P.B. Alaback, K.L. Waddell, and R.L. Slagle. (1986a) Data management procedures in ecological research. Pp. 93-113. *In* W.K. Michener, editor. Research data management in the ecological sciences. (Belle W. Baruch Library in Marine Science Number 16). University of South Carolina Press.

Stafford, S.G., M.W. Klopsch, K.L. Waddell, R.L. Slagle, and P. B. Alaback. (1986b) Optimizing the computational environment for educational research. Pp. 73-91. *In* W.K. Michener, editor. Research data management in the ecological sciences. (Belle W. Baruch Library in Marine Science Number 16). University of South Carolina Press.

Van Dyne, G.M. (1970) Progress, problems, and potential in the US IBP Grassland Biome Study. Memo October 14, 1970

Van Dyne, G.M. (1971) Grassland Biome, US International Biological Program. December Newsletter.

A Spatial Data Workbench for Data Mining, Analyses, and Synthesis

John Vande Castle and Deana Pennington
University of New Mexico, Department of Biology
Long Term Ecological Research (LTER) - Network Office
Albuquerque, NM, 87131-1091 USA

and

Tony Fountain
University of California, San Diego
San Diego Supercomputer Center, MC 0505
9500 Gilman Drive
La Jolla CA 92093-0505 USA

and

Cherri Pancake
Northwest Alliance for Computational Science & Engineering
Oregon State University
218 CH2M-Hill Alumni Center
Corvallis OR 97331 USA

ABSTRACT

Information managers at ecological research sites grapple with the complexity of diverse and heterogeneous datasets. The effective management of large geospatial datasets requires extensive hardware, software, and human resources that are often beyond the capabilities of smaller institutions. A major challenge has been the lack of tools capable of integrating very large geospatial datasets with more conventional ecological data. A data "toolbox" was developed to address this challenge. The toolbox, known as the "Spatial Data Workbench," consists of multiple layers of software tools that make it possible to access and integrate "toolbox" was developed to address this challenge. The toolbox, known as the "Spatial Data Workbench," consists of multiple layers of software tools that make it possible to access and integrate multi-temporal and multi-site geospatial data. The Spatial Data Workbench provides access to the large-scale data acquired for individual projects and makes them available in a user-friendly environment. The goal is to reduce duplication of effort and extend access to a wider research audience to facilitate integrative types of research, such as time series and cross-site analyses.

Keywords: Geospatial data, remote sensing, hyper spectral data, informatics.

1. ECOLOGICAL DATA MANAGEMENT

Ecological research sites such as those within the Long Term Ecological Research Network Program (LTER) of the National Science Foundation (NSF) are interested in the management of diverse ecological datasets. Information managers at ecological research sites grapple with the complexity of diverse and heterogeneous datasets. The effective management of large geospatial datasets, such as those generated by remote sensing, requires extensive hardware, software, and human resources that are often beyond the capabilities of smaller institutions. Significant amounts of staff time must be dedicated to daily data collection and requires extensive hardware, software, and human resources that are often beyond the capabilities of smaller institutions. Significant amounts of staff time must be dedicated to daily data collection and monitoring efforts. Important datasets, especially those in the multi-gigabyte size ranges such as hyperspatial or hyperspectral data, often must be stored off-line, so individual files must be moved back on-line when needed for analysis. The integration, analysis and synthesis of those datasets require significant levels of staff time and expertise. The capability and capacity of hardware and software systems are just beginning to permit the manipulation of these valuable datasets as online resources.

The Spatial Data Workbench is a collaborative effort between LTER (http://www.lternet.edu/) and the San Diego Supercomputer Center (SDSC), (http://www.sdsc.edu/) as part of the National Partnership for Advanced Computational Infrastructure (NPACI) Earth System Science initiative (http://www.npaci.edu/). Data are managed with the Storage Resource Broker (SRB) at SDSC. The SRB provides the basis for integrative analysis tools, for the storage and dissemination of geospatial LTER datasets through World Wide Web and other forms of access.

2. THE SPATIAL DATA WORKBENCH

A major challenge has been the lack of tools capable of integrating very large geospatial datasets with more conventional ecological data. Advanced remote sensing data such as AVIRIS (Airborne Visible Infrared Imaging Spectrometer) data (Figure 1) are particularly difficult due to their complex structure and data volume.

Figure 1. An AVIRIS thumbnail image representing the 224 bands of the full hyperspectral dataset.

The collaboration between LTER and SDSC developed a data "toolbox" to address this challenge. The toolbox, known as the "Spatial Data Workbench," consists of multiple layers of software tools that make it possible to access and integrate multi-temporal and multi-site geospatial data. In particular, the Spatial Data Workbench provides integrated support for large-scale data management and analysis using high-performance computing and storage facilities. It grew out of a need to manage and process the large AVIRIS hyperspectral datasets for several sites participating in the LTER program.

A single AVIRIS "acquisition" can encompass more than 50 gigabytes of data (Figure 2), typically distributed across a number of physical tapes. The desired datasets must be ordered from NASA's data archives, and then copied from tape to disk, a process itself requiring significant time and effort, before it can be analyzed. Once the analysis has been performed, the data are usually deleted because of the large data volume, to make room for other data.

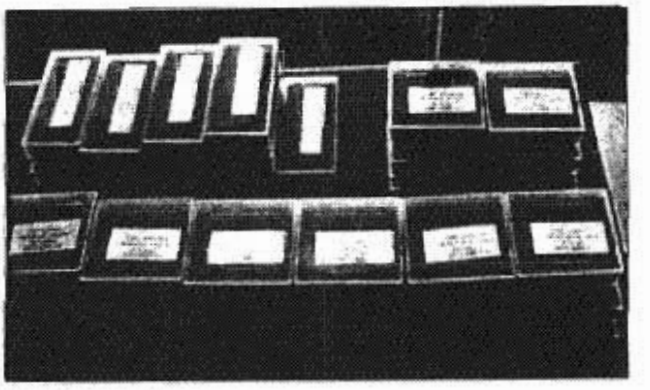


Figure 2. An example of the data volume from a single 2-day data acquisition of AVIRIS data distributed on high capacity tape cartridges.

The capability of integrating hyperspectral data with more conventional remote sensing data such as Landsat Thematic Mapper was a secondary motivation for the project. The concept of the Spatial Data Workbench, then, is to make large-scale geospatial data more generally available, thereby reducing duplication of effort and extending access to a wider research audience. To do this, the Spatial Data Workbench is in part a data archive and access tool to preserve the large-scale data acquired

for individual projects. It makes the data available in a user-friendly environment that facilitates more integrative types of research, such as time series and cross-site analyses. More than an access to datasets, the design of the Spatial Data Workbench permits analysis in a high performance computer environment.

3. DATA ACCESS

The Spatial Data Workbench manages AVIRIS, Landsat Thematic Mapper, Advanced Very High Resolution Radiometer, and other remote sensing data. These data have been acquired by the Network Office of the LTER Program since 1990. Because the kinds of integrated analyses needed by ecologists require an ever-expanding number of tools and systems, our software toolkit consists of 3 tiers (client, server, and analysis) that provide the software "glue" for building plug-and-play systems (Figure 3).

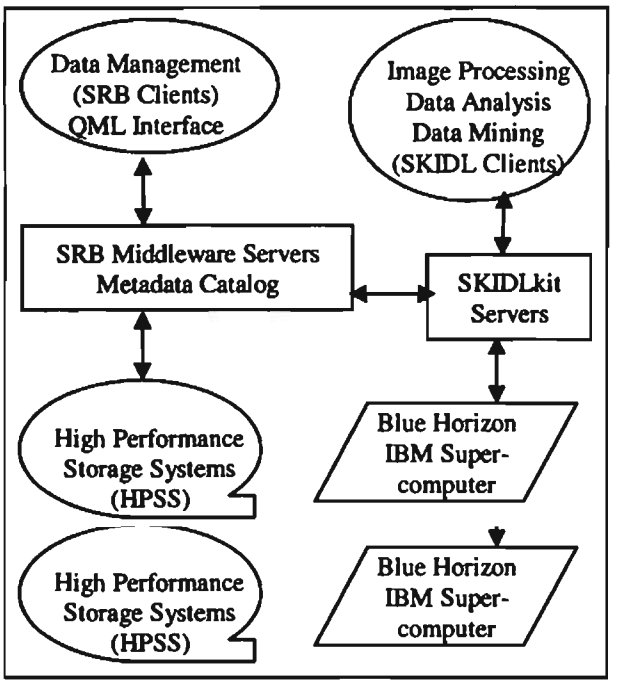


Figure 3. Architecture diagram and processing flow of the LTER Spatial Data Workbench.

Data residing within the Spatial Data Workbench are organized within a Storage Resource Broker (SRB), developed at the San Diego Supercomputer Center (SDSC). The SRB provides the middleware services needed to manage multiple, distributed and heterogeneous datasets as a single logical collection. The SRB provides seamless access to archival resources and file systems along with tools for authenticating users, controlling access, and auditing accesses. The SRB's metadata catalog, which includes both system-level and user-defined metadata, facilities query of individual datasets based on attributes rather than file names or physical locations. Metadata based on standards currently being developed specifically for ecological data (Ecological Metadata Language) by the NSF-funded Knowledge Network for Biocomplexity will be loaded into the SRB's metadata catalog. Advantages of the SRB include location transparency, improved reliability and availability [3]. The overall architecture is designed to be flexible in order to support a variety of configurations and applications.

Both metadata and datasets may be conveniently accessed through any web browser. The Spatial Data Workbench includes a variety of web-based search tools designed to reduce the level of technical expertise needed to locate remotely sensed data. The project web site, http://www.lternet.edu/technology/sdw/, includes documentation of available data cataloged by LTER site, links to data access interfaces, location information such as flight lines and imagery metadata, and low resolution browse images. An initial web-based interface was developed for direct access to the datasets although a more intuitive windows-based graphical user interface (Figure 4), to the Spatial Data Workbench is provided by a browser client to the data managed by the Storage Resource Broker at SDSC [1], [2]. We are also pursuing remote service access for the SKIDLkit tools by a Java multi-tier system and browser client to the data managed by the Storage Resource Broker at SDSC [1], [2]. We are also pursuing remote service access for the SKIDLkit tools by a Java multi-tier system and a second web services approach using XML, SOAP and WSDL. The services will be published so that developer can access them for custom applications, but the specific services will still need to be defined.

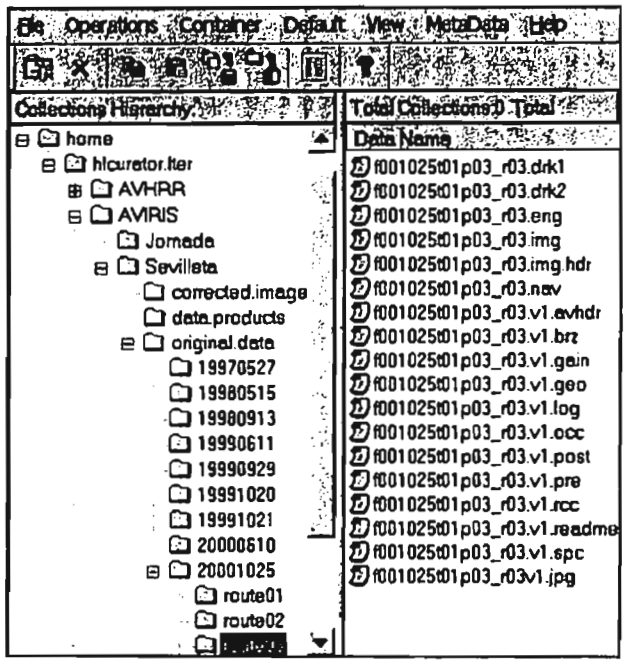


Figure 4. A windows-based browser interface to the Spatial Data Workbench data contents managed by the Storage Resource Broker at the San Diego Supercomputer Center.

We have also developed a specialized web-based interface to the high-volume, high-dimensionality AVIRIS hyperspectral data using Query Markup Language. The interface provides more simplified access to the more complex AVIRIS data sets. The custom interface is being extended to provide access to other imagery types.

4. DATA INTEGRATION AND PROCESSING

Due to the extreme size of these datasets, the Spatial Data Workbench must support server-side data

PROCESSING

Due to the extreme size of these datasets, the Spatial Data Workbench must support server-side data manipulation and analysis. To augment the management capabilities provided by the SRB, we have developed a processing and analysis integration toolkit. SKIDLkit (Figure 3) is a Java-based client-server application with utilities for loading large databases, performing statistical/numerical analyses, and operating efficiently in a Grid computing environment. "Grid-based" in this context means computational grid, not a data matrix. We are tracking the Grid developments, e.g., the Grid Forum actions, and designing a compatible architecture. SKIDLkit provides the software glue for building applications, including interfaces to commercial database systems (e.g., DB2, Oracle), hyperspectral processing libraries (e.g., IDL), data mining packages (e.g., IntelligentMiner), and GIS and data visualization support (e.g., Polexis). Combining high-performance analysis tools with SRB-based collection management services enables scientists to explore and add value to their projects seamlessly and efficiently.

Processing and analysis tools are currently under development for exploring these data collections to discover patterns, create models, and test hypotheses relating to biological/ecological processes. A processing and analytical pipeline is being developed off-line for the AVIRIS hyperspectral data and is currently under test. This pipeline will be completed and integrated into the SKIDLkit interface providing server-side processing for many of the compute-intensive and data-intensive processing/analysis operations. The data will be automatically downloaded into a database, retaining spatial and temporal information. An SQL tool will be provided for database query, along with a selection of pre-written SQL statements for common queries. We are developing scalable implementations of algorithms for data reduction and feature selection in high-dimensional data sets using concepts from Bayesian networks, genetic algorithms and support vector machines. Prototypes are currently under test.

CONCLUSIONS

The Spatial Data Workbench was initially designed as a tool for managing very large volumes of geospatial data sets in an ecological research environment. The integration of the data into a high

The Spatial Data Workbench was initially designed as a tool for managing very large volumes of geospatial data sets in an ecological research environment. The integration of the data into a high performance collection management system such as the Storage Resource Broker at the San Diego Super Computer Center provides the capability for distributed access and a pipeline to advanced analytic tools. The goal is to make large geospatial datasets more available in a user-friendly environment and to facilitate collaborative research.

This project uses tools including the Storage Resource Broker developed at the San Diego Supercomputer Center, University of California-San Diego. The supporting structure of the Spatial Data

Workbench, and software support to the Storage Resource Broker were made possible by the work of Arcot K. Rajasekar, of SDSC. This work is supported in part by a grant from the National Partnership for Advanced Computational Infrastructure (NPACI) Program: Earth Systems Sciences Thrust Award# 10152753 and by a grant from the National Science Foundation, #DEB-9634135 to the LTER Network Office.

REFERENCES

[1] NPACI, The SDSC Storage Resource Broker, 2002 - SRB web document at http://www.npaci.edu/DICE/SRB/

[2] Collection-based Persistent Archives, A. Rajasekar, R. Marciano, and R. Moore, *16th IEEE Symposium on Mass Storage Systems* , March 1999.

[3] Rajasekar, A.K. and M. Wan, 2002. SRB and SRBRack-Components of a Virtual Data Grid Architecture *Advanced Simulation Technologies Conference* (ASTC02), San Diego, April 15-17, 2002.

The 6th World Multiconference on Systemics, Cybernetics and Informatics

July 14-18, 2002
Orlando, Florida, USA

PROCEEDINGS

Volume XIX

Papers Contents and Author Index

Organized by IIIS
International Institute of Informatics and Systemics

Member of
International Federation of Systems Research IFSR

EDITED BY
Nagib Callaos

FOREWORD

Systemics, Cybernetics and Informatics (SCI) are being increasingly related to each other in almost every scientific discipline and human activity. Their common transdisciplinarity characterizes and communicates the, generating strong relations among them and with other disciplines. They work together to create a whole new way of thinking and practice. This phenomenon persuaded the Organizing Committee to structure SCI 2002 as a multiconference where participants may focus on an area, or on a discipline, while allowing the participants the possibility of attending conferences from other areas or disciplines. This systemic approach stimulates cross-fertilization among different disciplines, inspiring scholars, generating analogies, and provoking innovations; which, after all, is one of the very basic principles of the system's movement and a fundamental aim in cybernetics.

SCI2002 was organized and sponsored, jointly with the 8th International Conference on Information Systems Analysis and Synthesis (ISAS 2002), by the International Institute of Informatics and Systemics (IIIS), a member of the International Federation of Systems Research (IFSR). IIIS is an organization dedicated to contribute to the development of the Systems Approach, Cybernetics, and Informatics potential, using both knowledge and experience, thinking and action, for the:

a) identification of synergetic relationships among the three aforementioned areas and between them and society;
b) promotion of contacts among the different academic areas, through the transdisciplinarity of the systems approach;
c) identification and implementation of communication channels among the different professions;
d) supply of communications links between the university and professional worlds, as well as between them and the business world, both public and private, political and cultural;
e) stimulus for the creation of integrative arrangements at different levels of society, as well as at the family and personal levels;
f) fostering of transdisciplinary research, both on theoretical issues and application to concrete problems.

These IIIS objectives have directed the SCI/ISAS organization efforts, since 1995.

The success achieved in ISAS'95 in Baden-Baden (Germany), symbolized by the award granted by the International Institute for Advance Studies in Systems Research and Cybernetics (Canada), as the best and largest symposium at the 5th International Conference on Systems Research, Informatics and Cybernetics, encouraged its sponsors and session chairs to organize ISAS'96 in Orlando and prepare a more general Conference on Systemics Cybernetics and Informatics (SCI'97) in Caracas (Venezuela), SCI'98/ISAS'98,

SCI'99/ISAS'99, SCI 2000/ISAS 2000 and SCI 2001/ISAS 2001 in Orlando. The widely acknowledged success of these past conferences, by means of spontaneous verbal feedback and written comprehensive evaluation from 5200 authors of high quality papers, from 72 countries, encouraged the Program and Organizing Committee to make a definitive commitment to organize SCI 2002 jointly with ISAS 2002 in Orlando, in July 14-18, 2002. Many Program and Organizing Committee members from past international and world conferences joined us for SCI 2002 and ISAS 2002, including most of those who organized the World Conference on Systems sponsored by UNESCO and the United Nations' World Federation on Engineering Organizations (WFEO).

On behalf of the Organizing Committee, I extend our heartfelt thanks to the approximately 205 members of the Program Committee, from 48 countries, the 166 invited sessions organizers and to the 3820 additional reviewers for their reviewing efforts which made possible the quality achieved in SCI 2002/ISAS 2002. The 1800 papers published in these proceedings have been approved in the blind reviewing process from about 2850 submitted papers (including extended abstracts and condensed first drafts).

We also extend our gratitude to the invited sessions and focus symposia organizers, as well as to the co-editors of theses proceedings, for the hard work, energy and eagerness they displayed preparing their respective sessions. Our gratefulness is also extended to the sponsoring organizations which have made the conference possible, to Professor William Lesso, for his eternal energy and good nature while serving as the chair of the Program Committee, to Professor Belkis Sanchez who managed brilliantly the organizing process and Professor Jorge Molero, Eng. Maria Sanchez, Eng. Juan Manuel Pineda, Eng. Floramely Arias, Lic. Jeannette Castellano and Ms. Nancy Castellano for their knowledgeable effort in producing the hard copy and CD versions of the proceedings.

Our gratitude to Professors Bela H. Banathy, Stafford Beer and George Klir who dignified SCI2002 / ISAS2002 as its Honorary Presidents. Our expression of gratitude to Professor William Lesso for his eternal energy, continuous support and advice as the Program Committee Chair and as an old friend. We also wish to thank all authors for the quality of their papers. Thanks to this kind of papers quality we have grown from 45 papers (in 1995 Baden-Baden Conference) to about 1800 papers this year.

Professor Nagib C. Callaos
SCI 2002/ISAS 2002 Conference General Chair

CONTENTS

VOLUME VII

Information Systems Development II

Software Engineering

Testing Quality Assurance, Total Quality in ISAS

Design, Development, and Deployment of Enterprise-wide Real Time Information Systems – Invited Session
Organizer: Robin Qiu (USA)

Academic and Industrial Perspectives on Software Engineering – Invited Session
Organizer: Peter Hantos (USA)

Do what I mean: Mapping Intent to Functionality
Organizer: Babak Hodjat (USA)

Optical & Quantum Applications In Information Technologies – Invited Session
Organizer: Renat Letfullin (Russian-Federation)

The Ecoinformatics Challenge: Meeting Ecological Information Needs for the Site, Network, and Community – Invited Session
Organizer: John Porter (USA)

Agile Software Development – Invited Session
Organizer: Anne Fuller (Australia)

Process Technology – Invited Session
Organizer: Pierre Tiako (USA)

Information and Communication Technology in Rural Development – Invited Session
Organizer: Inampudi Ramesh Babu (India)

Process Technology – Invited Session
Organizer: Pierre Tiako (USA)