SCIENTIFIC DATABASES FOR ENVIRONMENTAL RESEARCH

John H. Porter

Department of Environmental Sciences, Clark Hall, University of Virginia,
Charlottesville, VA 22903

Abstract. The questions that scientists can answer are dependent upon the databases available to them. Modern genome research would not be possible without genome databases. Similarly, synthetic and integrative environmental research will be dependent on the quantity and quality of available databases. Examples of scientific databases include large "deep" databases such as Genbank and PDB, "wide" databases such as the National Geophysical Data Center and NASA Distributed Active Archive Centers (DAACs), and project-oriented databases such as those at Long-Term Ecological Research (LTER) sites. There are advantages and disadvantages for using database management systems that balance the capabilities gained against the costs of maintenance. The World Wide Web is a recommended interface for scientific databases. Such databases may be constructed on both UNIX and Windows NT workstations.

INTRODUCTION

There are several advantages to developing and using scientific databases (National Research Council 1997, Pfaltz 1990). First, databases lead to an overall improvement in data quality. Multiple users provide multiple opportunities for detecting and correcting problems in data. A second advantage is cost. Data costs less to save than to collect again. Often, environmental data cannot be collected again at any cost because of the complex of poorly controlled factors, such as weather, that influence population and ecosystem processes. However, the primary reason for developing scientific databases is the new types of scientific inquiry that they make possible. Such inquiries include: (1) long-term studies, which depend on databases to retain project history; (2) syntheses, which combine data for a purpose other than which they were originally collected; and (3) integrated multidisciplinary projects, which depend on databases to facilitate data sharing. Public decisions involving environmental policy and management frequently require data that are regional or national, but most ecological data are collected at finer scales. Databases make it possible to integrate diverse data resources in ways that support the decision-making process.

EXAMPLES OF SCIENTIFIC DATABASES

A useful analogy

A useful analogy in examining scientific databases is to consider individual data sets as "volumes" in a database "library." Libraries may have different sizes and different requirements for cataloging systems. For example, an individual might have a home "library" consisting of a relatively small number of books. The books would not be cataloged or organized, but simply placed on a shelf. An individual book would be located by browsing all the titles on the shelf. For an office library consisting of hundreds of books, a common model is to group books on the shelf by general subject so that only a subset of the library needs to be browsed. However, when the number of books in a library enters the thousands to millions, as for a public library, formal cataloging procedures are required.

This model also applies to scientific databases. If there are relatively few different data sets, a simple listing of the titles of the data sets may be sufficient to allow a researcher to locate data of interest. This is the prevailing model in single-investigator and small project databases. The

databases are typically in the form of esoteric World Wide Web (WWW) pages that do not conform to metadata (information needed to use and interpret data) standards.

Examples of databases

Some databases specialize in a single or few types of data and implement sophisticated searching and analytical capabilities. Examples of this type of database are large databases such as Genbank which serves as a primary archive of genetic sequence data for the human genome project, with over one billion nucleotide bases in approximately 1.6 million sequences (National Center for Biotechnology Information 1997) and PDB, the protein structure database which contains over 6,000 atomic coordinate entries for protein structure (*http://www.pdb.bnl.gov/statistics.html*). These are very large databases with funding in excess of one million dollars per year. In the library analogy, these databases are analogous to large, multi-volume reference works. They are highly "indexed," but focus on a restricted region of the data universe.

There are also various specialized types of databases that operate on a smaller scale. For example, MUSE is specialized software for managing herbarium specimens (Humphries 1997) and BIOTA is software for management of specimen-based biodiversity data (Colwell 1997). Like geographical information system software, these systems are commercially available and are used by a variety of institutions and investigators. In the library analogy, they would be books in a series that share format elements and address the same topic, but have different content. Like the large databases (Genbank, PDB), these databases are "deep" rather than "wide" (Table 1), providing in-depth services for a particular type of data.

Table 1. "Deep" vs. "Wide" databases.

"Deep" Databases	**"Wide" Databases**
• Specialize on one or a few types of data • Large numbers of observations of one (or few) type(s) of data • Provide sophisticated data query and analysis tools • Tools operate primarily on data content	• Contain many different kinds of data • Many different kinds of observations, but relatively few of each type • May provide tools for locating data, but typically do not have tools for analysis • Tools operate primarily on metadata content

"Wide" databases are data repositories that attempt to capture all data related to a specific field of science. For example, the National Geophysical Data Center (NGDC, *http://www.ngdc.noaa.gov/*) is operated by the National Oceanic and Atmospheric Administration (NOAA) and supports over 300 databases containing geophysical data (NGDC 1997). Such "data centers" use standardized forms of metadata (e.g., GILS, FGDC, DIF) for maintaining formal catalogs with controlled vocabularies for subjects and keywords. Similarly, the National Aeronautic and Space Administration (NASA) operates a series of Distributed Active Archive Centers (DAACs; See Olson and McCord, this volume), each of which specializes in supporting a particular area of earth or space science and have a varying number of different types of data sets. In the library analogy, these databases would be comparable to public libraries.

Additional "wide" databases are project-based databases. These are databases that support a particular multidisciplinary research project and may include a wide array of data focused on a particular site or research question. Examples of this type of database are the databases at individual Long-Term Ecological Research (LTER) sites (LTER Network Office 1997). These

databases contain data relating to a wide array of scientific topics (e.g., weather and climate, primary productivity, nutrient movements, organic matter, trophic structure, biodiversity, and disturbance), along with information that supports site management (e.g., researcher directories, bibliographies and proposal texts). Management of the databases requires approximately 15% of the total site funding and they focus strongly on long-term data. Within the LTER network, there are diverse approaches to data management dictated by the locations of researchers (at some LTER sites, most researchers are at a single university; at others, they are at many different universities), and the types of data collected (studies of aquatic systems have different data needs than studies of terrestrial systems). Although the LTER network uses individual metadata standards at individual sites, there are network-wide standards for minimum metadata content. These databases are fairly "wide", but not particularly "deep" in the sense that they provide access to a wide variety of data, but do not provide specialized visualization or analysis tools for most types of data. In the library analogy, these databases would be comparable to a large individual or small departmental library.

Some databases, such as individual WWW pages created by individual researchers may be neither "wide" nor "deep." The level of development of such pages varies widely, as does the quality and quantity of the associated metadata. In the library analogy, the pages from a single researcher would be comparable to a very small personal library with little need for searching and cataloging capabilities. As an aggregate, across all researchers, these databases constitute a valuable resource, but one that is difficult to exploit because they can be hard to locate and metadata may be insufficient or difficult to translate into usable forms. Additionally, WWW pages are notoriously ephemeral, so they are a poor choice for long-term database administration.

A strategy for evolving a database

In making the myriad decisions needed to manage a database, a clear set of priorities is the developer's most valuable friend. Every database has some things that it does well (although no part is ever perfect) and some areas that need improvement. The process of database evolution is cyclical. A part of the database may be implemented using state-of-the-art software, but several years later the state-of-the-art has advanced to a degree that it makes sense to migrate the system to new software. Therefore, database systems should be based on current priorities, but with a clear migration path, or at least opportunities, to migrate toward future systems. When making decisions about the types of software to use in implementing the database and associated interfaces, it is critical to consider an "exit strategy." Software that stores data in proprietary formats and provides no "export" capabilities are to be avoided at all costs!

The need for foresight applies to more than just software. The priorities of users may change. A keyword search capability may be a top user priority, but once it exists a spatial search capability may be perceived as increasingly important. It is not possible to implement a database system *in toto*, so the strategy adopted for development must recognize that, although some capabilities are not currently implemented, the groundwork for those capabilities in future versions must be provided for. Thus, even though an initial system may not support spatial searching, collecting and storing spatial metadata in a structured (i.e., machine-readable) form is highly desirable.

An important form of foresight is seeking scaleable solutions. Scalability means that adding or accessing the 1,000th piece of data should be as easy (or easier) as adding the first. The genome databases faced a crisis when the flow of incoming data started to swamp the system (which depended on some level of manual curation of inputs). The subsequent adoption of completely automated techniques for submission and quality control allows the genome databases to handle the ever increasing flows of data. Every system has some bottlenecks and their identification and elimination before they become critical, is the hallmark of good planning and management.

CHOOSING SOFTWARE

The choice of software for implementation of a database must be based on an understanding of the tasks you want the software to accomplish (e.g., input, query, sorting, analysis). Simplicity is the watchword as the world is full of sophisticated software that is expensive and difficult to operate, but that may provide little real improvement over simpler and less expensive software.

User interface

Although a variety of proprietary user interface options exist, it is hard to argue against using an interface based on World Wide Web (WWW) tools. Most potential users of a database will already have, or have access to, a WWW browser (e.g., Netscape and Microsoft Explorer) so there is no need to distribute specialized software. Most potential database users will already be familiar with a WWW browser, reducing the need for training. WWW tools continue to improve at a rapid pace. Important innovations have been the support of on-line forms and linking WWW servers to database engines so that WWW pages can be dynamically generated. The addition of programming languages (such as JAVA™) which allow secure operation of applications on the client-side greatly increases the types of operations that can be supported over the WWW. WWW tools can be used for input to a database, as well as for output. An advantage of this approach is that input of metadata and data can be made from many different locations, which can circumvent some potential bottlenecks.

Advantages and disadvantages of using a database management system (DBMS)

There are numerous advantages to using a DBMS. The first is that a DBMS has many useful built-in capabilities such as sorting, indexing, and query functions (Maroses and Weiss 1982). Additionally, large relational databases include extensive integrity and redundancy checks and support transaction processing with “rollback” capabilities, allowing one to recreate the database as it existed at a particular time. There has been substantial research into making relational DBMS as efficient as possible and many DBMS can operate either independently, or as part of a distributed network. This aids in scalability because if one computer starts to become overloaded, another can be added without having to substantially restructure the underlying system. Finally, most DBMS’s include interfaces that allow DBMS linkage to user-written programs or other software, such as statistical packages. This is useful because it allows one to change the underlying structure of the data without having to alter programs that use the data.

Despite these advantages, most DBMS’s are designed to meet the needs of business applications and these may be quite different from the needs of scientists (Maroses and Weiss 1982, Pfaltz 1990). For example, most commercial DBMS’s have few graphical or statistical capabilities. DBMS’s are typically designed to create standard reports that may be of little use to researchers. Additionally, DBMS’s are typically designed to deal with large volumes of data of a few specific types. They are less useful when dealing with relatively small volumes of data of many different types. Similarly, they can be relatively inefficient in dealing with sequential data. There are some functions, such as highly optimized updating capabilities, that are not frequently used for scientific data because, barring detection of an error, data are seldom changed once in the database. Additionally, not all analysis tools can be easily interfaced with a DBMS and proprietary data formats used by a DBMS may limit archival quality of data. A final disadvantage of a DBMS is that it requires expertise and resources to administer. In some cases, the resources required may exceed the benefits accrued by using a DBMS.

Even if you decide not to use a DBMS for data, you may want to consider use of a DBMS for metadata (documentation). The structure of metadata is frequently more complex than that of data

and conforms better to the model of business data (relatively few types of data, standard reports are useful). Most data are located based on searching metadata rather than the data itself so the query capabilities of a DBMS are useful. Similarly, metadata are changed more often than data, so that the updating capabilities of a DBMS are more useful for metadata.

CHOOSING A COMPUTER SYSTEM

At this time, there are two reasonable options for computer systems which support full-featured database creation: computers running UNIX and computers running Microsoft Windows NT. UNIX is a mature, full-function operating system. It has strong capabilities for multitasking and multi-user support. As a mature system, it is reliable and robust and there is a large body of WWW tools, many of which are free. On the down side, UNIX is difficult to learn and commercial software for UNIX is typically much more expensive than that for personal computer-based systems.

Microsoft Windows NT is a rapidly evolving operating system that has seen major improvements in operating system design that facilitate network access. Compared to UNIX, software and hardware are relatively inexpensive and most software is more “user friendly” than UNIX. The number of WWW tools for NT is growing rapidly. Limitations of databases on an NT are that they are more difficult to scale up than those on UNIX computers, and as a relatively new operating system, there can be problems with reliability.

The capabilities of these systems are similar enough that choice of a system may depend on the local computational environment. If UNIX computers are already in place and there is sufficient expertise to support them, UNIX may be the best choice. However, if those prerequisites are lacking, an NT system may be the better choice.

CONCLUSIONS

Development of scientific databases is an evolutionary process. Although databases evolve, they do not spontaneously generate! It takes the actions of an individual or group to bring them into being, often in a relatively simple form. It is not necessary that a new database try to incorporate all the features it will eventually encompass. Indeed, to do so is a prescription for disaster because it is extremely difficult to anticipate all the needs of the user community. Even if it starts in a simplified form, once in operation, a successful database generates its own momentum by coupling its user community into the development process.

LITERATURE CITED

Colwell, R.K. 1997. Biota: the biodiversity database manager. *http://viceroy.eeb.uconn.edu/biota*

Humphries, J. 1997. MUSE. *http://www.keil.ukans.edu/muse/*

LTER Network Office. 1997. Long-term Ecological Research. *http://LTERnet.edu*

Maroses, M. and S. Weiss. 1982. Computer and software systems. Pages 23-30 in G. Lauff and J. Gorentz, editors. Data Management at Biological Field Stations. A report to the National Science Foundation.

National Research Council. 1997. Bits of power: issues in global access to scientific data. *http://www.nap.edu/readingroom/books/BitsOfPower/* National Academy Press.

National Center for Biotechnology Information. 1997. GenBank Overview. *http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html*

Pfaltz, J. 1990. Differences between commercial and scientific data. Scientific database management, a report to the National Science Foundation. *gopher://lternet.washington.edu:70/00/newsletters/Reports/Miscell/uva_cs_90/cs_90-22*

Porter, J.H. and J.T. Callahan. 1994. Circumventing a dilemma: historical approaches to data sharing in ecological research. Pages 193-203 in W. K. Michener, S. G. Stafford and J. W. Brunt, editors. Environmental information management and analysis: ecosystem to global scales. Taylor and Francis, Bristol, PA.