TECHNIQUES FOR MANAGING VERY LARGE SCIENTIFIC DATABASES

Chairperson
William J. Campbell
Information Science and Technology Office
Code 930.4 NASA/GSFC
Greenbelt, Maryland 20771

Panelists:
Robert F. Cromp
Information Science and Technology Office
Code 930.4 NASA/GSFC
George Fekete
Hughes STX/NASA/GSFC
Ray Wall
Image Processing Applications and Development Section
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California
Michael Goldberg
Weather Sensors and Systems Department
MITRE Corporation
McLean, Virginia

Introduction

This panel will focus on the state-of-the-art in scientific data management, beginning with an overview of management of scientific data sets or databases, then explore the generic science requirements as well as a case example that drive the underlying data management system architecture, showing current technology limitations. A concept of intelligent information fusion is next presented with sufficient detail on how to integrate advanced technologies to enhance scientific production. Emphasis will be on user interfaces, spatial data structure, uses of neural networks for extracting information from scientific imagery, uses of object oriented database management systems (and why they're preferred over relational databases for scientific data), animation and visualization techniques.

Panelists Statements

Robert Cromp - Overview of an Intelligent Information Fusion System

NASA’s Intelligent Data Management (IDM) Project is conducting research into the development of data and information management systems that can handle the archiving and querying of data produced by Earth and space missions. Several unique challenges drive the design of these systems, including the volume of the data, the use and interpretation of the data’s temporal, spatial and spectral components, the size of the user base, and the desire for fast response times. The IDM group has developed an Intelligent Information Fusion System (IIFS) for testing approaches to handling the archiving and querying of terabyte-sized spatial databases. Major components of this system are the mass storage and its interactions with the rest of the system; the real-time planning and scheduling for processing the data; the extraction of metadata and subsequent construction of fast indices for organizing the data along various search dimensions; and the overall user interface.

The IIFS design is novel in a number of areas. Semantic data modeling techniques are used to organize the mass storage system to reduce the transfer times of the data to on-line devices and the mechanical motions of the supporting robotics. Neural networks and expert systems define how metadata is extracted to build up search indices to the underlying database. The metadata itself is organized in an object-oriented database which has special data structures for representing the multiple views of the data. A special data structure that maps directly between the Earth and a sphere organizes the data for efficient spatial querying. The user interface is configured dynamically at run-
time depending on the scientist's discipline and the current knowledge in the object database.

Gyorgy Fekete - Global Data Management and Visualization

One of the most pressing issues of our time is to answer the question of how industrial progress and modern lifestyle impact our global environment. This requires careful study of many sources of observational data. The analysis itself involves many scientists from many disciplines who sometimes do not interact with each other. Unfortunately, for whatever reasons, research groups seem to have evolved in partial isolation. Often the scope of contact is limited to meetings at conferences, reading of journal articles and occasional site-visits at each other's facilities. Virtually every project ends up employing a unique specialized representation method for the data to be studied, which in turn causes highly specialized and unportable analysis and graphing programs (visualization tools) to be written that are tailor-made to fit exactly one problem and one representation.

One source of problems is the lack of or too many standards (sic) in data interchange formats on all levels. Over the past few years there have been several notable attempts towards defining a common model and physical representation for data: CDF, HDF, netCDF, HDF V-Set, to name a few. To date, none of these has been successful in capturing the imagination of the absolute majority of scientists. It is possible that there will not be one best representation for all possible kinds of data, however, it is reasonable to expect that this number will be and should be small. Before any analysis can be performed, data has to be made accessible to investigators from many disciplines. Users have recognized that graphic browsing and visualization is essential to their data analysis needs. Although conventional methods for browsing spherically distributed global data fail to take the topological inconsistencies between the sphere and the plane into account, new methods of representation, such as the Sphere Quadtree (SQT) and associated algorithms make it possible to build incremental global browsing tools for spherical images.

Ray Wall - Computer Animating Space Imagery

Animations of planetary images have proven to be an innovative and useful modality for scientists analyzing large data sets. Whether touring the Valles Mariners on Mars, flying over Sif Mons on Venus or simply observing the atmospheric dynamics of the Great Red Spot on Jupiter, the effects are inspiring and thought provoking. What more convenient method could there be for scientists to observe spatial relationships, structures and process dynamics than through a touring metaphor? A fundamental tenet of visualization is data interaction. Broadcast quality color interactive navigation over very large data sets at video rates is probably beyond any general purpose computer today, not to mention the average scientist's desktop computer. Is the more expensive special purpose hardware necessary for science? Simple interactive navigation is only the tip of the computational iceberg. The computational demands really get high when we consider the need to overlay complex models into observational data and a desire to interact with these same models in response to the data. Provided there are no limiting performance bounds placed by the laws of physics, it is probably true that the technology will ultimately support these performance requirements—so the question is what is a reasonable goal to pursue at an accelerated pace?

The data volume potential of the next generation of space science instruments poses a true challenge to technology innovation. Data rates for high rate instruments that are chosen by dividing the number of potential products by the number of available investigators in the community provides a sobering realization that the ability to produce useful data far outstrips the ability to convert this data to scientific statement. Is animation a useful tool here? If we are to take full advantage of animating data as a means of coping with extremely large data sets, the ability to generate animations must be greatly streamlined. The tools used to construct animations, and access ancillary data must be simplified. The simplification must address the issues connected with selecting, merging and registering disparate data--currently a time consuming, tedious task.

Is the role of animation in science well understood? Dynamic data offers the clearest demonstration of scientific utility. But what about simulated flyovers and human scale terrain roaming? Existing capability to tour natural sciences provides an opportunity to explore these data sets through a virtual reality interface. The more a scientist can be immersed into a world of natural structures, processes and relationships described by the data,
the less they are distracted by the artificialities of
the observational environment. For scientific
analysis how deep must this immersion be? Is
virtual reality a valuable modality for science or
just another pretty interface? Linking these user
interface modes with compatible analysis tools,
providing a means to access and display related data
and providing a window to the source data is
essential if the payoffs of animation are to be
realized. Where do we start? How do we build on
existing technology? Do we build directly on what
exists or do we start over? Computer animations
have a definite role in the display and analysis of
scientific data. Defining and implementing the
capability that meets that role is the challenge.

Michael Goldberg - Automatic Content-Based
Cataloging

Archived satellite imagery is an important yet
underutilized source of data for Earth science
research. Currently, it is difficult for scientists to
find the images they need unless there exists
independent information about where and when
images of interest were captured. One reason for
this is that the catalog systems for the archives do
not sufficiently characterize the information content
of the images themselves. Instead, they tend to
index imagery only according to ancillary attributes
such as sensor name, geographic location, and
capture time. Given the enormous volume of data
being archived, images of scientific importance can
remain unnoticed indefinitely.

Automatic content-based cataloging is a practical
data management approach designed to improve
multi-disciplinary access to archived satellite
images. Spectral and spatial attributes are
automatically extracted from images and then
cataloged in a traditional database management
system alongside standard ancillary attributes.
Scientists can build database queries incorporating
these new attributes to reduce search time or to
discover previously overlooked images. A low-
resolution image browse summarizing extracted
attributes aids scientists in visualizing and
evaluating the criteria that were used in the database
search. Automatic content-based cataloging is
applicable to both "raw" and "product" images,
whether newly captured or previously archived.

Panel Participants

William J. Campbell is the Head of the
Information Science and Technology Office at
NASA Goddard Space Flight Center. His research
interests are in the development of user interfaces,
graphical data representation, object-oriented data
management, automatic data cataloging and
characterization, advanced data structures, image
processing, knowledge acquisition and resource
planning and scheduling. His previous work at
GSFC concentrated on development of large-scale
data information systems for Earth science and on
the integration and modeling of satellite remote
sensing data into geographic information systems.
He also served five years as the manager of NASA's
Eastern Regional Remote Sensing Center's remote
sensing training program. He has an M.S. degree in
Physical Geography from Southern Illinois
University and has over 35 publications relevant to
his work experience. He also serves as an
Associate Editor for the Photogrammetric
Engineering and Remote Sensing Journal, as a
member of the Committee on Information,
Robotics, and Intelligent Systems, National
Science Foundation and as a consultant and
Committee Member of the National Resource
Council of the National Academy of Sciences.

Robert F. Cromp is Principal Investigator of the
Intelligent Data Management project at NASA,
Goddard Space Flight Center at the National Space
Science Data Center. He received the B.A. degree
in computer science from the State University of
New York at Buffalo in 1982, and the M.S. and
Ph.D. degrees in computer science from Arizona
State University, Tempe, in 1983 and 1988,
respectively. He has published in the areas of
knowledge acquisition and representation, the
theory of strategies, information fusion, automated
extraction of metadata from images, machine
learning, parallel algorithms, automated expert
system development, geographical information
systems, natural language processing and discrete
mathematics. He also teaches at the University of
Maryland/University College and Bowie State
University. Dr. Cromp is a member of AAAI,
ACM, IEEE, INNS, ASPRS and MAA.

Gyorgy Fekete received a Ph.D. degree in 1988 in
Computer Science from the University of Maryland
at College Park. Before his graduate studies,
between 1979-1980 he was a research associate at
the Hungarian Academy of Science's Computer and
Automation Institute in Budapest, Hungary.
Beginning in 1986 he has spent two and a half
years as a specialist at the VLSI design lab of the
Systems Research Center at the University of
Maryland. Since 1989 he has been with Hughes-STX
at the National Space Science Data Center,
where he is involved with designing and
implementing visualization and graphic browse
tools for observational data.
Ray Wall received a Ph.D degree in Engineering from the University of California at Los Angeles in 1974. Currently he is the Manager of the Image Processing Applications and Development Section at Jet Propulsion Laboratory and an Associate Professor of Electrical Engineering and Computer Science at California State University at Los Angeles. The Image Processing Applications and Development Section provides research and development as well as operational support to the JPL community of planetary projects and science users. A key activity within the Section is the creation of 3D perspective animations of remotely sensed planetary data. These animations simulate overflights of planetary terrain. Examples include "Mars the Movie" and the current overflight animations of Venus constructed from SAR imagery produced by the Magellan Project.

Michael Goldberg has 15 years of experience in the processing, management, and analysis of satellite imagery. He has been successful in integrating image processing, database management, and geographic information system technology to meet end user needs. He has supplemented his practical experience with eight publications and with participation in the information processing activities of the American Meteorological Society. Mr. Goldberg has a B.S. in Mathematics from the University of Maryland and an M.S. in Computer Science from the Johns Hopkins University.