Scientific Databases – State of the Art and Future Directions

Maria Zemankova
MITRE Corporation
McLean, Virginia, U.S.A.
mzemanko@mitre.org

Yannis E. Ioannidis
Computer Sciences Dept.,
Univ. of Wisconsin-Madison,
WI 53706, U.S.A.
yannis@cs.wisc.edu

Conduct of scientific and engineering research is becoming critically dependent on effective management of scientific and engineering data and technical information. The rapid advances in scientific instrumentation, computer and communication technologies enable the scientists to collect, generate, process, and share unprecedented volumes of data. For example, the Earth Observing System Data and Information System (EOSDIS) has the task to manage the data from NASA's Earth science research satellites and field measurement programs, and other data essential for the interpretation of these measurements in support of global change research. Apart from being able to handle a stream of 1 terabyte of data daily by the year 2000, EOSDIS will also need to provide transparent access to heterogeneous data held in the archives of several US government agencies, organizations and countries. A single graphical user interface employing the Global Change Master Directory needs to help users locate data sets of interest among massive and diverse data sets, or find the appropriate data analysis tools, regardless of their location. Another major international effort in the area of human genome research faces some similar, as well as unique issues due to the complexity of the genome data, special querying requirements and much more heterogeneous collections of data.

Scientific databases can be viewed as critical repositories of knowledge, both existing and yet to be discovered. The emergence of subdisciplines in computational sciences attests to the trend of deriving scientific knowledge from data, as opposed to more traditional empirical approaches to research. In general, regardless of the discipline, success of research endeavor increasingly hinges on integration and management of numerous and diverse data sources, extensive data holdings, and complex data analysis products in a seamless environment that enhances or enables conduct of research.

Traditionally, research on the design, development, management, and use of databases has focused on concepts and requirements critical to business environments. Consequently, current database technology falls short of supporting the varied needs of scientific and engineering applications that are characterized by diverse and complex datatypes, very large volumes of data, special data processing needs, internationally distributed information sources and users, and extreme platforms heterogeneity. Furthermore, for the databases to be useful in scientific applications, it is important to realize that these systems need to designed for the scientists, rather than expecting that researchers are going to become well-versed in database technology details.

The pioneering projects in the area of scientific databases have shown that in order to make progress in scientific and engineering databases, the research must be driven by the needs of the discipline scientists or engineers and carried out in a close collaboration with computer and information scientists and engineers. Only this way, systems can be developed that enable a wide range of scientists and engineers to better utilize their data and other computing resources. Clearly, database researchers need to "suffer" an "overhead" of becoming familiar with the domain and understanding the researchers' needs. However, addressing the special characteristics and requirements

Proceedings of the 20th VLDB Conference
Santiago, Chile, 1994

of scientific and engineering databases should not be seen as a sacrifice on the part of the database researchers, as such efforts will open new areas for interesting database research and further database technology.

The aim of this interdisciplinary panel is to examine scientific databases achievements (or lessons learned) and explore future challenges for the database community. In particular, some of the topics discussed include:

- state of the art in scientific databases research and practice
- specific requirements in scientific data management (e.g., metadata, data calibration/validity, management of uncertainty in research hypotheses/results, designing and managing scientific experiments, and mobile/field databases)
- applicability and/or shortcomings of the current database technology in scientific or engineering domains
- priority of information processing needs in scientific communities (e.g., ease of access (e.g., Mosaic) vs. expressiveness of a query language (e.g., MSQL), integration of mass storage systems with database management systems, interoperability of database management systems with visualization and analysis tools, managing multi-media data, or allowing scientific collaboration)
- specific research or development issues that are very challenging for database researchers and/or can have extensive impact in conduct of research
- strategies for forming successful interdisciplinary teams

The panelists include experts in various subfields of database systems with experience in scientific databases research and/or development, as well as panelists from several scientific disciplines who use scientific data in their research and/or build or manage collections of scientific data.