9-30-1994

1994 IEEE Computer Society Press Reprint

James W. Brunt
University of New Mexico

IEEE COMPUTER SOCIETY PRESS REPRINT

RESEARCH DATA MANAGEMENT IN ECOLOGY: A PRACTICAL APPROACH FOR LONG-TERM PROJECTS

James W. Brunt

Reprinted from PROCEEDINGS OF THE SEVENTH INTERNATIONAL WORKING CONFERENCE ON SCIENTIFIC AND STATISTICAL DATABASE MANAGEMENT
Charlottesville, Virginia, September 28-30, 1994

IEEE Computer Society
10662 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1264

Washington, DC • Los Alamitos • Brussels • Tokyo

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

IEEE COMPUTER SOCIETY

Research Data Management in Ecology: A Practical Approach for Long-term Projects

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

Abstract

Effective management of ecological research data can insure the security and accessibility of data that cannot be collected again under the same conditions, and plays a key role in every aspect of the research project from experimental design to publication. Commercial relational database management software, developed primarily for business applications, does not provide an adequate solution for long-term scientific data management. An archive file format provides the standard around which a data file management system is implemented. The system works within the parameters of existing components of the operating system software. Data filters and data engines are used to communicate data to and from applications. This paper documents this working approach to research data management in use on the Sevilleta LTER project.

1 Introduction

Research data management in ecology, as in other disciplines, involves the gathering, keeping, and manipulation of research data, including the analysis and presentation of results [4] and provision for access to both data and results. The basic products of expensive research projects are the data that are collected. In virtually all research disciplines, a case can be made for retention of costly data that will become more valuable through time. In ecological research, where the field experiment or collections cannot be repeated under exactly the same conditions, the data are even more valuable [10].

With questions of global change increasingly being directed at ecologists, it is reasonable to expect that data will be used at longer temporal scales, scales that exceed the lives of scientists now collecting the data. Long-Term Ecological Research (LTER) is a program sponsored by the National Science Foundation's (NSF) Division of Environmental Biology [3] [6] developed to support research of ecological phenomena that occur over time scales of decades to centuries, periods of time not normally investigated with NSF research support [8]. In long-term studies, retention and documentation of the data are the foundation upon which the success of the overall project succeeds or fails. Twenty years after the fact is too late to discover that the data are, for any of a myriad of reasons, not available for the task at hand. A strong commitment to data management issues of quality assurance and long-term security is required to make data useful to those who did not collect it. Demonstration of this commitment is becoming a justifiable criterion for review by funding agencies.

Data management in large-scale, long-term projects like LTER provide value to the data by striving to meet the following goals:

Quality - provide the best quality data possible within budget.

Access - provide access to data for the investigators.

Security - provide security for data through archival storage, publication of data, and taking steps to protect the data from natural or man-made disasters, assuring data can be retrieved and understood by investigators now and in the future.

Support - provide computational support for the investigator's analysis of collected data, including but not limited to quality assurance checks.

These goals in support of ecological research are not new, similar goals have been described by Briggs [1], Michener and Haddad [11], and Stafford et al. [14]. Implementations of systems of protocols and computers to achieve these goals, however, is highly variable among the research community at large. This paper concerns itself with the practical application of approaches that work *now* to achieve these goals and elucidates needs for database developers and software engineers; specifically, the need for incorporation of documentation as a part of the data, and the need for an integrated, modular, approach to data management that reflects the nature of ecological data and the way in which research projects function.

2 Data Design

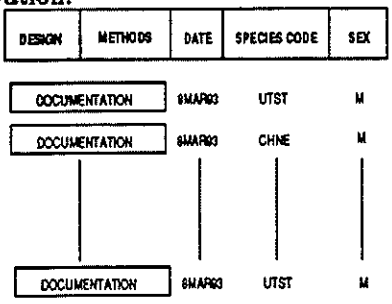
In any research project, the incorporation of documentation standards is crucial [11]. Any researcher who has tried to produce syntheses integrated over space and time using previously collected data, either by themselves or by others, has undoubtedly experienced the frustration of inadequate documentation of data. Regardless of strategies adopted to manage the data, there must be continuous effort by all parties to

0-8186-6610-2/94 $04.00

promote adequate documentation. Data should stand alone and contain sufficient information so that a future investigator who did not participate in collecting the data can use the information for some scientific purpose.

In business applications, all data are related and constitute a single logical database. In large, diverse research projects, there can be many logical databases as opposed to one. Each ecological database is acted on as a unit of data collected under the same monitoring or experimental design by the same sampling methods. The experimental design and sampling methods are part of documentation information for the data. Documentation is a fundamental and logical component of the data. An observation and the documentation of that observation constitute a single logical record. Put another way, the documentation makes up part of the unique key for each observation or record (Figure 1). For example, an observation of a lizard along a transect includes not only the attributes recorded but the methods and design under which it was collected. Figure 1 is just an illustration of the concept; in practice, the information does not need to be re-stated for every observation. Without the association of methods and design, however, the observation is worthless.

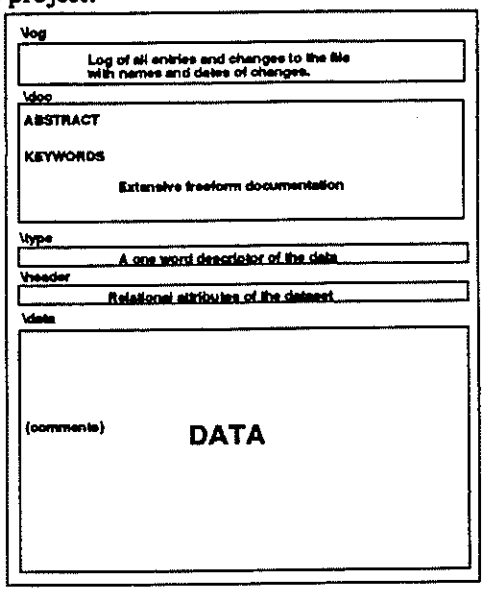
Figure 1: Conceptual relationship of the methods and design to the observation of ecological phenomenon. Because observations in ecology are collected by design and methods, the design and methods, represented here as objects, become part of the primary key to the observation.



In that documentation is a component of each observation, documentation should be kept as part of the database. It is then necessary to use a method of data storage that has adequate flexibility in handling documentation. Object-oriented database management system (DBMS) software can include free-form documentation as a key in the tuple, but these packages are not flexible enough to meet the other needs of the system, described later.

Because of this need to be flexible, the data design for the research data management system for the Sevilleta LTER is implemented on a datafile format as described in Conley and Brunt [4]. This format provides a computer data structure that contains full documentation and comments (Figure 2). The file, called the "Intersite Archive Datafile" (IAF) structure[4], is a generic ASCII data file that can be used on any hardware and software system, and that can be sent on any electronic network or file transfer system. The IAF structure has been adopted by a number of sites as a defacto standard for ascii data to facilitate intersite communication [2] and provides a mechanism for an orderly, although technologically unsophisticated approach to the design and implementation of a project research data management system.

Figure 2: Diagrammatic structure of the Intersite Archives dataFile (IAF) as adopted by the Sevilleta LTER project.



3 System Design

Ecological research data management at the LTER project level requires integrated computational services that can help meet the goals described above by providing access to relational data base operations, statistical and numeric operations, and technical text support [5]. These operations can be collectively supported as a research data management system. A properly instituted and supported system can increase the productivity of ecological researchers and "enhance the quality of the ecological science" [13]. To achieve this, a DBMS needs to *help* ecological researchers do science.

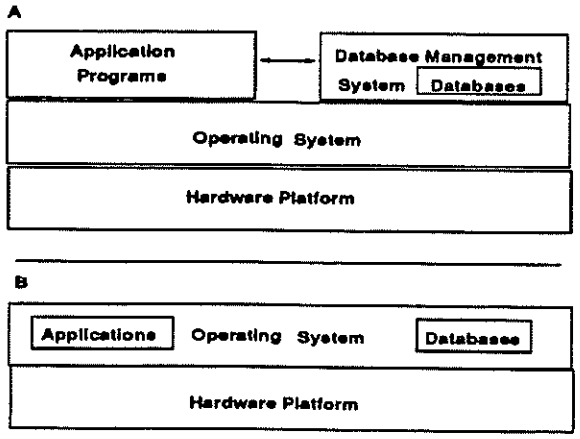
Putting data into a computer "where they belong" is a basic premise of the computerized database scientist. This simple, but empowering phrase is the guiding principle behind database management system (DBMS) software. Often, the new database manager finds himself with raw data in a commercial DBMS software package and can't do any of the needed processing or analyses. For example, much of the data collected on LTER projects are one or two dimensional observations tied to any number of geographic and edaphic keys. These and other scientific data take time to mature because they often require processing and analysis before meaningful database queries can

be performed;[16] ie., observational data are combined and transformed into information such as density of rabbits, percent cover of plants, etc. But the initial observations, or raw data, represent the data that must be protected, because it is the data with the least inherent error. These data might be needed by future scientists to recalculate the density of rabbits using some new density estimation algorithm. These factors and others point to the need for a modular system of research data management functions.

Unfortunately, business has had more influence over the design, development, and implementation of DBMS's than any other source. Most commercial data base management theory and software is directed specifically at business solutions, and seldom applicable to scientific applications. These solutions do not physically or logically meet the needs of the ecological research community. However, in practice, commercial DBMS software is often made to work in scientific applications, or more realistically yet, scientific data is "shoe-horned" into commercial DBMS software. This is an unacceptable solution.

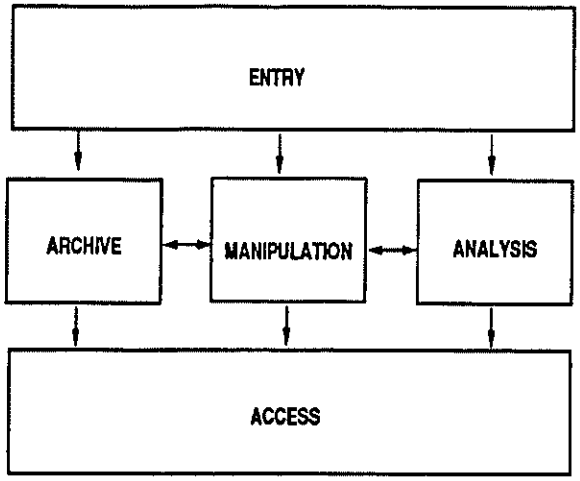
In keeping with the "practical" approach, we chose to develop a modular system around the file handling capabilities that exist in today's operating systems such as UNIX. Using shell and C programming under UNIX for application and manipulation programs eliminates the overhead of maintaining a commercial DBMS [9]. This approach is known as a "Data File Management System" (DFMS). Figure 3 diagrams the difference between the two approaches.

Figure 3: (A) how data are stored inside the DBMS software and exchanges are made between DBMS and application programs. The entire system runs "on top" of the operating system. (B) how a database management strategy can be implemented to take advantage of the operating system's inherent file handling and I/O capabilities.



The capabilities of the DFMS are available to meet all the requirements of a DBMS, as described by Hogan [7], through disciplined use. The contention is that a modular system that relies on a common data storage format is more functional for meeting project goals than DBMS software (Figure 4).

Figure 4: Relationships of the functional modules of Sevilleta research data management system.



The data that are archived in IAF files are observation-based entries as described above. Computerization of data from paper data forms, or data entry, is achieved by the use of SAS AF data entry programs that allow for the verification of field inputs with the documentation for those fields. Range checks, look-up tables, and choice lists are provided via these programs. The data are double entered or visually checked, then summary analyses produced for additional quality assurance checks and to produce the informational database files using /rdb and SAS BASE and STAT software. With the use of filters, the raw data are then translated into the archive file format and submitted for archival. Processed data are converted into queriable database files. As the data are processed, the information content goes up and the data are of greater use to a wider user-base.

Data stored in archive files are located in specific areas of the hierarchical directory structure of the computer system, with analytical and management tools stored in discipline specific areas. Certain areas of the directory structure are used for data entry, manipulation, and construction of the archive files. Once complete, the files are transferred to the archive area. Once the data are in the archive area, they are then accessible to the investigators via electronic network using SQL, mosaic, and gopher. Access to raw data is provided through an interface that allows the user to download the data in a variety of formats compatible with the analytical tools available. Other commercial software in use includes /rdb, and SAS.

The archive file is the reference version of the data. All subsequent work is done to the archive file or copies made from it and then replaced. This helps to control the problem of "offspring" files, files that contain different versions of the data proliferating on the project. The on-line data are copied and saved to tape or optical disk and put in two places off the premises to protect against natural and un-natural disasters.

4 Discussion

This solution has been called uninspiring, yet the fact remains it is a functional system that recognizes the way scientists work; it does not try to control the way they work. What scientists need from software and database engineers is fewer "omnipotent" database packages and more tools to integrate existing software [15]. The service providers must seek to provide easier access and communication of data and involve themselves in close collaboration with software engineers.

Acknowledgements

This work supported by grants from the US National Science Foundation for the Sevilleta LTER project (DEB-88110946). I acknowledge the contributions of colleague Walt Conley and express thanks for his inspiration and guidance in the early development of the system and to anonymous reviewers who contributed their comments. This is contribution No. 53 to the Sevilleta LTER program. UNIX is a registered trademark of Bell Labs; SAS, SAS BASE, SAS STAT, and SAS AF are registered trademarks of the SAS Institute, Inc. "/rdb" is a registered trademark of Robinson, Shaeffer, and Wright, Inc.

References

[1] J. Briggs, "Development and refinement of the Konza Prairie LTER research information management program," In: *Environmental Information Management and Analysis: Ecosystem to Global Scales*, Michener W.K., J.W. Brunt, and S. Stafford Eds. Taylor and Francis, New York, 1994.

[2] J. W. Brunt, and W. Brigham. "Data standards for collaboration in science," In: *Data Management at Biological Field Stations and Coastal Marine Labs: Report of a Workshop*, Gorentz, J. ed. pp 15-17. 1992.

[3] J. T. Callahan, "Long-term ecological research," *BioScience*, Vol 34, pp 363-367, 1984.

[4] W. Conley and J. W. Brunt, "An Institute for theoretical ecology? - part V: practical data management for cross-site analysis and synthesis of ecological information," *Coenoses*, Vol. 6, pp. 173-180, 1991.

[5] W. Conley, B. M. Slator, M.P. Anderson and R. A. Sitze. "Designing and prototyping a scientific problem solving environment: the NMSU science workbench," In: W.K. Michener, ed. *Research Data Management in the Ecological Sciences*, The Belle W. Baruch Library in Marine Science, no. 16, University of South Carolina Press, Columbia, pp. 383-410, 1986.

[6] J. F. Franklin, C.S. Bledsoe, and J.T. Callahan. "Contributions of the Long-Term Ecological Research program," *BioScience*, Vol. 40(7), pp. 509-523, 1990.

[7] R. Hogan, *A Practical Guide to Database Design*, Prentice-Hall, New Jersey, 194 pp, 1990.

[8] J. J. Magnuson, "Long-Term Ecological Research and the *invisible* present," *BioScience*, Vol. 40, pp. 495-501, 1990.

[9] R. Manis, E. Schaffer, and R. Jorgensen, *Unix Relational Database Management: Application Development in the UNIX Environment*, Prentice-Hall, Englewood Cliffs, N.J., 476 pp, 1988.

[10] G. R. Marzolf, and M.I. Dyer, "Future directions for research data management in ecology," In: W.K. Michener, ed. *Research Data Management in the Ecological Sciences*, The Belle W. Baruch Library in Marine Science, no. 16, University of South Carolina Press, Columbia, pp 411-420, 1986.

[11] W. K. Michener, and K. Haddad, "Database administration," In: *Data Management at Biological Field Stations and Coastal Marine Labs: Report of a Workshop*, Gorentz, J. ed., pp. 4-14, 1992.

[12] J. H. Porter, and J.T. Callahan, "Confounding a dilemma: historical approaches to data sharing in ecological research," In: *Environmental Information Management and Analysis: Ecosystem to Global Scales*, Michener W.K., J.W. Brunt, and S. Stafford Eds., Taylor and Francis, New York, 1994.

[13] P. G. Risser, and C. G. Treworgy. "Overview of research data management," In: W.K. Michener, ed. *Research Data Management in the Ecological Sciences*, The Belle W. Baruch Library in Marine Science, no. 16., University of South Carolina Press, Columbia, pp 9-23, 1986.

[14] S. G. Stafford, P.B. Alabach, K.L. Waddell, and R.L. Slagle, "Data management procedures in ecological research," In: W.K. Michener, ed. *Research Data Management in the Ecological Sciences*, The Belle W. Baruch Library in Marine Science, no. 16, University of South Carolina Press, Columbia, pp 93-114, 1986.

[15] S. G. Stafford, J. W. Brunt, and W. K. Michener, "Integration of Scientific Information Management and Environmental Research," In: *Environmental Information Management and Analysis: Ecosystem to Global Scales*, Michener W.K., J.W. Brunt, and S. Stafford Eds., Taylor and Francis, New York, 1994.

[16] D. E. Strebel, B. W. Meeson, and A. K. Nelson, "Scientific Information Systems: A Conceptual Framework," In: Environmental Information Management and Analysis: Ecosystem to Global Scales, Michener W.K., J.W. Brunt, and S. Stafford Eds. Taylor and Francis, New York, 1994.