Template-driven End-User Ecological Database Design

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

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

ABSTRACT

Historically, ecologists have collected and stored data in individualist 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 given few resources to hire 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 [NRC 1995, NRC 1997], many advances in ecology will depend upon effective information management. Database technology will be integral both to the management of ecological information and to the creation of new knowledge. Current database technology appears inadequate to this purpose. A workshop sponsored by the National Science Foundation, the USGS and NASA identified the need for a new "biodiversity and ecosystem informatics (BDEI)". 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 [NSF BDEI 2000, http://bdi.cse.ogi.edu , http://bio.gsfc.nasa.gov].

The Canopy Database Project [Cushing 2002] 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 more easily design, document, archive, and mine field databases. As with our previous work [Maier 1993], 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 [Nottrott 1999], 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 [Michener 1997, Spycher 1996]. Furthermore, even once data sets are archived and validated with adequate metadata, **idiosyncratic data representation makes cross study analysis difficult**, even for close collaborators. 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 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 [Michener 1998, Michener 2001]. 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.

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.

[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.

3. *THE CANOPY DATABANK*.

DataBank is a web-accessible database system designed for canopy researchers to integrate database technology more easily into their research. 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.

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.

Section 3a. DataBank Functional Requirements. In this section, 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 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 archivist. 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 accomplished 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 conducted 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 checking 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".

Section 3b. *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.

Consider an example where a researcher collects data about epiphytes as per-cent-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 per-cent-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 per cent 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.

Section 3c. *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.

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 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.

References

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

[2] Cushing, J. Nadkarni, N., Delcambre, L, Healy, K., Maier, D. and Ordway, E., 2002. 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.

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

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

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

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

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

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

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

[10] Nadkarni, N. & G. Parker. 1994. 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.

[11] Nadkarni, N. & J. Cushing. 1995. Final report: designing the forest canopy researcher's workbench: computer tools for the 21st century. Intl. Canopy Network, Olympia, WA. 50 pp.

[12] National Research Council. 1995. Finding the Forest for the Trees: The Challenge of Combining Diverse Environmental Data – Selected Case Studies. National Academy Press, Washington, D.C. 129 pages.

[13] National Research Council. 1997. Bits of Power: Issues in Global Access to Scientific Data. National Academy Press, Washington, D.C. 235 pages.

[14] National Science Foundation. 2001. 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.

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

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

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