Summary of the Final Report of the NSF Workshop on Scientific Database Management

James C. French, Anita K. Jones, and John L. Pfaltz

Abstract

The National Science Foundation sponsored a two day workshop hosted by the University of Virginia on March 12-13, 1990 at which representatives from the earth, life, and space sciences met with computer scientists to discuss the issues facing the scientific community in the area of database management. The workshop[1] participants concluded that initiatives by the National Science Foundation and other funding agencies, as well as specific discipline professional societies are urgently needed to address the problems facing scientists with respect to data management. This article presents a condensed version of the workshop final report emphasizing the technical research issues.

1. Introduction and Background

Over the next decade the problems posed by the exponential growth of data in a variety of scientific disciplines will become increasingly pressing. For this reason, an NSF sponsored, interdisciplinary workshop on scientific database management was organized to consider these problems and possible solutions. It brought together computer scientists and serious user/proprietors of scientific data collections in the space, earth, and life sciences. Our objective was to discuss the issues involved in establishing and maintaining large scientific data collections, and to identify opportunities for improving their management and use. More particularly, we sought to assess the current state-of-the-art, assess whether the needs of the sciences are being met, identify the pressing problems in scientific database management, and identify opportunities for improvement.

Many issues regarding scientific databases are similar to those found in conventional business environments, but the focus is different. The relative importance of the issues associated with any data management undertaking is determined by the characteristics of the data and the anticipated operational environment. Much scientific data can be characterized by large volume, low update frequency, and indefinite retention. In the past, it has been generally safe to assume that scientific data resulting from experimental observations was never thrown away. The future volume of data will be staggering. Mapping the three billion nucleotide bases that make up the human genome will result in an enormous volume of data. The *Magellan* planetary probe will generate a trillion bytes of data over its five year life — more image data than all previous planetary probes combined. This suggests that much scientific data will not even be on-line.

A recent article describing the state of NSFNET characterized the problem of access to the net as being hampered by a diversity within the computer world that "verges on anarchy."[2] This same diversity poses an equally substantial barrier to the access of scientific data by those who need it. Indeed, one of the significant problems with scientific databases is largely logistical. Perhaps the greatest problem facing the scientist is the bewildering array of commercial and custom database interfaces, computer operating systems, and network protocols to be mastered in order to examine potentially relevant data.

---

[1]The report of the NSF Invitational Workshop on Scientific Database Management is available as Technical Report 90-21 from the Department of Computer Science, Thornton Hall, University of Virginia, Charlottesville, VA 22903. The workshop was attended by Don Batory, Joe Bredekamp, Francis Bretherton, Mike Carey, Y.T. Chien, Vernon Derr, Glenn Flierl, Nancy Flournoy, Ed Fox, Jim French, Hector Garcia-Molina, Greg Hamm, Roy Jenne, Anita Jones, David Kingsbury, Tom Kitchens, Barry Madore, Tom Marr, Bob McPherron, Steve Murray, Frank Olken, Gary Olsen, John Pfaltz, Bob Robbins, Larry Rosenberg, Peter Shames, Arie Shoshani, Ferris Webster, Don Wells, Greg Withee, John Wooley, and Maria Zemankova. The workshop was supported by NSF grant IRI-8917544. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the panels and do not necessarily reflect the views of the National Science Foundation.

[2] "Waiting for the National Research Network," *AAAS Observer*, March 3, 1989.

From the point of view of the practitioner, there are some relatively simple questions that must be answered in order to enhance the scientific research environment:

What data is available to me?
Where is it located?
How can I get it?

To provide the scientific community with the means to answer these and other questions, database researchers must examine the issues peculiar to scientific database management and the sharing of scientific data.

The purpose of this workshop was to examine the issues of scientific databases in more detail with the goal of producing a planning document to guide the NSF as it considered a new research initiative in this area. In addition to the computer science representation, the workshop participants were drawn from among the various disciplines of the earth (e.g., oceanography, climatology, geology), life (e.g., microbiology), and space (e.g., astronomy, astrophysics) sciences. The overall representation was approximately 40 percent computer science and 20 percent from each area of the physical sciences. Besides NSF, a number of government agencies were represented, including Department of Energy (DOE), National Oceanic and Atmospheric Administration (NOAA), National Aeronautics and Space Administration (NASA), National Radio Astronomy Observatory (NRAO), and National Center for Atmospheric Research (NCAR).

The workshop began with invited talks from each represented area with the objective of exposing both common and distinctly different data management problems. Participants then met in one of four panels to examine the relevant issues more closely. Panel representation was proportional across all disciplines. The panel topics were: (1) Multidisciplinary interfaces: standards, metadata, multimedia, etc.; (2) Emerging and New Technologies; (3) Core Tools: access methods, operators, analysis tools, etc.; and (4) Case Study: Ozone Hole. This case study was used as a vehicle for investigating data management needs, successes, and failures in a real mission environment.

This paper is a condensation of the workshop final report.

2. Dimensions of Scientific Database Systems

The workshop observed that databases can be characterized in terms of at least three dimensions (which need not be independent). They are:

level of interpretation,
intended analysis, and
source.

Characterizing data of interest along these dimensions helps to clarify salient aspects so that data management issues can be more clearly enunciated and explored.

In the discussion that follows we use the terms "data set" and "database." By data set we mean data related to a single experiment or mission. We use the term database more generally to denote any aggregate of data.

2.1. Level of interpretation: A scientific database may be a simple collection of raw data, or real-world observations, or it may be a collection of highly processed interpretations. At least two of the panels observed that this dimension manifestly affects what one expects of the database, and how one employs it. One proposed subdivision of this axis is

raw/sensor data: (seldom saved) raw values obtained directly from the measurement device;

calibrated data: (normally preserved) raw physical values, corrected with calibration operators;

validated data: calibrated data that has been filtered through quality assurance procedures (most

commonly used data for scientific purposes);

derived data: frequently aggregated data, such as gridded or averaged data; and

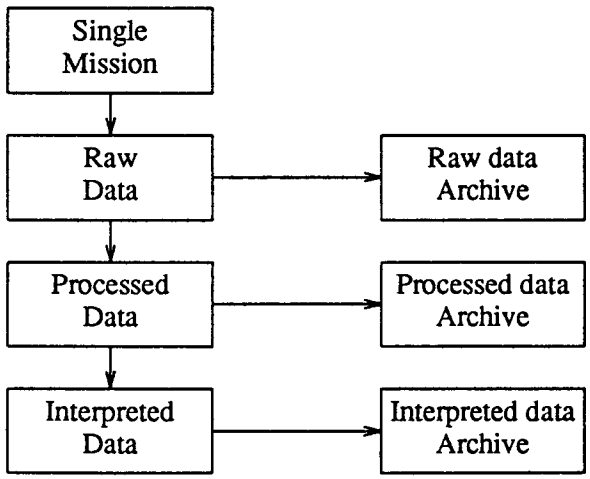
interpreted data: derived data that is related to other data sets, or to the literature of the field.

This sequence of successively greater interpretation need not be precisely correct. But it does indicate that the type of data in a data set can be highly dependent on its level of processing. Moreover, information about the processing must also be retained and distributed with any data set; this ancillary descriptive data, often called metadata, is vital to fully understanding and using a data set.

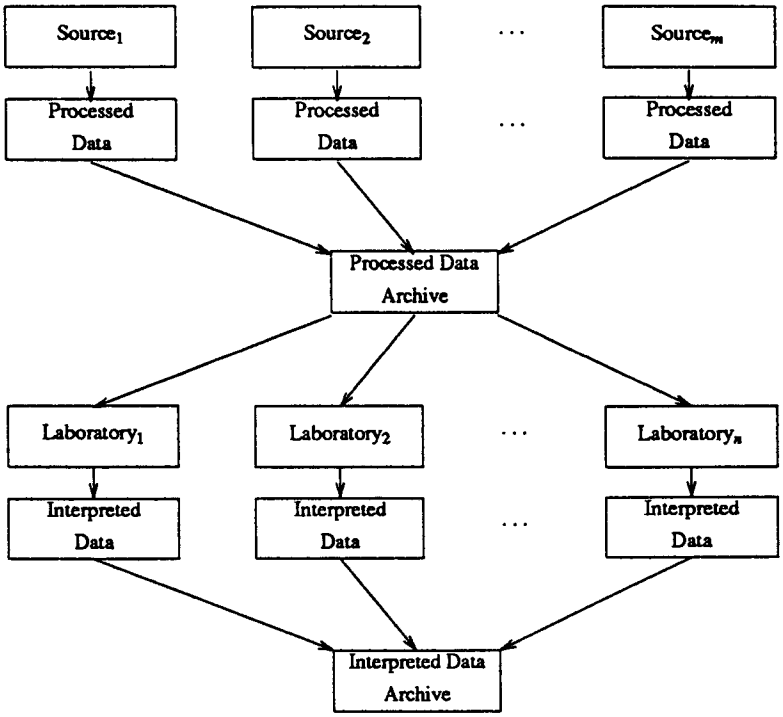
2.2. Intended Analysis: Our assumption is that all scientific data is subject to further analysis; otherwise there is little reason to retain it. The nature of such subsequent analysis frequently determines what particular representational format is most desirable. Much earth science data is analyzed statistically; time sequenced, multidimensional tables are common. A predominant activity in biological genome databases is elaborate pattern matching over linear, character data. Multispectral analysis in the space sciences apply transformations (e.g. Fourier) to very large two and three dimensional arrays. For each type of analytic processing, a database with different characteristics is most appropriate.

The criticism of currently available database technology is largely that it is designed for business data processing applications, and seems inadequate for scientific applications. Scientific applications pose a variety of complex requirements. Firstly, science data modeling requires new data types and support for multidimensional objects. For example, much of science data represents discrete sampling of functions of several dimensions, e.g., (x,y,z,t). Often the results are sequences, in which order is important. Secondly, the classes of transformations scientists wish to apply to data are larger and more complex than those within current database management systems and, in some cases, such as interpolation and Fourier analysis, rely on characteristics of the data which are not part of the data description. Thirdly, much more ancillary information is required. Fourthly, updating and correcting data sets is a very different process for scientific databases.

2.3. Source: This dimension, which is not generally mentioned in the database literature, may be the most fundamental. In Figure 2-1, we illustrate a familiar *single-source* database environment. Here we



Single-source Data
Figure 2-1



Multi-source Data
Figure 2-2

envision a single mission, such as the *Magellan* planetary probe, generating the data. Either raw or physical data may be retained in its original state in a raw data archive. Commonly, the raw data will be processed, by instrument calibration or by noise filtering, to generate a collection of more usable calibrated or validated data. Finally, this processed data will be interpreted in light of the original goals of the generating mission.

Both the syntactic and semantic complexity of the interpreted data will be much greater than any of its antecedent data. It will have different search and retrieval requirements. Possibly, only the interpreted data will be published.

In contrast to such a single-mission/single-source data archive one has data archives that are derived from multiple sources employing multiple data-generation protocols. Figure 2-2 illustrates a typical *multisource* collection of data. This structure would characterize the Human Genome project in which several different agencies, with independent funding, missions, and methodologies, generate processed data employing different computing systems and database management techniques. All will eventually contribute their data to a common data archive, such as GENBANK, which subsequently will become the data source for later interpretation by multiple research laboratories that also manage their local databases independently. In each of the local, multiple, and probably dynamic, databases, one would expect different retrieval and processing needs, as well as different documentation requirements.

This classification of databases in terms of level of interpretation, intended analysis, and source, however imperfect, helped clarify discussions at the workshop. In the following section, we list issues in scientific database management. The importance of an individual issue is often dependent on the position of the database of interest in this multidimensional classification.

3. Problems

All sciences have major data management problems; for example handling increasing data volume, metadata management, integration of database facilities with applications, finding data, access policy, ease of use. Different sciences seem to have different technical data management problems that are domain specific. In the following sections we have subdivided the problems raised at the workshop into main issues and lesser issues. This subdivision has been imposed to indicate a sense of relative importance to the reader without attempting a fruitless exercise of exactly ranking the problems.

3.1. Main Issues

The issues and problems discussed in this section received most of the attention of the participants.

3.1.1. Metadata: As discussed in section 2.1, scientific databases hold a wide spectrum of kinds of data. For the data to be meaningfully processed later, the metadata associated with the data must be preserved and accessible. This is the information required to identify data of interest based on content, validity, sources, preprocessing, or other selected properties. It is imperative that the metadata remain attached to the data. Metadata includes:

Who did what and when
Device characteristics
Transform definitions
Documentation and citations
Structure and format descriptions

3.1.2. Locating Data: Early in any scientific inquiry, the ability to find data becomes critical to the successful outcome of the investigation. Hypotheses need to be corroborated, or perhaps, archived data is to be mined for possible undiscovered properties. It becomes necessary to address questions such as

What data exists and where is it?
Is the data relevant to my interests?
Do useful data items exist?

This implies the need for a rather general data browsing capability providing facilities for locating and scanning data sets for indications of probable interest.

3.1.3. User Interfaces: To manipulate data and produce information, a scientist needs to access data and apply analysis tools in concert. Failure to integrate the data management and analysis environments restricts the productivity of the scientist. Integration becomes more important as one adds functionality to provide an automated assistant to the scientist. Such an assistant must track interlaced data and analysis steps.

The easier a user interface is to use, the more productive a scientist can be. In addition to being intuitive and easy to use, user interfaces need to

- be domain specific
- handle differing levels of user sophistication (novice/expert)
- browse across different database management systems (DBMSs)
- provide "hooks" for special application programs
- support tracking of data accessed across multiple DBMSs to maintain an audit trail of transformations applied to create each data set.

Achieving a system that will support multidisciplinary research across a variety of databases will be greatly facilitated by having sophisticated user interfaces hiding the heterogeneous reality and unifying disparate environments.

3.1.4. Flexible Representational Structures: Perhaps the single unifying cry of the workshop was that existing data models are inadequate for science data needs. The relational model has some advantages. However, the semantic gap between the relational model and what scientists need must be addressed. We must seek alternatives such as extending the relational paradigm, object-oriented database technology, extensible database technology, and logic databases for ways to efficiently support temporal, spatial, image, sequences, graph, and other more richly structured data.

3.1.5. Analysis Operators: One area of concern noted by most of the participants was the lack of appropriate operators within existing DBMS's for manipulating the kinds of data encountered in scientific applications. For example, more flexible comparison operators are necessary when attempting to match DNA sequences or retrieve image data. There was no agreement as to where these operators belong — within the DBMS as intrinsic operators or external to the DBMS as utilities or part of an analysis package. The approach used now is to have a commercial DBMS export data for use by external utilities. Often the data cannot be exported in a format compatible with the utility program so the scientist is forced to produce ASCII files that are subsequently massaged into an appropriate form. If results of the analysis are to be saved, the process must be reversed and the updated data imported back into the DBMS. Since there are no import/export standards this is a tedious and time-consuming process.

Extensible database research attempts to provide the mechanism for embedding custom operators into the DBMS. Object-oriented technology provides a basis for solution with methods and encapsulation. A philosophical question arises as to how much custom functionality is desirable within the DBMS. Rather than embed domain-specific operators in a DBMS, it may be more appropriate to create a standardized integrated analysis environment in which a DBMS can interact with a variety of useful tools.

3.1.6. Standards: Heterogeneity in data and operational environments is a fact of life. We must find ways to promote consistency within and across scientific disciplines. It is unreasonable to expect all disciplines to converge on some unifying standard for data model, data language, and communication; heterogeneity will continue to be a complicating factor. However, there are already instances of standardization within disciplines — the astrophysics community has endorsed FITS as its data interchange standard — and this trend should continue.

It was noted that the most successful standardization efforts arise when an organization creates a useful data format and associated analysis tools and then distributes them widely and also maintains them at no charge.

3.1.7. Standards for Data Citation: There was strong sentiment that data used in the conduct of an investigation should be cited prominently. A standard citation mechanism would allow researchers to locate and examine precisely the data used in the investigation. It would also give due credit to the data collectors.

It was noted that much of the interesting metadata is actually citations into the scientific literature. These citations too should be handled in a standard way so that where possible their content may be examined as part of a search for important data or to help assess the quality of data which is being browsed.

3.2. Other Issues

We have rated the following issues as less important issues because, either (1) there exist partial, although imperfect, solutions to the problem, (2) they seemed to be less frequently encountered, or (3) the problems are not readily amenable to a technological solution. While these may be less important from our perspective, there exist views in the scientific database management world (using the general dimensions described in Section 2) in which they can be very important.

3.2.1. Data Set Transmission: Data sets residing at one site (usually the collecting site or a designated repository) may have to be transmitted to the site where subsequent analysis will take place. Participants observed that there exist a number of wide area networks of sufficient bandwidth and reliability to handle most reasonably sized data sets. However, transmission of very large data sets may be slow. The delay in response time may be associated with the time to access and transmit the data set at the host site, as much as network delays.

3.2.2. Conversion of Data Set Format: A data set received from a foreign site may be in a format that is incompatible with the local analysis system. Subsequent data set conversion may depend upon adequate metadata to interpret the format and structure of the data, and more generally, upon the conventions expected by the local analysis system. Some discipline specific standards for data exchange already exist, such as FITS in the astrophysics community.

Analysis involving multiple data sets from disparate sources can be difficult. Relations obtained from Oracle, Ingres, or other relational DBMS need not be immediately comparable. With data coming from even less rigid data models, such as object-oriented DBMS, the problem is magnified. A straightforward technical approach involves converting all data sets to a local standard as described above. At a much deeper level, this problem involves the general issue of data fusion, which must take into account the semantics of the data in order to make meaningful comparisons.

Interoperability of multivendor DBMS's is essential to allow analysis programs running under the aegis of one DBMS to directly query/access data stored in a different DBMS.

3.2.3. Quality Assessment of a Data Set: Participants repeatedly noted the difficulty in assessing the quality of a received data set. While quality assessment has always been a fundamental scientific problem, many of the technical barriers arise due to insufficient metadata to interpret the data.

3.2.4. Volume of Scientific Data, Need for Permanent Archiving: The expected volume of sensor-generated scientific data is awesome. In the coming decade it will far outstrip the resources available to analyze it. The issue is: should (can) all of it be archived for possible later interpretation, or should (can) it be passed through some preliminary filter to determine what should be saved. The answers to these questions will be directly related to the cost-effectiveness of archival storage media.

That data production is often well-funded, while data management is poorly funded, if funded at all, was a recurrent theme.

3.2.5. Sociological Problems Data collecting Principal Investigators and their funding agencies have little incentive to release verified, but uninterpreted, data sets in a timely fashion. In fact, there are a number of sociological and monetary incentives not to do so.

It was repeatedly noted that data management is not an attractive career path within the scientific disciplines, whose primary goal is one of discovery. There is a need to educate both domain scientists and computer scientists in each others fields. However the time investment is viewed as a distraction from the major field.

This summary should convey to the reader the variety of problems faced by scientists in the management of their data. Unless these problems are addressed now, scientists in the 90's will find data management an increasing barrier to continued progress in their fields.

4. Recommendations

The discussions at the workshop clearly indicated the need for two distinct initiatives in scientific database management. We recommended that the federal agencies — in particular the National Science Foundation — create a broad research initiative directed toward the solution of the technical problems facing scientific database management. There are, however, many problems which fall outside the

purview of the NSF and which can most effectively be addressed by the scientific professional societies. Both roles are discussed below.

4.1. The Professional Societies

The professional societies, in their leadership role within disciplines, are the obvious vehicle for focusing attention on the data management problems within each discipline. They are also in the best position to represent the disciplines and to encourage cooperation in forums promoting interdisciplinary activity.

The following set of recommendations was directed to the professional societies. These recommendations may involve expansion of current activities and/or the creation of new initiatives.

- *Promote standardization of data interchange descriptions that use self-describing data formats.*
- *Standardize the description of data within each scientific discipline.*
- *Require appropriate citation of data and the deposit of relevant data into the appropriate archive before permitting publication in the societies' journals.*
- *Promote and fund workshops to investigate and recommend policy relative to the retention of data.*
- *Promote and fund workshops directed toward resolving the sociological problems hindering the development of sound data management policy.*

4.2. The National Science Foundation

Whereas the professional societies are particularly well positioned to influence the individual disciplines, the NSF can assume a more expansive leadership role in the effort to bring multiple disciplines into some state of conformance and cooperation. We recommended that the NSF launch a broad research initiative to address the technical problems impairing effective scientific database management. In some cases this will only require expansion of an existing NSF program. To be successful, this research initiative must involve multiple Directorates because the problems span many disciplines. A coordinated Foundation-wide research effort would allow all Directorates to share in the fruits of technical progress. As an effective starting point, NSF should stress exploratory implementations which produce prototypes of direct relevance to specific disciplines. This has the direct benefit of involving the domain scientists intimately and early in the design process. It is imperative to launch this research initiative now.

This workshop was chartered by NSF which asked the workshop participants what it should do to further scientific database management. While our recommendations are directed to NSF, many participants believe that other agencies, and even private funding sources, should participate in the aggressive execution of the recommendations that follow.

- *Perform research in methods for describing metadata and promote standardization in the management of metadata.*
- *NSF should include the matter of cataloging and publication of databases in its planning of NSFNET and other national networks. Research proposals for projects leading to databases*

should be required to include plans for publishing, cataloging, and either maintaining or transferring results to national archives.

- *Perform further basic research in storage device technology and in representation techniques for better utilization of the available capacity.*

- *Perform basic research in information science and information retrieval to improve the ability of interested researchers to locate relevant scientific data.*

- *Support the creation of one or more data analysis environments for science data that includes database and data archival processing. In these environments, special attention should be given to the integration of the user interface, the analysis component, and the database management system.*

- *Perform basic research in emerging database technologies such as object-oriented database systems, extensible database systems, and logic database systems. Further, explore alternatives to the relational model of data such as models directly supporting lists, sequences, and graphs.*

- *Perform research directed at solving the problems of heterogeneous data management environments.*

- *The NSF should coordinate a multiagency task force charged with promoting the creation of a harmonious environment which enhances the ability of scientists to exchange data freely and easily.*

Acknowledgements

This paper is a condensation of the final report of the NSF Invitational Workshop on Scientific Database Management. As such, it is clear that the participants of that workshop are responsible for the work reported here. Dr. Maria Zemankova deserves special recognition for her tireless efforts in promoting the importance of this work. We would also like to thank Dr. Won Kim for his careful reading of this paper and many constructive comments.