User Scientific Data Systems: Experience Report

Elaine R. Dobinson
Jet Pro ulsion Laboratory
California institute of Technology

introduction

This paper presents an abbreviated history of NASA science data management system development over the past ten years by selecting two case studies, each representative of a distinct era of science data management systems. The particular problems encountered by each of these systems, and the technical approaches to their solutions, have both taken advantage of and pushed the leading edge of data management technology. The special problems of managing science data and their associated metadata will be discussed.

In the early 1980s, influenced by the National Academy of Sciences Space Science Board CODMAC Reports, NASA funded several pilot data systems development projects to be based upon the key concept of the discipline data management unit. The data systems were organized into systems for separate science disciplines in order to serve the needs of particular science communities, and to provide each community with the data needed for its own research. These data systems included the Climate Data System, the Land Data System, the Oceans Data System, and the Planetary Data System. This paper will focus on the Planetary Data System, but the problems encountered are typical of the others as well. All were designed to service a particular disciplinary group, and all were originally thought of as being self-contained.

All of the initial pilot systems met with varying degrees of success, and are in some form operational today. The second era of data management systems, the era we are in at the moment, deals with the formidable task of integrating some of these stand-alone systems into a single service to provide data to the ever growing inter-disciplinary science research community. The broad community of earth scientists, focusing on the study of global change, is not only highly inter-disciplinary, but inter-agency, and even international. Part two of this paper will discuss some of the special needs of this research community for data, and the resultant challenges to data management technology of the data management system for the NASA Earth Observing System (EOS).

Case i: The NASA Planetary Data System (PDS)

Brief Background

The PDS was jointly designed by members of the planetary science community from around the country and data system developers at the Jet Propulsion Laboratory from 1985 through delivery of the initial version in 1989. This original version primarily concerned itself with science data already collected by previous NASA missions over the past two decades. Currently PDS has evolved, and continues to evolve, to archive and distribute data from current planetary missions as soon as the data are available. Work is also underway in the PDS project to plan for the archiving of data from future missions not yet flown.

The Early Vision

From early in its development the vision of the PDS was that of a system whose science data products would be localized at various planetary sub-discipline nodes and whose directory and catalog metadata would be centralized and managed at a central node at JPL. Detailed metadata about individual data granules as well as local physical metadata known collectively as inventories would be kept locally with the data products. The directory and catalog metadata would be be available to the local systems in a client/server mode so that a science user could access any of the relevant metadata from wherever he was located. The data products themselves would be labelled with self-describing metadata **in a standard form and distributed by either the discipline** nodes or the National Space Science Data Center (NSSDC), depending on the size of the products.

Largely because the prevailing data management technology of the 1980's was relational, the PDS metadata database was designed and built as a relational system. Important information

relating to the data products about the spacecraft, instruments, investigators, processing algorithms, etc., was all organized into relations and linked by relational operators to provide an ad-hoc query capability for data access. Testbed data sets from early missions were loaded and the system was released to the community for evaluation.

The Success and the Problems

The PDS was considered quite successful at doing what it was supposed to, i.e., making planetary data available to its community. However, the task of maintaining the system in its operational mode required the loading of many more data sets. This process required the data producers to provide the rich suite of metadata, which made the PDS so useful, in a highly structured form for ingestion into a relational database. In many cases the metadata already exist in the form of documents, journal articles, or data record headers. The scientist has to rework these metadata, and in some cases do some digging, to provide the PDS with its required inputs. This has been loudly complained about. Suddenly, the grand and glorious catalog that provides such a wealth of information is being called too expensive to maintain by the very community of scientists who designed it.

One solution to this problem is the automation of metadata collection. Planning ahead for the archiving of data in the early stages of the flight project would certainly help ensure that all of the required metadata were electronically present. Efforts in this direction are currently occurring with the Mars Observer and Cassini projects, Nevertheless, it is not always practical to carry along all of the metadata required by the ultimate archive system, and it is not always possible to identify all of the relevant pieces of metadata a priori, so the problem of ingesting other metadata as the system operates seems likely to occur even so. New technologies associated with object-oriented and multimedia databases may make the native forms of the metadata (science papers, videos, software, documents) more utilizable within the data system. Other ways of linking the vast amounts of textual information (such as WAIS) and integrating this information into the data system also need exploration.

Case 11: EosDIS (Version O)

Data systems belonging to the second generation of NASA systems go far beyond their predecessors of the single-discipline self-contained kind. These new systems, of which the EOS Data and Information System (EosDIS) is a prime example, must necessarily, for several reasons, build upon the systems already in place. These reasons have to do with cost (it's usually too expensive to start from scratch), logistics (most scientists do not want to give up their local capabilities), and the sheer volume of the data to be encompassed.

Version O of EosDIS has been chartered to prototype various approaches to interconnecting the underlying data systems without disrupting service to the local users. This experience has brought to light many challenges to current data management technology.

Data System Heterogeneity

Probably the most difficult and challenging problem faced by the EOS data system developers is that of integrating widely distributed! autonomous, heterogeneous data systems into a unified whole. NASA has identified eight institutions to serve as the Distributed Active Archive Centers for the earth science data collected in the past, the present, and the future. The DAACs as they are called are either the earlier discipline data systems built a generation ago or conglomerations of these. Each has a distinct coverage of earth science disciplines. The DAACS will upgrade their own data systems to handle their new data responsibilities, and the Information Management System (IMS) component of EosDIS will integrate these DAACs into a unified whole, providing any of the data to any scientist with complete location transparency. This requirement, known as "one stop shopping" in EOS circles, unveils all sorts of issues stemming from both system and data heterogeneity. Currently, a data dictionary is bein developed to document the local DAAC vocabularies so that approaches to the resolution of differences can be worked and true integration of the underlying inventories of data can be achieved.

In addition to integrating with all of the DAACS, the EosDIS also needs to couple with another data system to provide directory information to the earth scientists. The NASA Global Change Master Directory at the NSSDC is yet another source for data heterogeneity problems in that its

vocabulary serves an even broader community and needs to be merged with the terminology of the DAACs.

Metadata Generation and Utilization

The problems of automating the collection of metadata and of being able to utilize data and metadata in many different forms identified earlier in the discussion of the PDS are also present in EosDIS and even more critical because of the massive amounts of data to be generated. Multi-media and object oriented technologies to deal with the variety of data forms, and intelligent systems to generate the metadata from the content of the science data are all new technologies that may prove indispensable. In addition, new approaches to spatial and temporal searches, as well as sophisticated graphical interfaces and visualization of metadata, are needed to help locate data of interest from such a huge pool.

Summary

Problems of managing science data and associated metadata exist in both generations of data systems, though the new systems pose challenges on a much larger scale. This paper has raised some of the more pressing issues faced by the author currently. Progress in the solutions to these issues will benefit the science data management community as a whole as I'm certain that these are not NASA or space science specific.

Common Problems and Topics for Discussion

Multi-media Databases and Object-Oriented Approaches for Storing and Linking Science Data and Metadata

Planning for Metadata Generation in Mission Design

Evolvable, Extensible Systems

Interoperability between Heterogeneous Database Systems

Standards for Metadata

The research described in this paper was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.