2292

SOLVING PROBLEMS FOR VALIDATION, FEDERATION, AND MIGRATION OF ECOLOGICAL DATABASES*†

G. Spycher
Department of Forest Science, Oregon State University
Corvallis, Oregon, USA

J.B. Cushing
The Evergreen State College
Olympia, Washington, USA

D.L. Henshaw
Pacific Northwest Research Station
Corvallis, Oregon, USA

S.G. Stafford
Department of Forest Science, Oregon State University
Corvallis, Oregon, USA

N. Nadkarni
The Evergreen State College
Olympia, Washington, USA

ABSTRACT

The H.J. Andrews Long-Term Ecological Research site has supported production and access of ecological data for over 20 years. We have developed a flexible mechanism to generate validation code based on a standard set of metadata. This approach reduces the time required for data production and permits the maintenance of multiple databases. New challenges are migration of existing databases to include new data formats and federation of existing databases for general access. The more constrained environment of the Wind River Canopy Research Facility could be used to examine new distribution technologies and incorporate spatial data into the existing structures. Shared research sites offer incentives to individual researchers to use prescribed protocols and tools, and provide a test bed for solutions to migration and federation problems.

* Presented at Eco-Informa '96, Lake Buena Vista, Florida, 4-7 November 1996.

† Research support provided by National Science Foundation grant DEB-9011663.

The H.J. Andrews Long-Term Ecological Research (LTER) site, together with research groups of the USFS Forestry Sciences Laboratory in Corvallis, has supported an organized effort to maintain and store ecological data in a usable form for over twenty years (Stafford et al., 1986). The Forest Science Data Bank (FSDB) at OSU (Stafford et al., 1984,1988) stores over 200 long-term and opportunistic databases from diverse scientific disciplines for access by database owners and secondary users. The Wind River Canopy Crane Research Facility (WRCCRF) now plans to use FSDB services including its database validation system, and in return offers an opportunity for testing migration paths and federation of FSDB databases.

This paper first briefly describes our understanding of problems of validating and storing ecological data based on the experience of the FSDB, and a mechanism developed there to automatically perform validation checks based on standard metadata and specific database rules. We note both some data validation problems that remain and intrinsic limitations with the database technology we have chosen. We then propose some ways to overcome these problems and suggest how database research involving ecological data for the WRCCRF at The Evergreen State College (TESC) and Oregon Graduate Institute (OGI) might address these problems.

I. CONSTRAINTS IN MANAGING ECOLOGICAL DATA.

Environmental research sites intrinsically impose certain constraints on database management. Field and laboratory data are typically recorded on field sheets, field recorders or computers, or other monitoring instruments, and these collections are file-oriented rather than record-oriented. RDBMSs (relational database management systems) are a tool of choice in this context because database technology facilitates reorganizing data for new uses, and relational databases themselves are widely available, cost effective and provide transfer mechanisms between different products. However, we have experienced limitations with current RDBMS technology: lack of database tools for creating, maintaining and integrating different ecological databases; record orientation; and overhead for the individual researcher First, it is rarely, if at all, possible to achieve comprehensive quality control at the first point of entry using the database's built-in integrity features, and we have often had to circumvent the RDBMS record orientation for productive use. We collect some data on computers with validating data entry programs in the field, but rely on the office-based system described below for full-featured data validation. Second, RDBMSs do not provide mechanisms to organize and manage common data dictionaries and shared metadata. Given the rapid creation rate of new ecological databases, we need better ways to facilitate data production through the use of generic tools and to integrate data from several databases. While the many autonomous datasets are loosely coupled though central catalogs, common data dictionaries and shared metadata would allow the building of these integrated datasets and generic tools.

The user culture is the third important constraint with respect to database design knowledge and database software. Users prefer decentralized data access and enjoy control of their own data, and typically will import ascii data from the FSDB into various productivity tools. Also, the ecological community is very supportive of cooperative, multi-site efforts where data sharing is common. For these reasons, flat files rather than relational database tables are the preferred current storage format.

II. DATA VALIDATION AT THE ANDREWS LTER AND THE FSDB METADATA SYSTEM.

Early efforts at the Andrews LTER were directed toward a standard organization of ecological data and accompanying metadata using a mainframe tape library with paper documentation (Stafford et al., 1986). As the LTER data managers examined the documentation issue (Kellog, 1982), the suggested sets of required and desirable metadata were implemented in various forms by theAndrews and all other LTER sites. With organized structures in place, database validation became a key issue to ensure long-term database availability and preservation. Comprehensive data validation was somewhat deficient during the early years due to the high cost and lack of suitable software. The validation problem received more focused attention once both data and metadata became easily accessible on local file servers.

The FSDB metadata system includes database catalogs, table definition files, domain tables, and tables containing database-specific rules (Stafford et al., 1988). Standardized metadata structures are identical for every database. Rules, originally implemented as stand-alone programs, were recently incorporated into the system to consolidate data editing. Rules, are typically specific to individual databases and often have been 'discovered' with the help of database owners. Generic rules are common in time-series contexts, but most rules are only shared occasionally. All rules must be individually coded and are stored in annotated tables. For example, rules in the Andrews LTER Reference Stand monitoring study check that revisited trees have non-decreasing, incremental growth over time, remain the same species, and do not re-emerge after death.

The quality control system consists of a set of simple procedures providing generic data validation. A control program reads the relevant metadata for a data table and generates appropriate validation code relying heavily on embedded SQL (structured query language) statements. The control program executes the generated code and records the results on an error report. Validation includes checks for nulls (entity), domains (numeric ranges and codes), and database rules. All validation items can be turned off individually to avoid redundant checking. Referential validation is still performed through specific rules because our current FSDB metadata system does not support specifying relations. Planned metadata extensions will implement referential validation generically.

While the primary motivation for documenting databases was data preservation (Kellogg, 1982), the FSDB serves as more than a mere repository for the metadata. Besides supporting data production activities, the metadata are used to: 1) guide users in understanding database content, 2) support global queries of the data catalogs, 3) generate data set documentation reports, and 4) enable other generic access functions such as webpage creation and automatic import/export of flat files to RDBMs files.

III. FUNCTIONAL LIMITATIONS OF THE FSDB

Using the FSDB metadata system with RDBMs technology has significantly reduced the amount of time needed to validate ecological data. However, to gain the full benefit from this approach, we want to store both the data and the metadata in a database, provide automated query and retrieval programs, and encourage users to use common data types and validation programs prior to their own data analysis. These will require addressing the inherent problems in both the technology and the research culture described above.

The current FSDB approach lacks flexibility with respect to semantics and data sharing, extensibility, federation, and migration. Even if the science and technology were stable, the semantics of new data may differ from previous data. Thus, new metadata descriptions must be written for incoming data and compared to previous metadata before the data can be placed into existing tables in a relational database and thus be deemed "sharable".

Extensibility deals with adding new data structures to support new scientific questions. In cases where the science evolves, scientists need new data structures to describe their data. In these cases, we must write new tables (rather than reusing previous tables and associated metadata descriptions). We must assume that the RDBMS in use will be capable of handling these new data structures, or that we will simplify the data to fit the table structure.Thus, validation becomes a problem when extending the FSDB to support new data. There is, furthermore, no guarantee that new data tables will include foreign keys that match semantically or structurally the keys in the existing RDBMS, and we have lost the ability to "federate" the data, i.e., to integrate new data into queries across existing data is lost.

Problems introduced by new technologies, what we call "migration", are analogous to those above. Our problems are, however, compounded in that we may have no way to efficiently prepare data for distribution. Particular migration issues we face include incorporating spatial data into existing structures and validation schemes, and making databases available through new distribution technologies such as theWorld Wide Web. For example, we would like to have a mechanism to automatically move data from a central database to the web rather than generating static web versions several times a year. In effect, migration is one kind of extension, but solutions may require more significant research and development.

We see that once we have a method for data validation, new challenges arise. These include extending existing data structures to accommodate new research questions, federating existing databases for general access, and migrating existing databases to include new technologies.

IV. SHARED ECOLOGICAL DATA AT THE WRCCRF

We believe that problems of extensibility, federation, and migration of ecological data could be effectively explored in a more constrained and controlled environment than the entire Andrews LTER. Schemes to increase the capabilities for extensibility, federation, and migration inevitably require that scientists begin working with data archive personnel earlier in their research cycle and use prescribed protocols for data description. We believe that shared research sites could offer incentives to individual researchers to use prescribed protocols by providing tools that enhance researcher productivity. If that is true, then the shared research site could provide a test bed for solutions for extensibility, federation, and migration problems.

Computer and canopy scientists at The Evergreen State College (TESC), Oregon Graduate Institute (OGI), and the University of Washington are developing database models, data structures, and system architectures for improving extensibility federation, and migration through their work with the WRCCRF. The vision for this work evolved from NSF-sponsored workshops with canopy scientists (Nadkarni and Cushing, 1995), and draws on research integrating scientific applications and data (Maier et al., 1994), and metadata and database schema for scientific research (Diederich and Milton, 1991; Bretherton and Singley, 1994) An active collaboration with theAndrews LTER staff will give these computer scientists better understanding of ecological data structures and effective methods of validation. The FSDB validation procedures will be used as models of validation for the project.

This research project, whose goal is to dismantle barriers to data sharing, will link data sets of selected WRCCRF researchers. Restricting the scope to a specific site assures a higher return on investments in database resources, since the site prescribes a specific geographical area and existing site support systems can distribute tools and data. Shared sites typically provide data about the site that can serve as a basis for data sharing and researchers readily recognize the value of leveraging previous observations and results.

The database research approach addresses data-sharing barriers by: (1) providing site-specific data and common schemata (skeletons or models for database development) to support individual research and interfaces to analysis tools, (2) showing how analysis tools populated with site-specific data can enhance productivity and promote data sharing, and (3) conducting research to develop approaches for cross-scale linking, i.e., site-specific data that link individual datasets at different spatial and temporal scales. Domain scientists and computer scientists will jointly plan, design, implement and test new tools and data that are compatible with research protocols and tools already in use.

In the short term (October 1996 to December 1997), the research will address existing site-specific data as a foundation for a specific (hydrology) database. The canopy researcher will use the metadata and database (partially populated by site-specific data) to generate forms for data collection and to enter and validate field data. The new database will extend the existing database to support new scientific questions. Once a researcher has used the new database to enter and validate field data, we shall integrate the new data with site-specific data. From the resulting federated database we will generate files that can be used as input to some commonly used analysis programs demonstrating how one might migrate the federated database to new technologies and increase researcher productivity.

We expect four outcomes of the Wind River database research: 1) involving the ecology researcher in database design, (2) increasing researcher efficiencies in data management, (3) linking data at different spatial and temporal scales, and (4) linking structure and function data so new scientific questions can be addressed.

REFERENCES

Bretherton, F.P. and P.T. Singley. 1994. Metadata: a user's view. 7th International Working Conference on Statistical and Scientific Database Management. J.C. French (ed), IEEE Press.

Diederich, J. and J. Milton. 1991. Creating domain-specific metadata for scientific and knowledge bases. IEEE Trans. on Knowledge and Data Engineering, 3:4 (421-434).

W.K. Kellogg Biological Station. 1982. Data management at biological field stations. Report of a workshop; 1982 May 17-20. 46 p.

Maier, D., J.B. Cushing, D. Hanson, and M. Rao. 1994. Object data models for shared molecular structures. First International Symposium on Computerized Chemical Data Standards: Databases, Data Interchange and Information Systems. R. Lysakowski (ed), STP 1214, ASTM.

Nadkarni, N. and J. Cushing. 1995. Designing the forest canopy researcher's workbench: computer tools for the 21st century. Final Report. International Canopy Network. 94p.

Stafford, S.G., P.B. Alaback, G. Koerper, and M.W. Klopsch. 1984. Creation of a Forest Science Data Bank. Journal of Forestry 28(7): 423-433.

Stafford, S.G., P.B. Alaback, K.L. Waddell, and R.L. Slagle. 1986. Data management procedures in ecological research. In: Michener W.K., ed. Research data management in the ecological sciences. The Belle W. Baruch Library in Marine Science Number 16. Columbia, SC: University of South Carolina Press: 93-114.

Stafford, S.G., G. Spycher, and M.W. Klopsch. 1988. Evolution of the Forest Science Data Bank. Journal of Forestry 86: 50-51.