Requirements of a Database Management System for Global Change Studies

Y. Zhou
School of Geography
Clark University
Worcester, MA 01610

M.A. Gennert, N.I. Hachem, M.O. Ward
Computer Science Department
Worcester Polytechnic Institute
Worcester, MA 01609

ABSTRACT

Global change research is a form of scientific discovery. This paper identifies database requirements for global change research by examining the process of scientific discovery. We find that current database technology is inadequate for the task and a number of deficiencies are pointed out. Specific requirements are laid out in the areas of Data Management, Methodology Management, Flexible Environments, and User Interfaces.

INTRODUCTION

Global change research, a form of scientific discovery, is experiencing an explosive growth, due mainly to the availability of vast quantities and types of information. With this rapid growth comes a dramatic need to manage not only the data themselves but the algorithms, experiments, and hypotheses being generated by the research community. This paper identifies these needs by examining the process of scientific discovery in the context of global change research, and associate with it the implications in database technology for managing all aspects of the process.

Stages of Scientific Discovery

Many attempts have been made over the years to formalize the process of scientific discovery (Langley 87), though little consensus can be found among the various schools of thought and several researchers, philosophers, and psychologists believe that it cannot be formalized (Popper 61). Our view, developed through our own experiences and examining the findings of other researchers, is that the procedure of scientific discovery can be broken into a number of distinct stages, as described below.

Examine prior work: What was done? What data and operations (experiments) were used/performed? What is the meaning of the results? Do results validate the researcher's hypotheses? Is it repeatable? Is it applicable to other areas?

Create hypotheses: Possible starting points might include extending a previous hypothesis, perhaps to another domain, repairing a defect in a prior hypothesis, or creating a new model.

Design experiments: It is important that the experiments are necessary and sufficient to validate the hypotheses, and the researcher must be careful to check the potential for misinterpretation.

Select or gather data: What data are available? What is its source and structure? Are new data needed? If so, how will they be obtained?

Select or design algorithms: What algorithms are currently available? Are they sufficient, or do they need to be modified? Is it necessary to create totally new methods?

Perform experiments: It is critical to keep accurate records of all intermediate and final results.

Evaluate results: This involves both comparison to results of other researchers as well as performing several experiments to uncover trends and verify the robustness of the hypotheses.

Refine ideas and repeat as necessary: This may involve minor or major modifications and the use of different data or algorithms.

Report results: This must be done in a manner such that others may duplicate and evaluate the results. When results are reported in peer-reviewed publications, some degree of validation is performed.

Independent validation: In most cases, scientific discoveries are not widely believed until external validation has occurred.

Many endeavors in GIS and global change research can be viewed as following the template for discovery. For example, a hypothesis regarding land use/land cover change might involve gathering many forms of data (AVHRR, SPOT, maps), evaluating many classification schemes (principle components, maximum likelihood, linear mixture modeling), and performing experiments over diverse regions at different periods of time. Comparison of regions with similar climatic, socio-economic, or geographic characteristics may reveal a heretofore undiscovered relationship or trend. However, inconsistencies between different classification methods may prompt the development of entirely different techniques based on different forms of data. This in turn can lead to additional hypotheses and discoveries.

Observations

The potential for management appears in many aspects of scientific discovery. The most obvious aspect is data management, stemming from the increasing volume and types of data. Similarly, there is a growing need to manage the algorithms to be applied to the data. As there are standard mathematical and statistical software libraries available to the general scientific community, so too should there be common and consistent algorithms for all components of data analysis. To accomplish this requires the development of methods to manage the development, evolution, verification, and dissemination of algorithms. A third focus of management is in the scientific experiments themselves. The view of some types of discovery as iterative refinement dictates a need to monitor the progression of experiments to best identify the most promising future directions. This also helps in avoiding unnecessary repetition of experiments as well as using aspects of previously performed experiments in designing new ones. Finally, to facilitate both dissemination of results and external confirmation/verification, some form of management is needed. Some branches of science, notably biological sciences, have already identified this need, with standard formats for distributing data and reporting experimental results.

DATABASE MANAGEMENT SYSTEMS: CURRENT TECHNOLOGY AND ITS SHORTCOMINGS

Conventional DBMS

The needs and requirements put forth for database management systems (DBMS) evolved considerably over the last decade. The relational model (Codd 70) was developed in the early 1970s as the data model for business applications. One shortcoming of this model is in its limited expressive power to conceptualize the real world of business applications. The entity-relationship model has been proposed as a high level conceptual modeling environment for such conventional applications (Chen 76). Conceptual modeling tools similar to the ER model cannot efficiently model scientific applications.

New Technologies

The 1980s oversaw a dramatic evolution of DBMS technology with the advent of advanced semantic models for the conceptual design of computer-aided design and engineering databases (Hull 87, Peckham 88). Such systems model the structural aspects of data using basic data structuring constructs such as association, aggregation, and generalization. Some of these modeling tools were also proposed for specific statistical and scientific databases (Su 83). Object-oriented databases appeared as an outgrowth of object-oriented programming and were concurrently proposed with semantic models to overcome the shortcomings of relational DBMSs (Maeir 86). These systems rely on the concepts of abstract data types and encapsulation to model the behavior of database objects (Atkinson 89). Other research efforts sought to augment the relational model. One direction was to add deductive capabilities to DBMSs (Ullman 90), while the POSTGRES research group at Berkeley developed extensible post-relational systems (Stonebreaker 90). Deductive databases add the power of inferencing to the programming language of a DBMS while extensible systems such as POSTGRES attempt to integrate the advantages of object-oriented systems, rule-based system, and the relational model.

Current GISs

Current Geographic Information Systems generally concentrate on a single aspect of data manipulation. For example, the GRASS system is a GIS which provides a rich set of analysis operators for geographical applications. On the other hand, GEO (Oosterom 91) is a geographical query system which concentrates on the data management aspects of GIS. These systems manage the spatial properties of geographical data and let the user worry about the data semantics.

Shortcomings

Conventional DBMSs are, in general, not suitable for dealing with the requirements put forth by the scientific community (French 90) nor the GIS community (Aronoff 89). The diversity and size of data coupled with complicated manipulation and analysis impose quite different needs from the business management environment in which most traditional database techniques are applied. Some of the added requirements of GIS and global change research can be satisfied using the emerging technologies, while other requirements require further research and development efforts. We list some important issues and shortcomings of current DBMSs for these applications.

Object sizes: Spatial objects as well as scientific methods are stored and retrieved as large objects of variable sizes. Conventional DBMSs are not designed to efficiently deal with such objects.

Volume of the data: Scientific data are rarely expunged and are usually archived as their relevance to current needs decreases. For example, the Earth Observation System being planned for the 1990s may generate up to 10^{12} bytes per day. Current DBMSs cannot easily cope with the additional requirements put forth by the nature of scientific investigations.

Provision for spatial and temporal data objects: Global change research demands GIS environments that can manage the spatial and temporal aspects of the data. Time is essential in modeling the constantly changing world. However, conventional database systems represent the real world with only a tenseless snapshot that is inadequate for many applications where facts and data need to be interpreted in the context of time. Research on temporal databases is ongoing (Qiu 92, Soo 91, Stam 88), but the field is far from maturity and no actual systems are available that meet the need to manage spatial and temporal information. Spatial aspects of the data are equally important and current research efforts focus on extending current DBMS technology to include primitive operations for the management of spatial data (Guenther 90).

Artificial separation of analysis and queries: The management of scientific investigations, that is, analysis operators together with the management of data, is essential to scientific databases. Current systems such as GRASS and GEO focus on one aspect. What is needed is an integrated DBMS environment which views analysis operations as queries for which the data must be computed, rather than simply retrieved.

Difficulty expressing semantics: Current semantic and object-oriented systems do not provide the necessary mechanisms to express the full semantics of fuzzy definitions of scientific concepts (Morehouse 91). For example, *desertic region* is a concept with no precise definition. One user may relate it to the amount of rainfall, while another relates it to the average amount of vegetation over a time interval. Another concept with multiple definitions is *watershed*, which may be expressed in terms of steepest slope lines, immersion algorithms, or ridge lines (Vincent 91). Not only may data objects be fuzzily defined, so may processes. The software in (Prashker 91) includes 20 different algorithms for line simplification and smoothing.

Extensibility: In traditional DBMSs, the database schema is fixed when the system is initialized, and may be difficult to modify. However, it is not possible to anticipate the eventual schema in scientific investigations—new data types and relations among them will need to be continuously added. Examples of the diverse data types encountered can be found in (NERC 88, ESSC 88).

Lack of data integrity checks: Spatial data records are often highly interrelated, requiring a more sophisticated security system than the record locking approach used by general purpose DBMS.

Cooperative and shared environments: It is becoming increasingly difficult to store the needed data online at a single site. Thus, distributed databases will become commonplace (Özsu 91). There is a need to allow not only sharing of data in a distributed environment, but also sharing of operators, raising the issues of security and update policies for operators as well as data.

Lack of metadata management: Current systems exhibit only limited capability in expressing accuracy, precision and resolution of scientifically derived data. DBMSs for GIS and global change research should be able to provide some means for representing information about the accuracy and precision of data objects.

Inadequate output formats: Record-based DBMSs typically produce tabular output. However, for global change studies it is more important to be able to visualize a spatial distribution using a map, for example.

User interfaces: Interfaces are generally not intuitive to non-computer scientists. The use of special database languages, such as SQL, exacerbate the problem.

SPECIFIC REQUIREMENTS FOR GLOBAL CHANGE STUDIES

There are four types of requirements that must be met for global change studies: 1) Data Management, 2) Methodology Management, 3) Flexible Environments, and 4) Interface Considerations.

Data Management

Any study of global change must support spatial data access and analysis. Spatial data require a DBMS to

- Hold variable length records. Fixed length records are inadequate and inappropriate for representing the variety of spatial structures that may be encountered.
- Allow for spatial queries. Describing objects of interest by their location is more natural than specifying exact field values.
- Maintain integrity constraints. There are many interrelationships among data elements. Hiding implementation details from the user means that it might not be apparent how a given query will be processed—conflicting subqueries might be issued.

Different disciplines use different data formats (Townshend 91). Spatial data are commonly stored in either raster or vector formats. Each has its advantages and disadvantages. However, the format in which data are supplied depends more on the particular sensor type that was used than on the uses to which the data are put. Thus, satellite images are supplied in raster format while digital line graphs are vectors. However, the data user does not and should not care about this distinction; the specifics of data storage should be invisible. Thus, conversion between formats must be automatic and efficient. It should be treated as type coercion in high-level programming languages, working best when the user is unaware of it.

Detection of change requires that the DBMS access data using time as well as space. This introduces another set of issues, such as the need to

- Hold unevenly sampled data. Some temporal data are acquired at irregular intervals.
- Allow for temporal queries. Some, but not all, have counterparts in the spatial domain. Thus, temporal access primitives must be supplied (Qiu 92).

Data exist at many scales in space and time. There are several reasons. A researcher is typically interested in a specific range of scales, for example, days or weeks for Global Circulation models and years or decades for regional ecosystem changes. Not every query needs to refer to data at the finest level of resolution—it would be wasteful to estimate population at the same 10m resolution as SPOT imagery, for example. Also, cartographic products are available at a variety of scales; this functionality must be duplicated in any GIS. The multi-scale needs are

- Represent data at many scales in space and time.
- Link data at one level of resolution with data at higher and lower scales, if available.
- Provide interpolation operators to synthesize coarser scale data when needed.
- Allow for extremely low resolution yet rapid data browsing.

Methodology Management

Global change study is multidisciplinary science, where each discipline has its own methods. Global change study is also a new scientific horizon; new data, experiments, theories and technologies will result in new methods and results. Organizing these methods becomes a necessity.

The methods (or tools) used in global change study include:

- GIS functions
- Image processing
- Cartographic performance, e.g., projection transformation
- Global modeling
- Statistical analyses
- Expert systems

It is important that many of the implementation details be hidden from the user. For example, image processing operations typically need to access the image pixels is row-major order, yet the description of the algorithm should not refer to that order. Another implementation detail is file names; data should be accessed based upon semantics and the semantics should be understandable to the computer. File names fail this criterion. What is needed is a means of specifying what is desired at the appropriate level of description.

Flexible Environments

A scientific DBMS should provide a flexible and creative environment to allow users to manage their own methodologies. Recognizing that each discipline may have their own methods to analyze global environmental data, different users will want to use different methods in order to reach their goals. Therefore, a flexible DBMS should allow users to

- Organize their own procedures and data.
- Add new or override existing procedures and data.
- Share procedures and data once they have been verified.
- Browse for applicable data *and* operators.

Interface Considerations

In order to be usable to all levels of users, it is necessary to pay attention to the user interface. The interface must be intuitive, even for novices. It should be visually-oriented for input as well as output. Specific needs are

- The interface should be intuitive to all levels of users.
- The language used to describe queries and operations must be clear and unambiguous. Visual programming is one possible approach.
- Examine and repeat experiments. The experiments of others should be as easy to repeat as one's own.
- Generate graphical output. A temporal query might be displayed as a graph with time as one axis, for example. Since the result of a query will often be a geographic object, that object should be depicted as clearly as possible. This is more likely to be an image rather than a textual table.

CONCLUSIONS

We have examined the process of scientific discovery and related it to DBMS needs for global change research. A number of shortcoming with current DBMS technology were identified and specific requirements were given in several areas: Data Management, Methodology Management, Flexible Environments, and Interfaces. Meeting these requirements will pose a challenge to database researchers, but the payoff will be more usable tools for empowering global change researchers to understand the dynamic system they inhabit. We are now developing an architecture to meet these requirements (Hachem 92).

ACKNOWLEDGMENTS

The support of the National Science Foundation under Contract IRI-9116988 is gratefully acknowledged. The authors wish to thank Ke Qiu and Yuhong Zhang for their contributions to this work.

References

S. Aronoff 1989, Geographic Information Systems: A management perspective, WDL Publications, Ottawa.

M. Atkinson, F. Bancilhon, D. DeWitt, D. Maier, and S. Zdonik 1989, The Object-Oriented Database System Manifesto: Proc. Int. Conf. on Deductive and Object-Oriented Databases, pp. 40–57.

P.P. Chen 1976, The Entity-Relationship Model: Towards a Unified View of Data: ACM Trans. on Database Systems, Vol. 1, No. 1, pp. 9–36.

E.F. Codd 1970, A Relational Model for Large Data Banks: Comm. ACM, Vol. 13, No. 6, pp. 377–387.

Earth System Sciences Committee 1988, *Earth System Science: A closer view*, NASA, Washington, DC.

J.C. French, A.K. Jones, and J.L. Pfaltz 1990, Summary of the Final Report of the NSF Workshop on Scientific Database Management: *SIGMOD Record*, Vol. 19, No. 4, pp. 32–40.

O. Guenther and A. Buchmann 1990, Research Issues in Spatial Databases: *SIGMOD Record*, Vol. 19, No. 4, pp. 61–68.

N.I. Hachem, M.A. Gennert, and M.O. Ward 1992, A DBMS Architecture for Global Change Research: *Proc. ISY Conf. Earth and Space Science*, Pasadena, CA.

R. Hull and R. King 1987, Semantic Database Modeling: Survey, Applications, and Research Issues: *ACM Computing Surveys*, Vol. 19, No. 3, pp. 201–260.

P. Langley, H. Simon, G. Bradshaw, and J. Zytkow 1987, *Scientific Discovery*, MIT Press.

D. Maeir, J. Stein, A. Otis, and A. Purdy 1986, Development of an Object-Oriented DBMS: *Proc. Conf. Object-Oriented Programming Systems, Languages, and Applications*, pp. 472–482.

S. Morehouse 1991, The Role of Semantics in Geographic Data Modeling: *Proc. 4th Int. Symp. on Spatial Data Handling*, pp. 689–698.

Natural Environmental Research Council 1988, *Geographical Information in the Environmental Sciences*, Swindon.

P. van Oosterom and T. Vijlbrief 1991, Building a GIS on top of the Open DBMS Postgres: *Proc. EGIS '91 (European Conference on GIS)*, Vol. II, pp. 775–787.

M.T. Özsu and P. Valduriez 1991, *Principles of Distributed Database Systems*, Prentice Hall.

J. Peckham and F. Maryanski 1988, Semantic Data Models: *ACM Computing Surveys*, Vol. 20, No. 3, pp. 153–189.

K. Popper 1961, *The Logic of Scientific Discovery*, Science Editions.

S. Prashker 1991, Demonstration of the ZAPPER Software: *Proc. ACSM-ASPRS Annual Convention*, Vol. 2, pp. 270–278.

K. Qiu, N.I. Hachem, M.O. Ward, and M.A. Gennert 1992, Providing Temporal Support in Data Base Management Systems for Global Change Research: *Proc. SSDM '92*, Switzerland.

M.D. Soo 1991, Bibliography on Temporal Databases: *SIGMOD Record*, Vol. 20, No. 1.

R.B. Stam, R. Snodgrass 1988, A Bibliography on Temporal Databases: *Database Engineering*, Vol. 7, pp. 231–239.

M. Stonebraker, L.A. Rowe, and M. Hirohima 1990, The Implementation of POSTGRES: *IEEE Trans. Knowledge and Data Engineering*, Vol. 2, No. 1, pp. 125–142.

S.Y. Su 1983, SAM*: A Semantic Association Model for Corporate and Scientific-Statistical Databases: Inf. Sci. 29, pp. 151–199.

J.R.G. Townshend 1991, Environmental database and GIS: in Geographical Information Systems: Principles and applications, D.J. Maguire, M.F. Goodchild, and D.W. Rhind eds., London: Longman, pp. 201–210.

L. Vincent and P. Soille 1991, Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations: IEEE PAMI, Vol. 13, No. 6, pp. 583–598.

J.D. Ullman and C. Zaniolo 1990, Deductive Databases: Achievements and Future Directions: SIGMOD Record, Vol. 19, No. 4, pp. 75–83.