PB88-188446

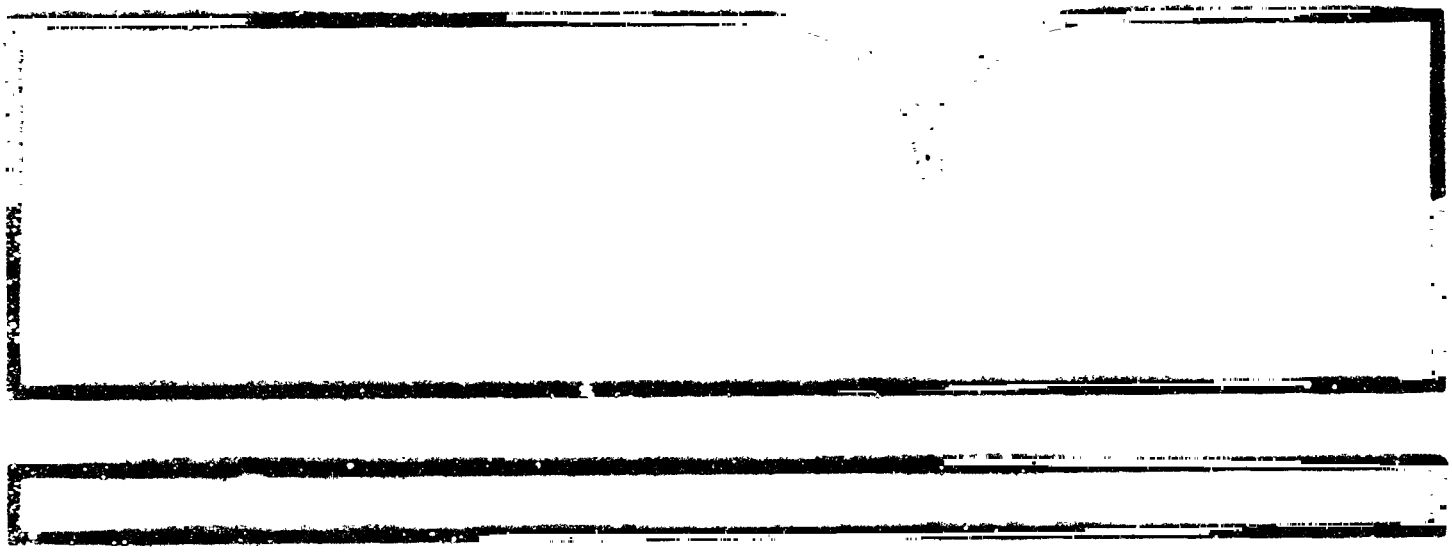
Issues and Recommendations Associated with Distributed Computation and Data Management Systems for the Space Sciences

National Research Council, Washington, DC

Prepared for

National Aeronautics and Space Administration
Washington, DC

1986



U.S. Department of Commerce
National Technical Information Service
NTIS

REPORT DOCUMENTATION PAGE	1. REPORT NO.	2.	3. Recipient's Accession No. PB88 188446/AS
4. Title and Subtitle: Issues and recommendations associated with distributed computation and data management systems for the space sciences.			5. Report Date: 1986 6.
7. Author(s)			8. Performing Organization Rept. No.
9. Performing Organization Name and Address: National Research Council Committee on Data Management and Computation Space Scienc Board Commission on Physical Sciences, Mathematics and Resources Washington, DC 20418			10. Project/Task/Work Unit No. 11. Contract(C) or Grant(G) No. (C) (G) NASW 3482
12. Sponsoring Organization Name and Address: National Aeronautics and Space Administration Washington, DC			13. Type of Report & Period Covered 14.

15. Supplementary Notes

16. Abstract (Limit: 200 words)

The primary purpose of this report is to explore management approaches and technology developments for computation and data management systems designed to meet future needs in the space sciences. This report builds on work presented in previous reports on solar-terrestrial and planetary reports, broadening the outlook to all of the space sciences, and considering policy issues aspects related to coordination between data centers, missions, and ongoing research activities, because it is perceived that the rapid growth of data and the wide geographic distribution of relevant facilities will present especially troublesome problems for data archiving, distribution, and analysis.

17. Document Analysis a. Descriptors

Space sciences; computation; data management systems; electronic data processing; distributed processing;

b. Identifiers/Open-Ended Terms

CODMAC ; NASA

c. COSATI Field/Group

18. Availability Statement	19. Security Class (This Report)	21. No. of Pages
Distribution unlimited	unclassified	124
	20. Security Class (This Page)	22. Price PC A19.95/MF A6.95

NRC NAS NAE IOM

Issues and Recommendations Associated with Distributed Computation and Data Management Systems for the Space Sciences

Issues and Recommendations Associated with Distributed Computation and Data Management Systems for the Space Sciences

Committee on Data Management and Computation
Space Science Board
Commission on Physical Sciences, Mathematics, and Resources
National Research Council

NATIONAL ACADEMY PRESS
Washington, D.C. 1986

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This report has been reviewed by a group other than the authors according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

The National Research Council was established by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and of advising the federal government. The Council operates in accordance with general policies determined by the Academy under the authority of its congressional charter of 1863, which establishes the Academy as a private, nonprofit, self-governing membership corporation. The Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in the conduct of their services to the government, the public, and the scientific and engineering communities. It is administered jointly by both Academies and the Institute of Medicine. The National Academy of Engineering and the Institute of Medicine were established in 1964 and 1970, respectively, under the charter of the National Academy of Sciences.

SPONSOR: This project was supported by Contract NASW 3482 between the National Academy of Sciences and the National Aeronautics and Space Administration.

Printed in the United States of America

COMMITTEE ON DATA MANAGEMENT AND COMPUTATION

Raymond E. Arvidson, Washington University, Chairman
Garry Hunt, Imperial College, London, England
David Landgrebe, Purdue University
Elliott Levinthal, Defense Science Office, DARPA
George H. Ludwig, University of Colorado
Thomas B. McCord, The Hawaii Institute of Geophysics
Ethan Schreier, Space Telescope Science Institute
Raymond Walker, University of California
Gio Wiederhold, Stanford University

SPACE SCIENCE BOARD

COMMISSION ON PHYSICAL SCIENCES, MATHEMATICS, AND RESOURCES

PREFACE

In 1982, the National Research Council published the results of several years of deliberations by the Space Science Board's Committee on Data Management and Computation (Data Management and Computation, Volume 1: Issues and Recommendations, NAP, 1982). Largely on the basis of a number of case histories of space missions, scientific processing facilities, and data centers, CODMAC (1) summarized the major problems that have been impediments to extraction of science information from space-acquired data, (2) recommended a number of general steps for improvement, and (3) developed a set of principles for successful management of scientific data. CODMAC also suggested how application of the principles in a variety of situations, ranging from data processing systems controlled by principal investigators, to the management of national data centers, could result in greater scientific yields from data sets.

Publication and distribution of the initial CODMAC document, followed by continuing dialogues among CODMAC, NASA, and members of the space science community, have served to uncover a number of further issues and problems related to space science data management and computation. A number of steps have been taken to meet some of the issues and to correct some of the problems identified in these discussions. These steps have included (1) consolidation of a number of management activities under the Information Systems Office of the Office of Space Sciences and Applications of NASA; (2) initiation of pilot data systems by that office to improve computation and management of data in various space science disciplines; (3) consideration of data problems and solutions by the solar and space physics community (Solar-Terrestrial Data Access, Distribution, and Archiving, NAP, 1984) and by

the planetary sciences community (The Planetary Data System, 1984); (4) initiation of a Computer Sciences Program in the Office of Applications and Space Technology, NASA; and (5) an evaluation and restructuring of the role of the National Space Science Data Center.

Although progress has been made, a major problem became evident during CODMAC discussions that followed publication of the initial document. The problem is that an overall vision is still lacking within NASA as to what requirements scientists will have on systems that are designed to handle, process, and store the significant quantities of data expected from future missions. For example, the 1984 NASA Space Systems Technology Model Executive Summary (NASA, 1984) devotes only two paragraphs out of a text of 278 pages to problems related to extraction of information once the data are on the ground. The importance of computation and data management associated with extraction of scientific information cannot be understated. In fact, the information and knowledge extracted from space science data should be the ultimate measure of mission success.

The primary purpose of the current document is to explore management approaches and technology developments for computation and data management systems designed to meet future needs in the space sciences. This report builds on work presented in the solar-terrestrial and the planetary reports cited above, broadening the outlook to all of the space sciences, and considering policy issues that transcend the individual disciplines. We stress aspects related to coordination between data centers, missions, and ongoing research activities, because we perceive that the rapid growth of data and the wide geographic distribution of relevant facilities will present especially troublesome problems for data archiving, distribution, and analysis. We note that our results are applicable not only to NASA, but also to other agencies, such as NOAA, that are involved in acquisition and analysis of large data sets.

A number of individuals need to be acknowledged who have contributed to this report, as participants in CODMAC's 1983 summer study, and as ongoing participants at CODMAC meetings and writing sessions. Those individuals include John Estes, University of California at Santa Barbara; Ted Albert, U.S. Geological Survey; Arthur Lane and Thomas Duxbury, Jet Propulsion Laboratory; George Pieper and Peter Bracken, Goddard Space Flight

Center; Michael Devirian and Caldwell McCoy, NASA Headquarters; and Lawrence Bolef, Washington University. Finally, Kristine Henrick, Susan Slavney, and Carol Martin, Washington University, should be thanked for their steady support in manuscript and figure preparation.

CONTENTS

1. EXECUTIVE SUMMARY

1.A. COMPUTATION AND DATA MANAGEMENT CHALLENGES IN THE 1980s AND 1990s

Projected rates of growth of spaceborne data over the next decade for the planetary sciences, astronomy, solar and space physics, and the earth sciences can be modeled with exponential growth functions, with doubling periods averaging only a few years. For example, the 30 or so terabits of solar and space physics data now in the National Space Science Data Center will probably increase by over an order of magnitude within a decade. Earth sciences data could increase by over 2 orders of magnitude during the same period, if such instruments as imaging radars and spectrometers are flown for extended periods of time and even if only data from science-dedicated experiments are retained.

Based on a number of example research scenarios envisioned in the space sciences for the 1980s and 1990s, requirements to search through, select, acquire, process, and store a wide variety of data sets to solve given problems will grow significantly. In some cases, data acquired over long periods of time will need to be analyzed. In addition, a wide range of laboratory and in-situ information will need to be integrated with the spaceborne observations. In some cases, the data needed to solve given problems will need to be obtained from geographically dispersed sites, including archives that are not under NASA's direct control. These complex requirements have come about in part because the space sciences are moving from a period of exploration, such as mapping the surface of the sun or Mars, or surveying stellar infrared sources, to a mode of intensive scientific analyses. In this intensive mode, a variety of

quantitative data need to be analyzed to constrain the increasingly sophisticated models.

The rapid growth rates, the geographically distributed nature of space science data, and the increasingly complex data uses imply that data handling, processing, and storage requirements will increase dramatically with time, at least at the same rates as those for growth of spaceborne data. Likewise, management of computation and data management systems designed to meet the complex needs will tend to be more complex than in the past. New, innovative ways must be found to meet both the management and the technology challenges imposed by these rapidly growing requirements on future computation and data management systems. We perceive that the management challenges are far greater than the technology challenges. The purpose of this report is to suggest reasonable approaches to meet these challenges. We stress management recommendations, and we consider technologies that should be utilized or developed to implement our recommendations.

1.B. DISTRIBUTED SPACE SCIENCE DATA MANAGEMENT UNIT SYSTEMS AS AN APPROACH

We delineate three major types of Space Science Data Management Units (SSDMUs) that are key elements to meeting the computation and data management challenges over the next decade: data centers, data repositories, and active data base sites. Data centers are defined as facilities housing data sets and associated information that require long-term maintenance because of the likelihood of use in future research activities; data repositories are facilities maintaining relatively large volumes of data in temporary buffers (e.g., mission data repository); and active data base sites house data being used intensively in research. Based on past experience, data acquired by an instrument science team associated with a given mission are usually held in a repository for the length of the mission. The data in the repository are usually reduced to a form appropriate only for initial analyses, to serve as a basis for more detailed analyses at active data base sites. The active data base sites are usually located close to research scientists at NASA centers, universities, and other research-oriented institutions. Upon completion of missions, data repositories have migrated to data centers, but often without proper documentation. Usually, data from active data

bases have disappeared once the research group controlling the data disbanded or moved on to other activities. Loss of these data is particularly unfortunate since in many cases these data are of high quality and of interest to other researchers.

Explicit delineation of the functions of data centers, repositories, and relevant active data base sites, including definition of the interfaces between them, would help ensure transfer of more complete, better documented data to the data centers. In fact, where appropriate, data centers, repository sites, and active data base sites should be considered as segments of geographically distributed information systems designed to serve the space science disciplines. As noted, the key to successful implementation of distributed information systems involving data centers, repositories, and active data bases is to rigorously delineate the roles and responsibilities of each SSDMU segment of the systems. We recommend that space science data centers (e.g., National Space Science Data Center), in coordination with their user communities, play a major role in organizing and overseeing SSDMU systems, including managing any information networks connecting the segments, providing directories and catalogs of relevant data, and stipulating standards and protocols to be used within the system. An approach that actively involves all three types of SSDMUs is quite a departure from past practices, where data centers were typically the last places for delivery of data, often without local expertise or enough supporting information to be fully useful. When combined with a renewed emphasis on the part of the space science community to produce high-quality, documented data, this approach, where data centers are involved in activities ranging from missions to in-depth data analyses, will help to ensure that useful data will be entered into the centers for use by the broad space science community. The present management structure at NASA Headquarters can be utilized to guide development and operation of distributed SSDMU systems. The Information Systems Office should manage those aspects of the systems that are of facility class in size or that transcend missions or disciplines.

Developing geographically distributed information systems involving data centers, repositories, and active data bases should be accomplished in an evolutionary fashion, and should include experiments aimed at deriving the best management methods and technologies to utilize.

The current pilot activities (Pilot Ocean Data System; Pilot Climate Data System; Pilot Planetary Data System; Pilot Land Data System) funded by NASA's Information Systems Office are designed to develop experimental computation and data management systems for use by the space science community. These pilot activities offer the advantages of the involvement of the community and an evolutionary approach to SSDMU implementation. We recommend that these pilot information system efforts continue and expand to include other space science disciplines, but with an overall goal of moving toward implementation of distributed information systems. We also recommend a focusing of pilot activities to experiment with management and technology issues associated with geographically dispersed SSDMUs that are linked to form distributed computation and data management systems. The pilot studies should also be focused to test whether or not discipline-oriented data centers offer better service and data quality than universal or space-science-wide data centers.

1.C. TECHNOLOGY RECOMMENDATIONS

NASA's posture in terms of new computation and data management technologies for SSDMUs and SSDMU systems should largely be to maintain an awareness of advances, and to assist the space science community in utilizing the technologies in meaningful ways. Awareness is the key because most technology advances in computation and data management will probably arise through industry-supported activities. There is an important area where unique requirements of the space science community suggest that technology development efforts are needed. That area deals with development of portable software packages that are designed for wide use in the space science community. Portability means developing software in higher level languages, in reasonably machine-independent form, and with use of acceptable standards. Widespread use of such packages should alleviate some of the current problems in transfer of data between SSDMUs and should facilitate distributed archiving and processing of data. An expert systems approach to analyses of imaging spectrometer data, and coupling of advanced data base management software with spatially and temporally tagged vector and array data, are but two examples of

needed software development efforts that are not being vigorously pursued by industry at this time.

Technology solutions to computation and data management problems must be tailored to the type of SSDMU under consideration. For example, for SSDMUs involving small research groups, simple work stations, consisting of microprocessor-based terminals, with modest storage devices and other appropriate peripherals may be sufficient. Other groups will require minicomputers and a number of special-purpose peripherals. On the other hand, SSDMUs with a charter for pipeline processing of significant quantities of data, or for long-term maintenance of data, will require more sophisticated and capable data handling, processing, and storage systems.

Computation and data management capabilities must be significantly upgraded at data centers, repositories, and at active data base sites, even if just to maintain the current processing status quo. This statement is based on comparing projected rates of increase of processing speed, storage capacity, and communication rates that will be available at any given time, with computation and data management needs required to extract scientific information from the expected space science data sets. The analysis was conducted by assuming that (1) at least the same fraction of data currently processed would be processed in the future, and (2) the new technologies would be acquired using current funding levels, scaled for inflation. In most cases, data growth rates outpace the rates of growth of computation and data management power at constant cost. NASA should work closely with the space science community on upgrading computation and data management systems in ways that will best meet the rapidly increasing demands in reasonable ways. For example, minicomputers with advanced, high-speed work stations offer one means of significant upgrading the processing capabilities for relatively small SSDMUs.

High-speed parallel processors and other large computational machines will continue to be beyond the funding levels of most research groups during the next decade, although based on data processing requirements, the need for access to these machines will probably grow in scientific research. These machines will be located at a few sites. Remote access to such large systems will be needed and will be a problem unless attention is given to how to remotely access and actively utilize these large machines. NASA should provide effective access to the high-speed machines at NASA centers, such as the

massively parallel processor (MPP) at the Goddard Space Flight Center (GSFC) or the Crays at the NASA Ames Research Center. Such use should incorporate the ability to process data remotely and to input and extract data in reasonable ways. NASA should also continue development of software that takes advantage of the computational speed of these large, facility-class processors. Finally, NASA should vigorously pursue investigation of low-cost concurrent processors that can be placed at local sites, thereby upgrading systems under control of the local SSDMUs.

Electronic communications will be essential in SSDMU information networks, given the geographically distributed nature of data centers, repositories, and active data base sites. Communications will be needed for access to large computer systems, searching directories and catalogs, browsing through data sets, delivery of selected data, and support of mission operations and cooperative research activities. Communications will also be needed for coordination and management of the systems. NASA should aggressively pursue an evolutionary approach to communications networks that would interconnect the various SSDMUs that make up the coordinated, geographically distributed information systems. The first step might be through dial-up lines, followed by higher speed (e.g., 56 kilobits per second (kbps)) links, and in some cases by satellite-rate (megabits per second) connections. The networks should be flexible, allowing for a range of needs, from simple dial-up to high-speed lines, and they should expand and contract as requirements vary. Augmentation of NASA's planned Program Support Communications Network (PSCN) to include support of research and analysis functions would be one method of developing such a system. At present, the plan for the PSCN calls largely for supporting communications between NASA centers. Alternate solutions should also be examined, including use of direct broadcast systems.

NASA should work cooperatively with the space science community in developing useful standards and protocols that can be applied to software development, system interfaces, data formats, directory/catalog formats, and documentation. Standards are key elements to have in place for information systems. On the other hand, standards that are an impediment to research will not be adopted by the space science community. Thus emphasis should be given to standards that can be developed in an evolutionary manner, being first tested and commented

upon by the space science community, before formal adoption.

NASA should play a leading role in developing a capability for scientists to access a distributed directory and catalog system that includes NASA and relevant non-NASA data. A major impediment in space sciences research is the lack of information about what data sets exist, what their characteristics are, and how to obtain calibrated versions of the data. A major step in alleviating this impediment would be construction of directories and catalogs of space science data. Directories and catalogs should be remotely accessible. In the earth sciences, especially, access to data from other federal agencies, from state agencies, and from other governments will be needed to properly address the science issues of the next decade. Thus directories and catalogs of non-NASA data must be developed and made accessible. Data centers should play a major role in developing such capabilities.

2. INTRODUCTION--PURPOSE AND SCOPE OF REPORT

Look not mournfully into the past. It comes not back again. Wisely improve the future.

Henry Wadsworth Longfellow

A Space Science Data Management Unit (SSDMU) was defined in the Space Science Board's Committee on Data Management and Computation (CODMAC) initial deliberations (NRC, 1982) as a group of researchers and support staff who have some data management and computational facilities, and who extract information from space science data. SSDMUs can range in size and scope from small university-based research groups, to teams associated with facility-class space instruments, to large data archive facilities such as the National Space Science Data Center (NSSDC). Research within the disciplines covered by the space sciences is moving into a new era, one in which a large volume of data will be acquired and a number of data sets will be needed within a variety of SSDMU settings to solve the increasingly complex questions that are being addressed. Generally, the data volumes and data uses will grow much faster than the number of researchers examining the data. In addition, the data needed for any given task may not have been collected by any one researcher. Thus ready access to high-quality, well-documented data, together with the ability to handle, process, and store the data are key ingredients for successful management of space sciences data in the 1980s and 1990s. Advances in computation and communications will allow such data management to be done in new, innovative ways. However, as noted in the NRC (1982) report (see Table 2.1), technology is not the main impediment to better data management within any given SSDMU environment. Rather, institutional arrangements, a lack of continuity of management philosophy, a lack of attention to generation and retention of quality data, together with lack of funding, have been the key stumbling blocks.

TABLE 2.1 Brief Summary of CODMAC (NRC, 1982) Findings of Relevance to Data Management

Area of Concern	Common Problem	Recommendations
Data system planning	1. Lack of involvement of science community 2. Adequate funding included in plans 3. Lack of overall planning in general	Adequate planning, funding, and end-end, active involvement of science community in data management
Data	1. Most research groups are processing underfunded in terms of data processing. As a result, they are not able to fully utilize new technologies to alleviate data problems.	
Data distribution	1. Long delays in delivery 2. Users do not know what data exist 3. In some cases, delivered data not properly documented	Much more attention should be paid to providing data in form useful by secondary users. Directories and catalogs are needed.
Data standardization and fidelity	1. Wide variety of formats used 2. Extent of documentation widely variable 3. Insufficient ancillary data	
Software	1. Not documented or portable. Largely developed to meet only immediate goals.	Emphasis on portable software and higher degree of inheritance

In this document we move beyond the general recommendations of the NRC (1982) report and develop guidelines for planning, implementing, and operating SSDMUs, given the expected space science data and the probable user requirements in the 1980s and 1990s. We first summarize the characteristics of the expected data sets and the user requirements that should be levied on systems designed to handle the data. We then consider existing and projected technologies that can be brought to bear on meeting the requirements, and we recommend technology areas that NASA should augment or develop because of the peculiar needs of the space sciences. We then discuss several examples of SSDMU arrangements, including institutional configurations, existing and planned, that involve the space science community, utilize technology in a reasonable manner, and significantly improve the capability to access and analyze well-documented, quality data. From the requirements and the examples we derive guidelines for the future, stressing the roles of data centers, repositories, and sites housing research data sets (active data bases) as part of coordinated, geographically distributed information systems.

3. DATA SETS AND RESEARCH SCENARIOS FOR THE 1980s AND 1990s

You will always underestimate the future.
Charles F. Kettering

3.A. INTRODUCTION

The intent of this chapter is to briefly describe the computation and data management problems that NASA and the space science community will face in the 1980s and 1990s, based on current data volumes, expected rates of data growth, and ways data will be utilized. The challenges produced by data volume and rates of data growth can be clearly delineated. An equally important challenge, however, lies in satisfying the increasing demands researchers will have on data handling and processing functions, particularly on the ability to obtain data from a variety of sources. We explore these demands through examples of research scenarios involving various disciplines and SSDMU environments. The examples are not meant to be inclusive of all possible situations. Rather they serve to help develop an envelope of user needs, together with providing indications of how space scientists will conduct research in the coming decade.

3.B. CURRENT DATA VOLUMES AND PROJECTED RATES OF GROWTH

Tables 3.1 to 3.4 list the quantity of existing digital data, the quantity expected from approved missions in each space science discipline, and estimates of the volumes that will be produced from probable, but not yet approved missions. The projections are necessarily approximate, with uncertainties of perhaps a factor of 2 for missions only in the planning phases. The trends are perhaps best visualized in graphical form. Figure 3.1 is a plot of the cumulative number of bits returned as a function of time for each space science discipline, based

TABLE 3.1 Data Expected From Future Missions in Astronomy and Astrophysics

Mission	Status	Year	Data Expected
Infrared Astronomy Satellite	Completed	1984	10^{11} bits
Space Telescope	Approved	1986	10^{12} bits/yr
Roengten Satellite	Approved	1987	10^{11} bits/yr
Cosmic Background Explorer	Approved	1987	10^{11} bits/yr
Gamma Ray Observatory	Approved	1988	$5x10^{11}$ bits/yr
Extreme Ultraviolet Explorer	Planned	1987	10^{11} bits/yr for 5 years
X-Ray Astrophysics Facility (AXAF)	Planned	1991	$5x10^{11}$ bits/yr
Far Ultraviolet Spectroscopic Explorer	Planned	1990s	$2x10^{13}$ bits/yr
High Throughput Mission (large X-ray collector)	Planned	1990s	$5x10^{12}$ bits/yr

NOTE: Current volume of digital astronomy data is approximately 10^{13} bits, and current volume stored at NSSDC is approximately $3x10^{12}$ bits.

on the data in the tables. Note that the data volume axis is plotted on a logarithmic scale. The trends in growth of space science data can, if averaged over several years, be modeled with exponential functions, with data doubling intervals ranging from 2 to 5 years. The rapid growth and resultant large volumes indicate that the space sciences are moving into an era that will significantly challenge scientists' ability to handle, process, and store data. In addition, based on past trends in funding NASA investigators, it seems improbable that the number of researchers will grow at the same rate as the data volumes. As a consequence, we expect that the ratio of data to researchers will grow rapidly, indicating that enhanced data management and computation procedures are

TABLE 3.2 Data Expected from a Number of Future Missions in the Planetary Sciences

Mission	Status	Encounter Date	Data Expected, bits
Voyager	Ongoing	1986/Uranus 1989/Neptune	4.5×10^{11} 10^{11}
Galileo-Jupiter Orbiter and Probe	Approved	1989	$\sim 10^{13}$
Venus Radar Mapper	Approved	1988	$\sim 10^{13}$
Comet Rendezvous	Planned	1990(?)	$\sim 10^{13}$
Lunar Geoscience Orbiter	Planned	1991(?)	$\sim 10^{13}$
Mars Geoscience Climatology Observer	Approved	1992	$\sim 10^{13}$
Titan Flyby/Probe	Planned	1990s	$\sim 10^{13}$
Saturn Flyby/Probe	Planned	1990s	$\sim 10^{13}$
Mars Aeronomy Orbiter	Planned	?	$\sim 10^{13}$
Mars Probe Network	Planned	?	$\sim 10^{13}$
Venus Atmospheric Probe	Planned	?	$\sim 10^{13}$
Multiple Main-Belt Asteroid Orbiter and Flyby	Planned	?	$\sim 10^{13}$
Saturn Orbiter	Planned	?	$\sim 10^{13}$
Earth-Approaching Asteroid Rendezvous	Planned	?	$\sim 10^{13}$

NOTE: Existing digital data in planetary sciences totals is approximately 10^{13} bits, and current volume stored at NSSDC is approximately 4×10^{11} bits.

TABLE 3.3 Data Expected from a Number of Missions in Solar and Space Physics

Mission	Status	Year	Data Expected, bits/yr
IMP-7,8	Ongoing	Ongoing	2.4×10^{10}
DE-High	Ongoing	Ongoing	3×10^{11}
ISEE	Ongoing	Ongoing	8×10^{11}
Active Magnetospheric Particle Tracer Experiment	Approved	1984	10^{11}
Solar Optical Telescope	Approved	1990	10^{12}
Upper Atmospheric Research Mission	Approved	1990	2.5×10^{12}
International Solar-Terrestrial Physics Project	Recommended	1990s	4×10^{13}

NOTE: Current volume of data is about 10^{13} bits, and current volume at NSSDC is approximately 3×10^{12} bits.

mandatory, even to analyze the same fraction of space science data that are analyzed now.

3.C. ASTRONOMY SCENARIOS

In this section we discuss two astronomy scenarios that illustrate demands on SSDMU data bases in terms of the need for remote access and on-line browse capabilities.

3.C.1. Interdisciplinary Study of the Structure of Galactic Jets

In this study the object is to understand the structure and environment of jets of material emerging from active galaxies. In particular, it is assumed that radio galaxies have been observed that exhibit jets of various forms: e.g., continuous versus knotted; straight

TABLE 3.4 Data Expected From a Number of Missions in the Land, Ocean, and Atmospheric Sciences

Mission	Status	Year	Data Expected
GEOS, G,H	Ongoing	Ongoing	1.5×10^{13} bits/yr
NOAA F-J	Ongoing	Ongoing	10^{13} bits/yr
ERBE	Approved	1984	10^{12} bits/yr
LANDSAT D,D'	Ongoing	Ongoing	10^{14} bits/yr
TOPEX/POSEIDON	Planned	1988	10^{12} bits/yr
Geopotential Research Mission	Planned	1991	10^{12} bits/yr
SIR B,C,D	B=Funded C,D=Planned	1984, TBD	6×10^{14} bits
Shuttle Imaging Spectrometer	Planned	1989	10^{13} bits
Earth Observing System	Planned	1990s	10^{12} bits/day

NOTE: Current volume of Landsat data is approximately 10^{14} bits, while 2×10^{13} bits of other data exist. Current volume at NSSDC is approximately 7×10^{12} bits.

versus kinky; or one-sided versus symmetrical. By detailed comparison of the radio data with image data at other wavelengths, it may be possible to determine jet emission mechanisms (via overall electromagnetic spectrum) and jet kinematics and confinement (via presence of gas, etc.). As a first step, optical data from the Space Telescope (ST) and X-ray data from the Advanced X-ray Astrophysics Facility (AXAF) might be compared with the jet morphology derived from the radio data.

Comparison of radio, optical, and X-ray data for a well-defined set of targets is relatively straightforward. With a limited and well-defined set of targets, catalogs of galaxies observed by ST and AXAF would be consulted. Browsing of summary data sets to ascertain if suitable images exist that include the galaxies of interest would also be highly desirable and would supply more detailed information. If the researcher was not located at the facilities supporting the catalogs and browse files,

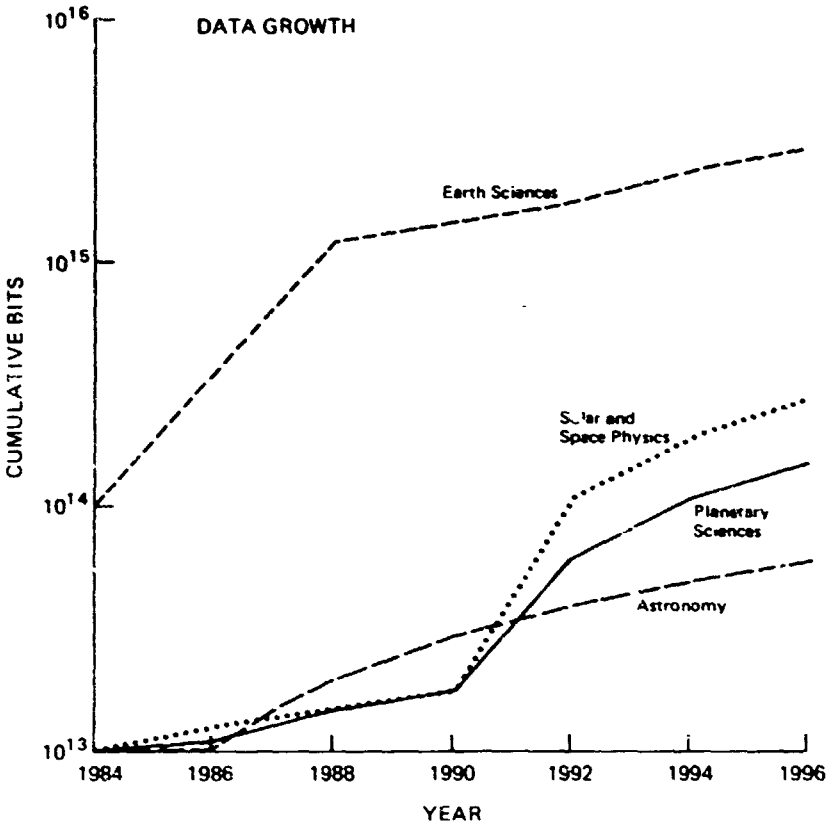


FIGURE 3.1 Projected growth rates for space science data, based on data from tables. Earth orbital missions assumed to last for 5 years, except for operational satellites and the space telescope, which are projected as continuing data producers.

electronic access via a communications link would be highly desirable so that the user could work at his home institution. Or, up-to-date information could be distributed on a regular basis on, for example, floppy disks (attributes) and videodisks (browse data).

After catalog searches, copies of calibrated, digital data in a standard format would probably be requested for use at the researcher's home institution, using a standard format such as the Flexible Image Transport System (FITS)

that has been developed by the astronomy community. It is important to note that FITS would allow the tapes to be read with a minimum of programming effort if used at both the data facility and the researcher's home institution. The ST and AXAF images would then be compared with the radio data to look for correlations using standard data analysis tools such as cross correlation techniques. At some point in the analysis and interpretation, details of the conditions of the various observations would be requested to determine such parameters as true resolution (i.e., Was there excessive smearing due to jitter?), true magnitudes (i.e., true radiometry, including knowing how the calibration performed, and whether there were any unusual background effects), etc. Thus some overall quality measures should be included with the ST and AXAF data, along with information documenting calibration procedures. More subtle effects may even have to be considered and would require accessing the observing logs and user manuals.

Given calibrated data, indications of data quality, and the appropriate analysis tools, the investigator would be able to conduct the study and thus constrain the nature of the jets.

3.C.2. Galaxy Distribution as a Function of Magnitude and Color

In this study galaxies are to be categorized based on observed color and magnitude. The study, although not requiring data from different subdisciplines, is taxing on the SSDMU data management facilities, since it is an involved data base research project. In the study, ST images covering as large an area as possible would be needed. In addition, the images would have to satisfy certain criteria as to exposure, background, and filters used.

We assume that the data reside at the Space Telescope Science Institute (STSCI), which will be a data center for Space Telescope Data, and that the researcher resides at his home institution. Thus, it must be possible to search the data base remotely.

The data base would first be queried for a list of all ST exposures that satisfy the following criteria:

- Galactic latitude of >40 degrees (to avoid the obscuration by dust in our own galaxy).

- Ecliptic latitude of >40 degrees (to avoid zodiacal light).
- Ecliptic longitude of >40 degrees (to avoid sunlight).
- Exposure time >30 minutes. It is assumed that the data base contains the exposure time, and not just total observing time.
- Exposure is a nonproprietary science exposure (i.e., not a calibration exposure).
- Exposure is calibrated.
- Exposure is from the wide field camera, the planetary camera, or the faint object camera (FOC) in f/48 512 x 512 imaging mode.
- If the exposure is a wide field camera or planetary camera image, then the filter must be a set of user-specified modes.
- If the exposure is an FOC image, then the filter must be in a particular set of user-specified modes.

The queries to the data base assume that the user is knowledgeable about both the data base query language and the ST instruments. This assumption may not be true for all users, so a HELP capability for the query system must be provided on-line. The HELP functions should include information on the instrument characteristics, the query procedures, and the variables that can be searched.

The result of the queries would be a list of all frames that satisfy the criteria specified above. The list should appear not only on the user's remote terminal, but in a file as well. Then, the list could be available as input to the query system for further searches by the investigator or by other researchers.

The list of sky areas would then be divided into three classes: those covered by two different filters (from which a V magnitude can be derived), those covered by only the V filter, and those covered by only one filter, which is not the V filter. The latter two lists form the basis of another query. In this query, the researcher relaxes the exposure time requirement to 15 minutes in the hope of uncovering additional exposures that can provide colors for some of the galaxies in the V band exposures or V magnitudes for the other exposures. Any of the non-V band exposures that remain unmatched are discarded.

The researcher should have the capability to peruse the data file header information. Perusal of this information would allow discarding of unsuitable

exposures, e.g., those in which the pointing was not sufficiently stable. A remote data browse capability would be extremely useful, since the capability for a quick look at each exposure allows further narrowing of the list of exposures to be delivered. That capability could also be implemented, for example, with a videodisk player at the investigator's home institution. The player, containing a disk of ST images, could be controlled remotely (i.e., by the STScI) as part of the search procedure.

The researcher would also be interested in other exposures related to those on his list (e.g., special calibration exposures, exposures through other filters, exposures in the same field with a spectrograph), and might perform an additional search of the archives to find related images. A list of related exposures would be kept, but whether any of these exposures would be added to the list of exposures to be retrieved depends on what is found.

Having generated the list of exposures, the user would then request that the data on the list be extracted and delivered to the researcher's home institution. Depending on the manner in which the ST data management facility is set up, especially whether remote processing is supported, the researcher may not be directly involved in the extraction process. For example, if the data sets are based on tape, the appropriate data would be located, copied, and sent to the researcher's home institution. If the archives are on-line (on optical disks, say), copy commands to transfer the data sets electronically to the investigator might be invoked. In addition to the archived data, the user might want the lists resulting from the data base search, and possibly some special purpose analysis software.

For this particular study, a reasonable number of images would be on the order of 100 and each image would be on the order of 10^7 bits, so roughly 10^9 bits would need to be sent to the researcher. If the researcher comes from an institution that is connected to the STScI with high-speed (56 kbps) lines, the data might be scheduled for overnight transmission, requiring just under 5 hours of transmission time. If the researcher does not have access to a network, the data might be sent on tape. Final analysis would probably take place at the researcher's home institution, where the appropriate hardware and software should reside to do the work.

3.D. PLANETARY SCIENCE SCENARIOS

In this section we discuss the Mars Geoscience Climatology Observer (MGCO) Mission, stressing the variety of data to be produced and the need for comparison with Viking observations. The large data volume, the complexity of the data, and the need to conduct timely analyses during the mission present significant computation and data management challenges.

The MGCO spacecraft is scheduled to be put into orbit about Mars in 1992. The nadir-looking MGCO spacecraft would be in a nearly circular polar orbit, making selected measurements of atmospheric and surface features over a 2-year period. The three instruments most likely to be included and that would produce the most data are the visual and infrared mapping spectrometer (VIMS), the gamma-ray spectrometer, and the radar altimeter. In addition, the approximately 50,000 digital images (each image roughly 1000 x 1000 x 8 bits) that make up the Viking Orbiter data set would be used as a base data set, partly because there will not be a high-resolution framing camera for evaluation of surface morphology.

The three MGCO experiments would produce a more or less continuous stream of data throughout the 2-year nominal mission. The VIMS will depart most from this routine operating mode because of the high data rates that need to be generated to cover selected areas in detail. In addition, VIMS will produce some support imaging to be used in conjunction with other MGCO instruments. The gamma-ray spectrometer and the radar altimeter are scheduled to make measurements at a uniform rate throughout the mission in order to build up a global map of the surface radioactivity and elevation. Although the spacecraft data system will likely be configured to acquire a standard set of observations, contingency plans must exist to change to a different observation strategy if the surface becomes obscured by dust storms. Typically, local dust storms spread planet wide with a time scale of about one week. Thus, timely examination of the data is needed if MGCO is to be commanded to monitor dust storm growth.

The best estimate of the total data return is obtained by considering the spacecraft communication rate restrictions and assuming full use of the tape recorder facility. One tape recorder is to be read out once per day. The recorder holds 5×10^8 bits, so that 3.4×10^{11} bits for the 680-day mission would be returned in this manner.

In addition, once per day for about 4.5 hours, a 32-kbps downlink would be available to the imaging spectrometer, for an additional return of about 3×10^{11} bits for the mission. This is surely a low estimate for the total data return because it is likely that more passes than expected at NASA's Deep Space Network will occur, increasing the data return by perhaps a factor 2 to 4. A total of over 10^{12} bits is thus likely to be returned during the nominal mission.

The imaging spectrometer could produce a spectrum of reflected solar radiation in approximately 256 spectral channels between 0.35 and 5.0 μm (micrometers) for at least a 1-km-wide strip along the spacecraft ground track (one spectrum every one-third second) on the sunlit hemisphere. This acquisition sequence would produce about 10^8 spectra during the normal mission. In addition, a number of spectral bands will be chosen to construct full images to provide the geological context for interpreting the full spectral data. In addition, there would be special observations, e.g., full spectral maps of selected areas such as the permanent polar caps. It is estimated that this experiment will contribute about 75 percent of the 10^{12} data bits to be returned.

The 10^{12} bits probably underestimates the useable data to be returned by the imaging spectrometer, for there may be a considerable amount of data compression and intelligent editing on-board the spacecraft. These "unnecessary" bits will be restored during ground processing when analyzing and displaying the data as spectra and images. This procedure may lead to perhaps an order of magnitude increase in the number of data bits to be handled on the ground as compared to those transmitted from the spacecraft.

The gamma-ray spectrometer would have an uncompressed data rate of up to about 2,000 bps and would operate throughout the mission. The data would consist of energy flux spectra that would be integrated, perhaps on-board the spacecraft, to produce higher and higher signal-to-noise spectra for smaller and smaller surface areas as the global integration continues. These gamma-ray data would probably be processed and reprocessed as the mission continues.

The radar altimeter is likely to make specific readings once every 2 seconds for an approximately 2 x 2 km footprint throughout the mission. These data would be converted to distances between the surface and the space-

craft, and then to surface heights above a center of mass as the orbit and spacecraft positions are determined.

The altimetry and gamma-ray data would be processed into maps of topography and elemental concentration and then be made available for comparison and integration with other data sets.

The VIMS multispectral map data would probably be processed as a stack of registered images, and if enough spectral bands were included, spectra could be extracted and combined with the more complete spectra for analysis aimed at mapping mineral chemistry. The reflection spectra would have to be individually analyzed to determine mineralogy. Mineralogy maps would be developed as separate, derived data bases.

A wide variety of science disciplines and communities would wish to use the MGCO data, for studies related to Martian climate, volatile cycles, surface evolution, weather, and polar cap history. All the data sets must be available in a uniform format. The most important results would be obtained by digitally combining and overlaying these varied data sets. For example, the gamma-ray data would yield elemental composition, while the reflection spectra would characterize mineral chemistry; these must be used in conjunction for greatest return.

All the data sets should be registered to an image base map. The Viking Orbiter digital image data set should be available, but at the moment it is not properly processed. The effort needed to decalibrate the Viking data alone is a large one, with 4 Tbits of data being available.

There will be a number of investigator home institutions participating in the mission data processing and postmission data analysis, because of the wide range of science disciplines and measurement techniques. Data manipulation should occur at these institutions, and exchange of data sets and even remote processing will probably be required.

In summary, the major challenge of the MCGO mission is in the data handling. For ground-based analysis, perhaps over 10 Tbits will be involved. A variety of global data sets will be produced, which must be registered to image data from a previous mission (Viking). The MCGO global data sets will probably be produced at several institutions as the mapping mission is under way, and most data sets will require several reprocessing interactions during the mission. The global data sets must be available as resources to a wide variety of scientists at

different locations during and after the mission for comparison and consideration, in order to reap the full scientific benefits of MGCO.

3.E. SOLAR AND SPACE PHYSICS SCENARIOS

In this section we concentrate on how the science community would acquire data from the Global Geospace Study, where large data volumes, a variety of data, and a number of facilities and institutions will be involved.

Several missions have been suggested in the solar and space physics (SSP) area for the late 1980s and 1990s. These missions would differ from previous SSP missions, both because the particles and fields instruments would be much more sophisticated and because there would be much greater emphasis on auroral imaging observations. Not only will the volume of data increase, but to achieve the scientific objectives it will be necessary to study simultaneously data from several spacecraft and from several instruments on each. These missions will place increased demands on SSDMUs charged with handling, processing, and storing the data.

Of all the SSP activities, the Global Geospace Study (GGS), which is part of the six-spacecraft International Solar Terrestrial Physics (ISTP) mission (see Table 3.3), will be the most complex. During the nominal missions (2 to 3 years depending on the satellite), the telemetry stream from ISTP will produce 4×10^{13} bits of data for GGS. About one-third of the data will be imaging data. The GGS archive will contain more than an order of magnitude more magnetosphere data than is currently stored in the magnetosphere and upper atmosphere archives at NSSDC. This mission will set the upper limit on the SSP data system for the 1990s.

Since the goal of the GGS mission is to study the flow of energy and momentum through the solar wind-magnetosphere-ionosphere system, the data system must be one that facilitates the exchange of data from both satellite-borne experiments and ground-based instruments. The GGS will include four spacecraft: three from the United States and one from Japan. In addition, many of the investigators associated with GGS are from Japan and Europe. The data system must support those investigations as well as those in the United States. Clearly, exchange of information and data across federal agencies and international boundaries is key to the success of GGS.

Computation and data management plans for the GGS are relatively mature. We therefore discuss them in terms of how users would acquire and analyze GGS data. Present plans call for the GGS data system to consist of a Central Data Handling Facility (CDHF) plus 26 Remote Investigator Facilities (RIF) at investigators' home institutions. Some of the RIFs will be located at non-NASA sites, including other countries. The RIFs will be linked to the CDHF by 9600 baud communications lines. The CDHF is scheduled to have a master data base that will hold all edited (removed from telemetry stream and placed in instrument format) data plus higher-level data products produced by the investigators at the RIFs. The science repository will consist of edited data, software to process the data, processed data, and a key parameter archive. The key parameter data set would consist of low time resolution data (up to 10 parameters per instrument) that will provide a browse capability so that a user can select the data type and data intervals he needs for his study. The digital key parameter data should be accessible over the 9600 baud communications lines. Key parameter data will also be available on microfiche. The CDHF will probably provide each RIF with optical disks that contain all edited data from all instruments on a given spacecraft plus definitive orbit and altitude data. These data plus the processing software from the CDHF archive will provide the users of each RIF access to all of the data.

In addition to the key parameters and edited data, the CDHF will contain event data. The event data will consist of portions of the data selected because they are of special interest. These data will be thoroughly processed at the RIFs and returned to the CDHF, which will make them available to the community. The data should be available via the communications lines.

Scientific access to the GGS data will be through the RIFs. Users not at an RIF site will communicate with the RIFs either by using dial-up telephone lines or by using a communications network similar to the SPAN network (see Chapter 6). The RIF will provide analysis software and graphics support for its users. Frequently, the science users will browse the key parameter files to select the intervals for study. If the user wishes to use data from one of the interesting special "events" that have been designated by the community for detailed study, he will be able to obtain fully processed event data from the CDHF through the RIF for his study. If he requires

high-resolution data for an interval not in the event data set, he can obtain them from the edited data file stored on optical disks at each RIF. Each principal investigator will provide the CDHF with software to process the data from his instrument. The user will be expected to process the edited data by using this software.

3.F. LAND, OCEAN, AND ATMOSPHERIC SCIENCES SCENARIOS

The land, ocean, and atmospheric sciences in the next decade offer major challenges to computation and data management, since a variety of NASA and non-NASA data, both of spacecraft and ground-based varieties, will be needed to answer many scientific questions. We illustrate the potential complexities with two scenarios, one dealing with vegetation biomass and one with the Earth's radiation budget.

3.F.1. Vegetation Biomass, Productivity Estimation, and Large Area Inventory

In this section, we discuss an earth science research project where diverse data sets are needed to achieve objectives. Even so, the project is of relatively small scope, involving remote sensing data available at present, and not including advanced sensor data that we may be available in the 1990s in orbit (see Butler et al., 1984, and Earth Observing Data System description in Chapter 6). The work is complicated further by the involvement of a number of institutions and the need to transfer data and information between institutions.

The goals of the project would be (1) to develop methods to measure directly, by remote sensing, biomass and net primary production of terrestrial vegetation in a boreal forest setting using these data and (2) to employ satellite images, primarily Landsat and NOAA Advanced Very High Resolution Radiometer (AVHRR) data, in assessing and improving the current, operational accuracy of major accepted sources of continental-scale land cover information. This research would lead to an improved understanding of vegetation characteristics and processes, such as biophysical characteristics (leaf area index, biomass, net primary productivity, canopy tempera-

ture, and albedo), and plant physiological processes (evapotranspiration, photosynthesis, and respiration).

The ability to infer key vegetation characteristics from remotely sensed data is principal to the economy of large-scale research. As a first step, close-range spectral signatures of vegetation, collected from a low-altitude platform, would be correlated with such laboratory measurements as leaf reflectance. These data would then be used as a basis for comparison with higher-altitude measurements from aircraft and spacecraft (Landsat, NOAA AVHRR), where atmospheric conditions attenuate and distort the characteristics of these signatures.

As a second step, manual interpretation and machine classification of Landsat and NOAA AVHRR data would be done to stratify vegetation and other land covers into broad, physiognomic categories (based on vegetation structure) suitable for global comparisons. Aerial photographs, field reconnaissance and other sources would be used for accuracy verification. This approach provides both a comparison for current information sources, and an assessment of the methodology of very large area vegetation mapping.

The data management and processing tasks that would be involved in this project are very complex and would involve diverse data sources and distributed investigators. Preliminary investigations of this type with university and NASA center participation have demonstrated that the complexity limits the rate and efficiency of analysis and magnitude of effort in several ways.

1. The necessity of transferring graphic and tabular data sets between institutions for proofreading, registration, etc., will be a major limiting factor on the speed and efficiency of data analysis. At several stages, forms, listings, or tapes must be mailed between institutions and formats converted. Analysis of some data could be delayed by several months, impeding planning for further work. In addition, considerable human resources could be consumed in essentially nonproductive work.

2. The size and completeness of this study and others like it would be limited by the ability to access and calibrate large data sets. Correlative data--topographic, meteorological, historical, etc.--are crucial in understanding patterns studied. Independent acquisition of such data is impractical, and most existing data bases are not under NASA control. In many

cases the data are very difficult or time-consuming to obtain. Ready access to (or even knowledge of) data from parallel studies in other areas would be very valuable for verification of generality of patterns.

3. Discovering, obtaining, registering, and analyzing remotely sensed data other than those gathered specifically for the project would be of such difficulty that valuable types of data may be unused due to lack of knowledge of their existence or resources for making them useable.

An information systems approach could benefit this project in many ways. Some of the most important areas of support might be in the following:

1. Data input: Direct transmission of field data between field sites and processing centers at NASA centers and universities could cut processing time by an order of magnitude (from months to a few days). Entry or conversion of correlative data (topographic, soils, climatic) to acceptable form would add to the potential of the project.

2. Preprocessing: Registration (band-to-band and sensor-to-sensor) and common formatting of sensor data (from Landsat Thematic Mapper (TM) or Multispectral Scanner (MSS) data, AVHRR, scatterometer, radiometer, imaging spectrometer data, etc.) would be of great value and high priority. Efficiency of work would be vastly improved if this could be accomplished within a few weeks of data acquisition. Also of value (but less important) would be the capacity to digitize photographs with interactive input from remote principal investigators (PIs).

3. Analysis: Efficiency of analysis could be increased if real-time interaction between centers and remote investigators in the analysis process were possible.

4. Storage and cataloging: A directory, with documentation of correlative data sets held within NASA, and elsewhere, would be of great value and is of high priority.

5. Distribution and networking: Access to data sets referred to in item 4 above, and ability to overlay them digitally in common format would be a high priority. Data, besides being in compatible format, must carry documentation of quality and type. The time scale for such access should be on the order of a few days.

Networking of computers and availability of peripherals at NASA centers to provide access by remote investigators should be valuable.

3.F.2. Studies of the Earth's Radiation Budget in the Earth Climate Program

In this scenario we discuss the research methodology to be used as part of the Earth's Radiation Budget Experiment and thereby illustrate challenges imposed by data obtained from a number of spacecraft, housed in a variety of locations, and under various agency controls.

Models to predict the future climate of the Earth must include the role of changes in the Earth's radiation budget. The budget is dependent on the relative magnitudes of solar radiation absorbed by the atmosphere/surface system and that reflected and re-emitted to space. These radiative quantities are associated with the driving mechanisms of the general circulation and involve a complicated interaction between the external radiation from the sun, and the interaction of radiation with the clouds, oceans, surfaces, and the possibly changing composition of the atmosphere. The observational data base for an investigation of the radiation budget would involve accurate measurements of the external solar radiation, together with the radiation reflected and emitted from the Earth in the UV to the far infrared parts of the spectrum. Radiative processes in the atmosphere also have a strong diurnal signature. To account for these variables, the Earth Radiation Budget Experiment (ERBE), will involve observations from sun-synchronous polar orbiting weather satellites under NOAA's control, a drifting NASA satellite, with supplementary measurements made from the set of geostationary satellites to provide a more complete set of measurements to reduce the possible temporal and spatial sampling errors.

In one research scenario, an investigator would have access to maps of the global radiation budget parameters (outgoing longwave flux, albedo, absorbed solar radiation), for all the available satellite observations for each month of the year, and on a spatial scale compatible with numerical general circulation models. The data on the radiation flux from the sun would also be available. Periods with anomalous measurements would be investigated in a more detailed manner through access to the daily measurements and those from the individual spacecraft.

Researchers will need to be able to access ERBE data in ways that efficiently summarize the spatial and the temporal data obtained during a particular period of the mission. With such access, the presence of any unusual climatic or budget features may be investigated in conjunction with changes due to solar forcing, thereby increasing the understanding of solar effects on weather and climate.

In another research scenario, involving the effects of clouds in the radiation budget, researchers would require access to the data base housed as part of the International Satellite Cloud Climatology Project (ISCCP). The correlative measurements of the geostationary satellites should also be available, perhaps by facilitating access to the data bases held by the operational satellite agencies (e.g., NOAA in the United States) supporting the weather forecasting programs.

3.G. SUMMARY OF COMPUTATION AND DATA MANAGEMENT TRENDS

The large volume and rapid growth of space science data, doubling every few years, is clearly one way of gauging the extent of the data management and processing problems that need to be dealt with in planning, implementing, and operating SSDMUs in the 1980s and 1990s. In addition, as illustrated by the example research scenarios, data from a variety of instruments, missions, and sources will be needed to conduct much space science research during this era. Data handling and processing needs in the space sciences will probably grow by more than a linear proportionality with respect to the data growth. In addition, some of the data, especially in solar and space physics and in the earth sciences, will come from agencies other than NASA and some must be collected in the field. Challenges clearly await in terms of having the ability to search a data set, to access the data, and to process the data, in addition to problems related to long-term data curation. Major challenges await in developing the management structure that will facilitate efficient, timely access to data from various NASA missions, data from non-NASA sources and, in some cases, data from other countries. Such a system must access a geographically distributed set of data bases.

4. USER REQUIREMENTS FOR SPACE SCIENCE COMPUTATION AND DATA MANAGEMENT SYSTEMS

The buck stops here.

Harry Truman

4.A. INTRODUCTION

We now draw on information presented in the last chapter about space science data volume growth rates, and the probable uses of the data, in order to extract sets of requirements that should be levied on computation and data management systems. We also begin to use a number of terms, some of which appear for the first time in this report. The terms fall into two basic categories: definitions of general levels of data processing and data types (Table 4.1) and our definitions of data bases (Table 4.2). It is hoped that standard definitions of these terms, if followed by the community, will alleviate some of the confusion associated with the semantics of data management and computation as applied to the space sciences.

In the following sections we first consider a general model of data flow in the space sciences, distinguishing between archives, repositories, and active data bases. We then discuss specific issues related to contents of data sets, management of data sets, data directories and catalogs, and we end with requirements on data search, access, and process functions. As noted earlier in this report, we stress those aspects related to computation and data management once the data are on the ground. This stress should not be construed as an indication that mission operations activities are not important or without associated issues. The realm of mission operations, including instrument control and the role of on-board processing, will be dealt with in a later report.

TABLE 4.1 Definitions of Space Science Data Levels and Types

Data Level or Type	Definition	Utility
1. Raw data	Telemetry data with data embedded	Little use to most of science community, except for radio sciences
2. Edited data	Corrected for telemetry errors and split or decommutated into a data set for a given instrument. Sometimes called Experimental Data Record. Data are also tagged with time and location of acquisition. Corresponds to NASA Level 0 data.	Wide use, especially for researchers familiar with instrumentation
3. Calibrated data	Edited data that are still in units produced by instrument but that have been corrected so that values are expressed in or are proportional to some physical unit such as radiance. No resampling, so edited data can be reconstructed. NASA Level 1A.	
4. Resampled data	Data that have been resampled in the time or space domains in such a way that the original edited data cannot be reconstructed. Could be calibrated in addition	Wide use, especially for secondary users

TABLE 4.1 (continued)

Data Level or Type	Definition	Utility
	to being resampled. NASA Level 1B.	
5. Derived data	Derived results, as maps, reports, graphs, etc. NASA Levels 2 through 5.	General way in which information is transferred
6. Ancillary data	Nonscience data needed to generate calibrated or resampled data sets. Consists of instrument gains, offsets; pointing information for scan platforms, etc.	Needed to be able to convert edited data to calibrated, resampled, or derived data sets
7. Correlative data	Other science data needed to interpret spaceborne data sets. May include ground-based data observations such as soil type or ocean buoy measurements of wind drift.	Crucial data in many cases to provide ground truth calibration for spaceborne data
8. User description	Description of why the data were acquired, any peculiarities associated with the data sets, and enough documentation to allow secondary user to extract information from the data.	Important aspect associated with the data that will be even more important for facility-class instruments and for secondary users of data

NOTE: We define a secondary user as a researcher not involved with instrumentation design, development, or data acquisition. A secondary user would normally go to a data archive to obtain the required data set.

TABLE 4.2 Definitions of Selected Data Management Terms

Term	Definition	Utility
Data archive	Long-lived data base, maintained as a national resource at a data center	Provides long-term access to data
Data repository	Short-term data base that serves as way station or clearing-house for data	Variety of uses, such as a mission data base to support operations and compilation of initial results
Active data base	Subsets of data or complete data bases that are being actively used by science community	The data to use in doing scientific research
Data base	The actual data, either part of an archive, repository, or active data base	Needed to do research
Data catalog	Descriptions of data base in sufficient detail to retrieve subsets of data. Searchable by data fields or attributes, down to some level of granularity.	The way to look through a data base
Data directory	Top-level index containing information about location, ownership, contents of data	The first step to determining what types of data exist for given time period, location, etc.

4.B. STYLES OF DATA MANAGEMENT--REPOSITORIES, ACTIVE DATA BASES, AND ARCHIVES

It is useful to consider a general model of data flow from receipt of the data from the spacecraft, to a mission data system, and eventually to data bases that can be accessed by the space science community. With this flow model, we can distinguish three different styles of SSDMUs:

1. Repositories, which are facilities that are temporary buffers for new data, usually existing only as long as the mission producing the data. The data are distributed to investigators associated with the mission for analyses related to mission operations and first science results. Or, the data are processed centrally and accessed by investigators.
2. Active data base sites, which house data actively used in ongoing research. Active data bases generally outlast a given mission and are maintained as long as the science requirements and funding permit. We envision active data bases as generally being under the direct control of and housed with the science community, in contrast to mission repositories, or the next data set type, an archive.
3. Archives, which consist of long-lived collections of science, operational and related ancillary data, located at a data center, and supported with adequate cataloging, protection, and distribution functions. Archives are stable data bases that ensure long-term access to the data by the general space science community.

It is important to note that the boundaries between the three types of SSDMUs sets can overlap. In some cases, the three styles can be supported by one SSDMU, depending on both management considerations and the technology available. On the other hand, if the operations requirements conflict with the science needs, it may be necessary to implement a mission data repository in a separate SSDMU from an active data base site. Or, as is the case even today, if the scientific community capable of maintaining active data bases is geographically dispersed, a data center supporting a centralized data archive could be physically separate from active data base sites that contain subsets of the data that are topics of ongoing research. Some of these concerns can be illustrated in Figure 4.1, which shows a simplified

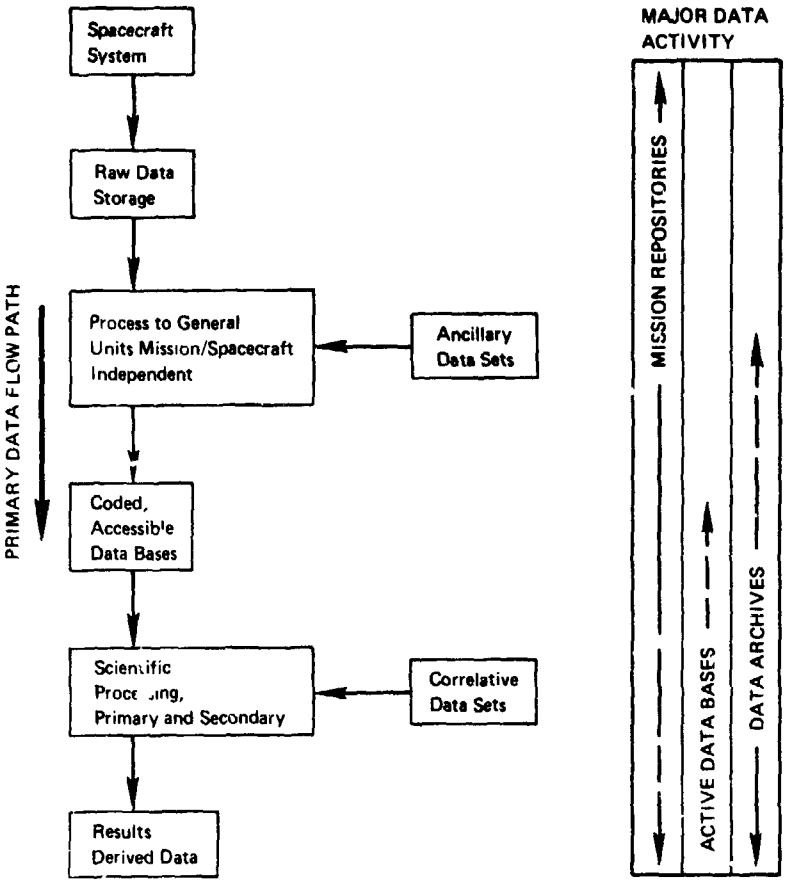


FIGURE 4.1 Conceptual model of data flow and activity in the space sciences.

data flow model, with data flow stages of particular relevance to repository, active data base, and archival data types. We will use the term SSDMUs in the rest of the chapter to refer to a collective system that meets all the requirements, whether physically one entity or three (or more in the case of multiple active data bases).

4.C. DATA SET CONTENTS

A space science data base in a repository, active data base site, or an archive, should contain a minimum of three categories of data: basic science data in various forms; ancillary data, needed at some level to interpret the science data; and the basic software tools needed to access the data, as well as to perform basic analyses. Additional desirable categories of data include mission planning data, derived data, correlative data, and technical mission specification data. The data bases may also be classified as public domain data, temporarily proprietary data, and nonpublic data.

4.C.1. Science Data

Basic science data must include at least two types: edited and calibrated. Edited data, or "Experiment Data Records," should include all the science information generated by the experiment. Effects of multiplexing, packetization, tape recorder playback, error-correcting routines, or other telemetry transmission/capture processes should be removed, but no information should be lost. Retention of useful edited data in most cases can be justified since the cost of storage is far less than the cost of reprocessing or of re-acquiring the data. Data should also be "edited" or structured to a level suitable for indexing in a catalog and relating to a planned observation (e.g., sorted by time, instrument, observer, etc.).

Calibrated data have instrument signatures removed as far as possible. The data have been converted into values that are in proportion to standard physical units. These corrections often involve temporal or environmental dependence, as well as algorithmic dependence. Thus the conditions of the calibration--software utilized, parameters of the instrument, etc.--must be archived with the data. The purpose of archiving data is to establish a "best-effort" standard data set that most investigators, especially nonexperts in the experiment or discipline, can utilize. It has to be expected that individual researchers, for special purposes, may apply unique calibrations, starting with the edited data. Thus access to edited data is needed for secondary users, but need not be as convenient as access to the more commonly requested calibrated data. In many cases, calibrated and

resampled data should be stored, along with full descriptions of the processing steps involved in reduction. The resampled data are often of the widest interest to secondary users because they provide summaries of the trends or patterns in the data set.

4.C.2. Ancillary Data

Ancillary data include those data that are necessary to calibrate and analyze the basic science data. Obvious examples include spacecraft and instrument housekeeping (engineering) data, orbit/ephemeris/attitude data, instrument transformations and alignments, timing data, and possibly environmental models. The need for calibration data should be especially noted. The goal is to have available in a repository, active data base, or archive all data needed to fully analyze any given piece of science data. Engineering data must be accessible both in a fashion suitable for science data processing and interpretation, and also in repositories for mission-operations-related trend analyses and contingency analyses. For this latter category, command histories, schedules, and similar data that may be necessary to reconstruct the configuration of a science instrument or the spacecraft itself when the science data were taken, may also be needed. The attitude data referred to above may also reasonably include such items as star catalogs, if necessary to reconstruct pointing data.

4.C.3. Software

Software made accessible with the data becomes essential as data bases (active data bases, repositories, and archives) become ever more multidisciplinary and interdisciplinary and are maintained for long periods of time. The utility of maintaining a set of calibration software is obvious, given the fact that calibration procedures and parameters change with time. Basic analysis software which is often instrument or discipline dependent must also be available for the nonexpert user of the data base. This software should, in principle, be developed (or coordinated) by the same SSDMU that is responsible for maintaining the data base, with direct scientific involvement, according to reasonably state-of-the-art software principles, and should encourage the use

of higher level languages and a reasonable degree of machine independence. The capability should exist to add user-developed software of a more advanced nature as it becomes available. This software base is a key ingredient in efficient utilization of the data by an on-site investigator, by an external investigator performing remote processing, and by an external investigator who requests both data and software for home institution processing.

4.C.4. Mission Planning Data

Mission planning data include information that went into designing and implementing a given observation. To first order, this includes the basic attributes necessary to identify the data, including observer, target, scientific program, instrument, etc. Mission planning data are useful within SSDMU repository environments for the mission planning and scheduling processes, and in all three environments for potential users who need to know what data may be available for research. Mission planning data may be contained in the catalog of the data base and not in the data base itself. However, other useful entries, such as proposal and user records, scientific justification, planning constraints, instrument configuration specifications, etc., may also be desirable, either in the data base or in the catalog.

In some cases (e.g., X-ray astronomy--Einstein Observatory), a single catalog may be used to track science proposals, observation scheduling milestones, scheduling and timeline data, data acquisition and processing milestones, data location, and certain reduced results. In others, the structure of the data or the mode of acquisition may dictate separate catalogs. The details and exact contents of these catalogs will, of course, be discipline dependent.

4.C.5. Derived Data

Derived or reduced scientific data are not always suitable for archiving, due to the varying desires of researchers, the varying quality of the analyses, and the potential bulk of data. However, provisions for adding analyzed data to an SSDMU data system are a desirable feature of a data management system. Obvious examples include special calibrations or reduction procedures that

produce particularly useful data sets, such as summary maps or plots. The processing could also involve, for example, sophisticated resampling tasks that would be difficult for small research groups to accomplish. This capability becomes particularly important in active data base and archive environments.

In an archive environment, a system that houses derived data can also evolve into an "electronic library" for a specific discipline. As data become more and more sophisticated (e.g., multidimensional arrays, deep images), normal publication media (e.g., journals) become limited in their data presentation capabilities (e.g., tables and image reproductions). One can extrapolate the "electronic library" of derived data to include publications, which include by reference a reduced image in a controlled data base, and then to include an actual publication within an archive, with perhaps only abstracts circulated in print.

Perhaps a less fanciful rationale for retaining derived data involves those projects that have large numbers of data sets, large consortia of investigators, and a major commitment of a given discipline's observational resources. Certain reductions may be performed routinely and then archived for the benefit of all as derived data sets. In astronomy, one can, for example, perform source detection or image classification, whereby catalogs of astronomical objects can be produced and archived.

4.C.6. Correlative Data

The question of correlative data--other scientific data that may be used in analyzing and interpreting the data--is discipline dependent. As noted in Chapter 3, in some disciplines, particularly the earth sciences, access to the correlative data sets, many of which are not under NASA control, is essential. For other disciplines, such as astrophysics, use of correlative data has typically been simplistic--source catalogs or overlays for identification of new sources. However, as useable data base management systems proliferate, there will be multiple sets of comparable data, e.g., data taken at different wavelengths. As discussed in the last chapter, the scientific utility of simultaneous access to many data sets (e.g., for spatial correlations of different "color" images, or time variability studies) is obvious. With adequate arrangements and appropriate technologies, these

correlative data bases can be treated separately and accessed as needed from geographically distributed active data bases, repositories, or archival sites.

4.C.7. Technical Mission Specification Data

Of particular importance to data repositories, technical mission specification data involve such documentation as mission requirements, hardware and software specifications and requirements, and design data. Instrument descriptions--experimenter notebooks--which contain instrument responses and describe how an instrument operates, should be included. This documentation is often necessary to interpret data, and is just as often voluminous and difficult or impossible to track from the inception of a project through the utilization of archived data years or decades later. With the growing acceptance of word-processor-generated documentation, and graphics software standards, it should be feasible to add instrument descriptions to the data repository and to later transfer the information to active data base and archive sites. This requirement, with standardized procedures, should be considered seriously for use in future missions.

4.D. MANAGEMENT OF DATA SETS

It is axiomatic that an active data base should be under the control of an SSDMU where the data are actively used in research. The management structure may vary depending on the type and scope of the SSDMU and the data bases: A PI institution for an active data base for a given experiment; an "Institute" for a facility-class mission that colocates active data bases and archives; or some other site or sites for aggregate active data bases composed of a variety of data. The main driver is that the active data base should be under the direct control of active researchers, to help ensure scientific utility, with support from a professional staff. The SSDMU housing the actual data base should also be responsible for certifying the validity (or quality) of the contents of the data base. The site must therefore have the appropriate scientific reputation, as well as an adequate level of support.

There must also be provision for maintaining the long-term integrity of archives of data. For data

located at PI institutions, or other active data base sites, provisions must exist to tranafer the data to a NASA center or other suitable national facility if and when the active data base sites become "less active." In turn, the archived data should be reviewed periodically to determine utility and need for data retention and/or continued archiving.

It is important to note that an SSDMU acting as an archive for a given discipline may be geographically distributed. There could, for example, be multiple SSDMU sites, each with an archive from a different mission, with different data types or different operational requirements. Appropriate catalogs, data base technology, and managerial attention, make the physical location of the data unimportant to a remote user. However, until technology is appropriately implemented, this should not encourage the splitting of different portions of the same archive (e.g., science data and engineering data), unless necessary for operational or other reasons.

4.E. SECURITY AND INTEGRITY OF DATA SETS

We assume from the outset that we are dealing with space science data, and that there are no undue requirements for security. There are then only two high-level requirements: preserving the integrity of the data, and preserving the proprietary nature of an observer's data. The requirements related to access and charging are perhaps better left to a discussion of policies but will be briefly mentioned below.

4.E.1 Integrity

The utility of a data base depends to a large extent on the standards applied to data--essentially quality control. It is important not only that the tools to fully interpret the data be present but also that the processing applied to any calibrated or derived data be very well defined. Any data going into a repository, active data base, or an archive should be subject to at least minimal quality control, be it visual inspection, limited computer processing, or both. Additionally, any decisions for reprocessing and/or recalibrating data must be taken carefully. One of the reasons for having active bases in addition to archives is to supply this quality

control. A significant management responsibility for any SSDMU will rest with the individual who must approve additions and modifications to the data base, especially if that active data base is to become part of an archive.

4.E.2. Preservation

Data archives will likely be essentially "write-once" facilities; e.g., they will consist of data sets that will not be modified as frequently as data in active data base sites. It is assumed that some combination of levels of password protection and protection against remote archive updating will be adequate to safeguard the on-line archive, as long as an off-line, physically separate duplicate copy of the archive is maintained as well. It is important that "security" measures do not significantly affect archive utility.

4.E.3. Proprietary Data

Most missions, whether consisting of PI or facility-class experiments, allow an observer to have sole access to his or her data in the repository for a given amount of time. Thus it must be possible to preserve proprietary rights for the required time. Similar considerations will apply to proposal data, and possibly to certain software. Again, it is assumed that password protection is adequate, although encryption of data might be considered desirable in some circumstances. Any data that is truly secret is probably inapp opriate for a scientific data base.

4.E.4. Operational Security

It is probable that portions of a data base in a repository may be needed on-line in a mission operations environment. These subsets include planning and scheduling data, command groups, guide star data, etc. These data must be specially safeguarded. Furthermore, operational needs may also drive requirements for redundancy and reliability in the data base management system. In order to balance operational needs with remote scientific access, it may be necessary to keep

operational data separately and/or copy it to an on-line system as appropriate.

4.E.5. Remote Access Management

The need for wide access to data sets must be balanced with the allocation of limited resources and the need for accounting. At the current time, NASA typically funds individual observers to carry out data analysis. In addition, NASA has the responsibility to support data dissemination to the public (i.e., the primary NSSDC function). As remote access (and remote computing) become more powerful, the distinction between the various types of users and modes of utilization will decrease. Moreover, sophisticated catalog searches, browsing, and data selection and transmission could use a substantial level of computer resources.

It is clear that at one extreme, basic inquiries and data requests must be supported on a level of effort basis, while at the other, sophisticated access and remote processing must be made available to the funded users. Detailed policies in this area will obviously depend on relative levels of funding and the cost-effectiveness of various technologies. We return to this topic in Chapter 6.

4.F. DATA CATALOGS AND SEARCH FUNCTIONS

4.F.1. Directory

A data directory (see Table 4.2) satisfies the need to let potential users know about the existence of a data set. There should be directories that document the existence of important space science data sets. This function is especially important in an archival SSDMU environment. These directories should contain high-level descriptions of the data set contents, including such information as to types of data, sizes of data bases, and time coverages. Detailed instructions on how to access the detailed data base catalogs are also needed. Whether or not the directories are centralized at one SSDMU or distributed among a number of them does not matter, although the size of the directories should be quite small. The important points are that there must be a "central directory" that is well publicized, and the

means to interrogate it must be very simple. In some disciplines, particularly solar and space physics and earth sciences, access to relevant non-NASA data would be greatly facilitated if users could examine directories that include these data and then be directed to the appropriate catalogs. Transparency to the user is important, so that a minimum of query languages need be learned.

4.F.2. Data Base Catalogs

There should be sophisticated catalogs of the individual data base contents for repositories, active data bases, and especially for archival data bases. These data bases must have basic "smart" capabilities for browsing: attribute searches, attribute relations, and nonexpert friendliness as well as expert efficiency. The capability for user-specific processing is also desirable. The attributes used in the catalog for a given data base will be discipline dependent in many cases; it is necessary for the user community to define the appropriate attributes as well as to specify the required granularity of the catalog. Whether or not these catalogs are totally distributed or are redundantly kept at a central SSDMU location is again not a major issue. Science control of the catalog is the important point.

For a sufficiently cohesive discipline domain, it may even be possible to use a natural language artificial intelligence type of query system that could translate requests at some modest level and deposit the user at the proper level in a structured query system. However, any system must be sensitive to the beginning user as well as the intermediate or very sophisticated user. The user must be able to set the support level of the system to the capabilities he feels he possesses at the time of any query session.

Search capability in a rather large data base can be very taxing on an SSDMU's computational capabilities. For example, in a large library system, the card catalog usually is organized along three specific lines: subjects, authors, and titles. A query that asked this system to identify and locate books on planetary satellite surfaces, or texts authored by A.G.W. Cameron, or the location of "The Physics of Planetary Interiors" would probably be successful. However, if one were to ask this same system to find all the works written by

English meteorologists in the period from 1923 to 1932, it is not likely that any response would be forthcoming without enormous effort, because the data base is not organized to handle this kind of query efficiently. Therefore it is most important to (1) design the catalog/query system to be able to handle efficiently the most common types of requests in that discipline, and (2) supply enough computer resources to perform this search for that discipline community.

Another aspect of this issue deals with the need to support catalog queries along new and "nonclassical" lines. Since there are a near-infinity of ways to organize the data base, many reformulations are likely over its lifetime. It must be possible to add new relations relatively easily, without major disruption.

4.F.3. Remote Access

There is a need for remote access to the directories and data catalogs, especially in an archival SSDMU environment. There should be no need to physically visit an archive location to determine whether a data set exists. For the purposes described above, as well as for data requests, a normal low-rate modem (300 to 9600 baud) is adequate. However, for data browsing (see below) there will be communications/remote processing/data distribution trade-offs.

4.G. DATA BROWSING, ACCESSING, AND PROCESSING FUNCTIONS

4.G.1. Browsing

The browse capability involves, at a minimum, interrogation of the catalog via attribute searches as discussed above. However, data utility is more readily established by inspection of the data proper. Depending on the discipline, the type of data, and the type of study, there is a vast spectrum of types of browsing. For low-volume data (e.g., several bytes per measurement), entire data sets may be scanned. For large data volumes (high resolution spectra, images), different strategies are needed. Single frames may be selected from a large set to determine feasibility of a study. Depending on the available communication link, certain data may be extracted and transmitted (e.g., a low-resolution or

low-dynamic-range image). Such browse-level data may also be made available on widely disseminated media (e.g., video disk) for local inspection via inexpensive computer and/or image display systems. This latter type of approach is feasible with current technology. On the other hand, we can also imagine a requirement to interactively browse through data to help decide what new set of observations to acquire during a mission. This repository-style browsing may not be consistent with the long lead time needed for distribution of video disks. Electronic access may be required.

All of the above require the existence of certain standards and protocols for directory, catalog, and data access. These will likely be discipline dependent and thus should be established with the involvement of the science community, hopefully following more generally established guidelines.

4.G.2. Data Accessing

After establishing the utility of certain data, it may be necessary to obtain a subset or even the full set of data meeting the user's needs. The data request mode again should depend on the type and amount of data, and the communications capability. A data request may involve direct transmission of a data record, subsequent mailing of a tape, floppy disk or optical disk, or subsequent transmission of the data via a wideband link. These will again involve trade-offs between speed of response and cost. This is especially true when technology allows large numbers of data to be kept on-line.

All of the above discussion on access implies the existence of communications capability. Minimum requirements for efficient browsing, and remote processing range from 300 to 9600 baud, while large-volume data transmission probably requires access to at least 56-kbps communications. Transfer of array-oriented data could involve megabit capability if near-real time access is required. Since most requirements do not involve continuous communications, it appears highly desirable to establish some wideband shared communications network, joining the appropriate space science data bases with each other as well as with their user communities. It is important, however, to keep cost and timeliness requirements in mind when discussing electronic communications. Any such network should not be implemented at the expense

of data analysis and basic research in a given discipline. In addition, the most efficient means of sending data may very well be to distribute high-volume data sets widely with such technologies as high-density magnetic or optical media. The main use for wideband electronic communications may be to support rapid looks at data from repositories to support mission operations.

4.G.3. Data Processing

Once the data are acquired, the complex step of data reduction and science processing begins. A recurring issue within NASA and its research community involves adequate support for data analysis. Certain functions clearly should be supported in an SSDMU that has in its charter the housing of an archive, repository, or active data base: these include, at a minimum, simple directory and catalog queries, and requests for small amounts of data. However, support for decalibrating and analyzing data, whether at the data base site or at a user's home institution, must clearly exist if the system is to have real value.

Allowing an archive or an active data base to be established at an SSDMU implies that some funding for local data analysis will exist. This support could be increased to cover "approved" outside users. This will be the case for the Space Telescope Science Institute, which is expected to supply support to both general observers and archival researchers at the STScI and at their home institutions. It is likely that larger SSDMUs will develop software for data reduction and analysis. Care should be given to develop the software in higher level languages and in reasonably machine-independent form. Then, the software can be distributed to smaller groups, thereby easing software development costs and enhancing communication and data transfer among the science community.

Remote processing alleviates some (but by no means all) of the requirements for large-scale data distribution and redundant computer facilities. Data processing tasks that require large numbers of data and/or special-purpose hardware (e.g., large array processors, supercomputers, dedicated algorithm processors) may be more efficiently done at a central site than at a researcher's home institution. In addition, there will likely always be some researchers without access to adequate computer

facilities. The capability for remote processing should be inherent at any data base site housing data available to the research community; the technical details and management policies will depend on the discipline and the resources available.

In summary, adequate computational capabilities should exist to support directory and catalog searches, data browsing and accessing, and data processing. These capabilities should meet the needs of missions and research and analysis programs and the needs associated with long-term data curation.

5. TECHNOLOGY TRENDS AND ISSUES RELEVANT TO SPACE SCIENCE DATA MANAGEMENT UNITS

For I dipp'd into the future, far as human eye could see . . .

Alfred, Lord Tennyson

5.A. INTRODUCTION

Based on the projections made in Chapter 3 regarding data growth, together with the developing complexity of user needs expressed in Chapters 3 and 4, we see a rapid escalation of demands being placed on software and hardware technology. An increase in multidisciplinary and interdisciplinary studies that require that data from more than one source be combined places particularly stringent demands on technology, since the data are likely to be geographically distributed. The cost of identifying, bringing together, and jointly analyzing data from more than one instrument or spacecraft or from more than one archive is considerably greater than processing data from a single source. The increases in communications, processing, and storage needs due to such complexities are hard to predict. We use an envelope of demands derived from the research scenarios discussed in Chapter 3, together with basic concepts from information theory, to project these increases.

We first discuss current capabilities for technologies of interest to SSDMUs and make projections of hardware and software improvements. Later we consider how the demands made by increases in data volume and user expectations, documented previously, match the capability increases we expect from technology. Finally, we identify problems and bottlenecks, as well as technology opportunities, and conclude with recommendations for future technology investment by NASA.

We use the term "demands" here in a sense that is close to the term "requirements" but with the proviso that we cannot be sure whether all demands placed on technology can be met.

When we make estimates about future scientific computation and data management capabilities, we consider only those we can expect to be available to the space scientific community. This includes technologies that, with a high level of confidence, will be in widespread commercial use or can be easily adapted from commercial products. We do not see that space science requirements will be driving the progress of technology in the computer hardware and software areas to a major extent. Exceptions where opportunities exist for NASA to provide developments to meet its unique needs will be noted.

We address the question of the cost of new computing technology indirectly. We assume, for purposes of projecting increases in available technology, that funding for data management and computation to support SSDMUs, including individual research projects, academic laboratories, and large centralized computing organizations, will remain approximately constant, scaled for inflation. The constant funding will enable the researchers, working in a variety of institutions, to buy hardware of significantly increasing capability in the future, but it is not clear if an era of constant funding for computation and data management will suffice to meet the computation and data management requirements discussed in this report. Thus we pay particular attention to whether user demands can be met in an environment of constant funding.

5.B. EXISTING AND PROJECTED HARDWARE CAPABILITIES

5.B.1. General Technology

The rapid progress of computer technology is a well-established and well-documented fact. Computer systems are improving, and we can expect this development to continue into the foreseeable future. In this section we will project technological capabilities in five areas of computer hardware technology: (1) processing speed, (2) input-output bandwidth, (3) storage volume, (4) communication speed, and (5) data display and presentation.

A balance of these capabilities must exist at the various system categories that we project to be typical in the years to come if the demands for data search, access, and process functions described in Chapter 4 are to be met. Our projections depend on expectations for commercially available technology rather than the capabilities of raw computer chips or novel specialized architec-

tures. Our projections also include factors that account for the expected net effectiveness of these technologies. Specific assumptions made will be indicated.

The next three subsections deal with three categories of processing systems and their capabilities: work stations, local multiuser systems, and high-speed scientific processors. These three system categories are usually associated with certain categories of SSDMUs.

5.B.2. Single-User Work Stations

Single-user work stations are a relatively new development in computer technology. Work stations are distinguished by being oriented toward highly interactive local use, typically by a single user at a time. They have become essential tools for mechanical and electronic design activities. Only in a few cases are they currently being used to process data from spacecraft observations. The work stations currently found in use in the space sciences have typically been assembled by research groups having greater than average computational and engineering capability in-house.

We see a great level of commercial activity in making these work stations more broadly available in the near future. This activity is being driven by the cost reductions being made possible through Very Large Scale Integration (VLSI). Table 5.1 shows our projection, and Figure 5.1 shows the results in graphical form. Note that the rate of growth of work station processing capability can be modeled with a doubling interval of 2 to 3 years. Since this technology is not yet very mature, we see a rapid increase in performance, slowing somewhat after 1995, but still proceeding at a rate greater than for general scientific computing hardware. A number of commercial products, for instance: the Apollo, PERC, SUN systems, MICROVAX, etc., are now being marketed. Such work stations typically use modern 16- or 32-bit VLSI processors and are delivered with a fair amount of simple but powerful software. Included in the software are systems that permit user interaction at a level that requires little detailed programming knowledge. There are often advanced graphic display capabilities for rapid hypothesis formation and analysis, and access to computer networks in order to share both data and software.

TABLE 5.1 Projected Advances in Computational Capabilities Assuming Constant Cost

Processor Type	Typical Current Cost, $1,000	Performance, millions of operations per second 1983	1986	1995
Work station (68000)	25-50	1	6	50 (integer)
		0.05	1	15 (floating point)
Multiuser (VAX)	100-300	1	4	15 (integer)
		0.8[a]	3.6	12 (floating point)
Scientific processor	500-5000	100	300	10,000 (integer)
		50	150	5,000 (floating point)
		2 Mbps	10 Mbps	100 Mbps (I/O rate)

[a]If equipped with optional floating-point hardware.

NOTE:

- We assume that for other than floating-point arithmetic, four work station instructions are used to perform the equivalent of one scientific processor instruction.
- We assume that for other than floating-point arithmetic, two VAX-type instructions are used to perform the equivalent of one scientific processor instruction.
- The validity of these ratios depends greatly on the type of computation being performed.

An important feature of these systems is that they have low support staff requirements. Work stations are largely operated by the researchers themselves. Maintenance and programming help is obtained only as needed. Much of the high degree of effectiveness of these systems is thus due to the favorable ratio of hardware to

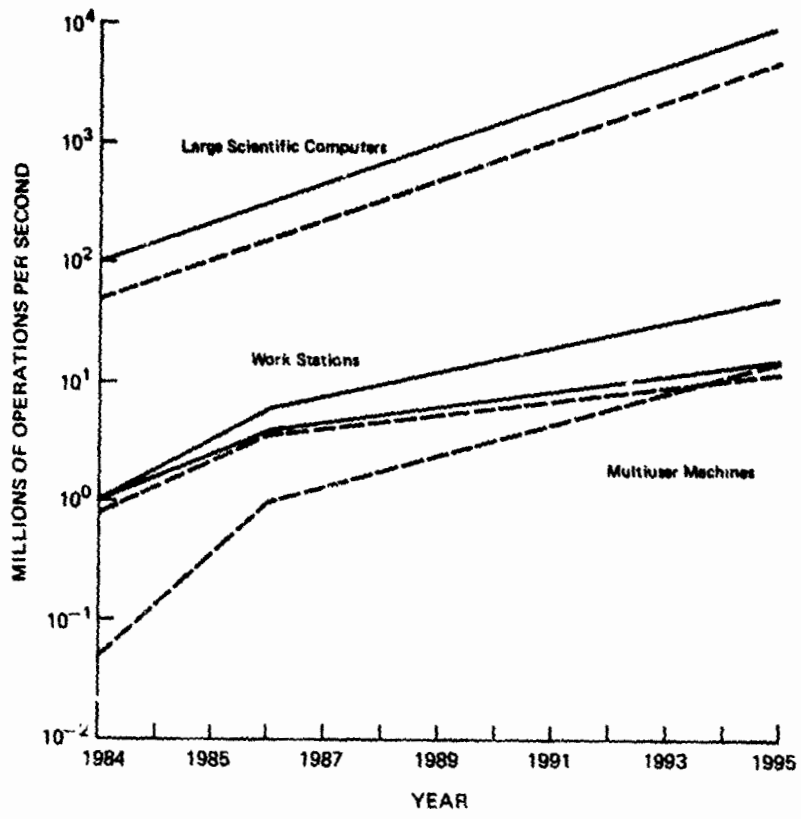


FIGURE 5.1 Projected growth of processing capabilities at constant cost. Upper bounds are for integer and lower bounds for floating point operations.

personnel costs. Another advantage is that hardware costs may be reduced by sharing access to expensive peripheral devices, since work stations can access tapes and printers through communication networks. On the other hand, the limited definition and acceptance of high-level communication standards in scientific computing poses problems for the space science community for such access.

The effectiveness of these systems will further increase as more software becomes available and is shared among these systems. The sharing of software is somewhat inhibited by the variety of work station designs now

becoming available, even though many systems are based on the same set of underlying chips. Differences in the manner in which these chips are connected, the machine architecture, will often inhibit sharing of software between different systems at the machine level. Having a common operating system, such as UNIX, now available on many of these systems, can hide the differences. Furthermore, programs written in common languages, such as C and Fortran, will be shareable over differing systems, with minor costs of adaptations to specific systems.

Growth in work stations will be driven by requirements for office systems and computer-aided design and manufacturing (CAD-CAM) applications. These stations will require reliable and fast data management, and graphic output. CAD-CAM demands for floating point computation are not as great as the requirements found in space research. We hence see today fairly weak floating point computational capability, but expect augmentation of these systems with economical floating point VLSI hardware. The requirements of the CAD-CAM marketplace will cause work stations to become available with powerful graphics capability. Transformation from three-dimensional data to the two-dimensional images will be done by specialized chips. Some of the graphics will be beneficial for space data processing as well. Although grey-scale image processing may lag, the 8- to 16-bit image processing required for scientific data analysis follows relatively easily from the 4- to 8-bit requirements for graphics displays.

The effectiveness of the high-data-rate human interface, which is provided by work stations with integrated work environments and bit-mapped graphics, appears to make these systems very suitable for interactive data analysis. It is difficult for multiuser systems using remote terminals to compete in this arena. As noted, the stations need not be isolated, since multiuser systems may be accessed via the networks to provide file services.

The computational capabilities of work stations that are projected in Table 5.1 use as unit of performance the internal computational rates of the underlying VLSI processors, with consideration of their instruction set and the access speed to the working memory of the computers. It must be realized that these instruction execution rates are not directly comparable to instruction rates on other types of processors. The factors in the expected rapid growth of capabilities in the work stations are due to greater integration of floating point hardware

and the use of multiple processors within single systems, in order to carry out subsidiary functions required for effective computing, such as input-output, display control, and memory management.

5.B.3. Local Multiuser Systems

Local multiuser systems in the environment of space data processing are typically 32-bit scientific processors operating in a time-sharing mode. Examples of such systems are the DEC VAX's, the Data General Eagle (the 32-bit successor of the Eclipse processors), the Prime systems, and certain IBM processors. In addition to having more capacity for sharing processing and larger memories than the work stations, they are also distinguished by having substantial sets of peripherals. Such peripherals include magnetic tape units useful for space data entry and archiving, disk storage units that contain the individual users' data, as well as data being shared by the set of users of such a system. Other peripherals include output devices, which can be too costly for individual researchers, such as laser printers and imaging equipment.

The terminals associated with even small multiuser systems are often geographically dispersed to some extent. Some researchers even now use "smart" terminals (e.g., personal computers) to conduct local operations, and then they transfer results to the multiuser system to utilize shared peripherals, such as letter-quality printers. High-speed links to computers in other departments of the institution are often available through multiuser systems, again providing shared access to equipment that is too costly for small groups to acquire and maintain.

Local multiuser systems are typically owned by an SSDMU consisting of a relatively small to medium-sized group in a research or academic institution. We envision that these SSDMUs will house many of the active data bases in our distributed computation and data management approach. Although the systems are shared, the fact that they are usually owned by a single group means that costs of management to provide privacy, protection, and resource allocation can remain minimal. Typically, a small technical staff is associated with each of these systems. The staff keeps the shared software up to date, communicates with the suppliers of the software, and advises users on the best way to utilize these systems. Manage-

ment control is typically exercised by scientific personnel.

We project that these systems will see steady growth of capabilities as technology improves, although the range within individual configurations will probably vary greatly. The plot of growth for these systems, shown in Figure 5.1, shows these multiuser systems increasing in capability toward the early 1990s at a slightly slower rate than work stations. The systems then begin to level off.

Since these machines are configured for scientific computation, the ratio of processing rates to floating point computational rates favor numeric computations more than the ratios seen in many of the work stations available at present.

5.B.4. High-Speed Scientific Processors

Major high-speed scientific processors (i.e., parallel processors) are currently found associated with large facilities. Typical machines at this time are the large Control Data Corporation Cyber, the Cray-1, and multiprocessor configurations of large IBM equipment. These machines are distinguished by having parallel processing capabilities. Arrays of data can be brought into the processor, and instructions are available that operate on all elements of a vector simultaneously. The complexity of these machines makes it often difficult to program in ways that realize their full potential. Software written for serially oriented machines does not usually take advantage of the parallel processing capabilities. Also, certain computations lend themselves more to the exploitation of the parallel processing capability inherent in these machines than other types of computations. In the projections in Table 5.1 and Figure 5.1, we consider a typical mix of these problems, and it may be that certain computations that are very suitable for a given machine architecture could be executed at a considerably higher speed. Even so, we project a capability doubling interval of only several years.

Facilities operating the current class of machines typically need considerable staff to maintain the hardware and software systems associated with such an institutional service. The size of these operations is such that there is also a professional management staff associated with these computation centers. Many users

will want to access these facilities remotely and thus will require formal training, documentation, and advice at a level that is provided only informally at the smaller computational facilities.

Currently, there is investment by NASA in the development of these supercomputers, specifically the massively parallel processor at the GSFC and the Crays at the NASA Ames Research Center. We expect that this investment will pay off over the long term. A parallel investment in software for space research is necessary in order to obtain the full benefit from the hardware investment. Such an investment requires recognition of commonality of problems and solutions to be effective. Such work may be best accomplished by vendors or specialized software groups. Our long-term projections assume that these investments will be made.

5.B.5. Input-Output Data Rate and Storage

The large volume of space sciences data demands that we consider issues related to input-output data rates and to storage capacity. Once large volumes of data are stored, they also have to be accessible at a reasonable rate. For example, in many initial analyses large numbers of data are first filtered for significant events, a process requiring rapid access to data bases.

The data rate of input-output devices determines the data transfer speed of the interface between the processors and the data entry and storage units. We see today few fundamental limits on the data rate available for input-ouput. The input-output bandwidth only leads to bottlenecks if unusual systems are to be configured, say, a work station with a storage capability of a size normally associated with a large scientific processor.

For data storage we distinguish three types of operational requirements:

- For archival purposes we require a low demand rate and a high volume.
- For the repository data base we require a high entry rate and a moderate volume.
- For the active data base and working storage we require a high demand rate and a volume that is related to the size of the data of interest and the size of the immediate user community.

TABLE 5.2 Projected Advances in Data Storage, Assuming Constant Cost and Access Time

Function	Cost/ System, $1000	Access Time, ms	Typical Fetch Size, bytes	Total Size, bytes 1983	1986	1995
Working storage	5	100	1K	2M	20M	200M
Data base	20	30	10K	500M	2,000M	8,000M
Archive	500	30,000	100M	10^{12}	10^{13}	10^{16}

Each type of requirement can be mapped to a certain category of storage, as follows, with projections tabulated in Table 5.2 and plotted in Figure 5.2.

5.B.5.1. Working Storage Level. We define working storage as the secondary memory used for immediate, local processing. Working storage is needed in repository, active data base, and archival environments. Working storage will hold, for instance, arrays of selected values for some analysis. At the working storage level we typically see small, fixed disk storage devices that currently cost above $1,000 for capacities of several tens of megabytes. These devices are predominately used at the work station level, although they are also available with intelligent terminal capabilities. Access times to data are on the order of 100 milliseconds (ms), and data quantities obtained per retrieval are modest, say 5,000 bytes per fetch. At this cost level we mainly see rapid increases in storage capacity and more modest increases in access speeds. When larger computers are used, this requirement is satisfied by shared usage of larger disks. While larger disks permit more flexibility in access, the effective access times and unit storage costs on larger, shared multiuser computers do not differ greatly from those seen on work stations.

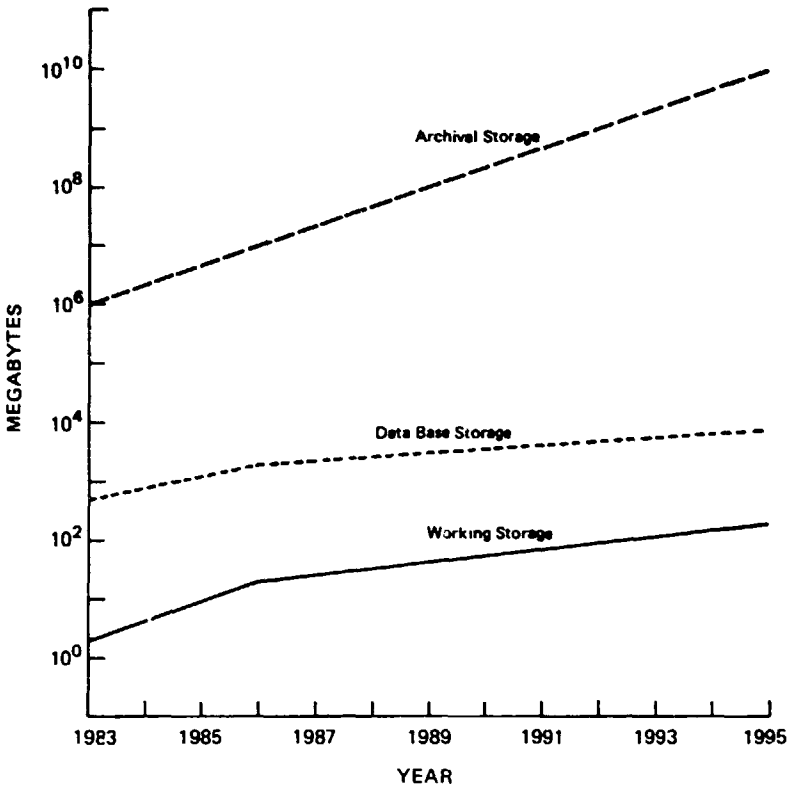


FIGURE 5.2 Projected growth of storage capacity at constant cost.

5.B.5.2. Data Base Level. At the data base levels where data are to be stored for some period of time for frequent access, we see disk units providing substantial storage capacities. Typical costs for these units are currently in excess of $10,0 [illegible]r hundreds of megabytes, so that these devices are ty[illegible]ly found attached to multiuser systems. They provi[illegible]access times on the order of 30 ms. In order to have rapid access to large scientific data quantities, the data may be organized in units of about 10 bytes per fetch. Rapid technological process is continuing in the magnetic recording area on which this technology is based. We do see the technology maturing and expect less rapid progress in the longer term.

5.B.5.3. Repositories. For repositories of data that would last for some specified period, with fairly frequent access, we find either tape units, similar to those traditionally used for archives, or disk units of the data base variety to be reasonable.

5.B.5.4. Archival Storage. To date, major archival storage facilities have been based either on conventional magnetic tape, or on magnetic tape libraries with automated reel or cartridge retrieval. The automatic devices cost on the order of $1,000,000. In low-volume operations, reels are mounted and dismounted manually, and the investment in equipment is more modest.

The mechanical devices used to manage tapes and cartridges have access time delays on the order of 3 s. The data quantities made accessible per retrieval are on the order of 100 Mbytes. In some devices the retrieval is staged, which means that data are brought in from cartridges or reels and placed on some intermediate storage medium, typically on disks, for further processing.

We expect repository and archival SSDMUs to be enhanced by the use of optical recording techniques. Early generation devices are already available, but the greatest growth will appear once the systems designers recognize this capability and provide the incentive for maturing of this technology.

The increase in storage capacity made possible by write-once optical disk technology does not yet provide a solution to many of the user requirements for massive storage. High-density optical storage technology is oriented toward writing on a particular location of the disk only once, although it can then be safely read many times. Conventional software to support active files utilizes both data and access structures on the storage devices. The access structures permit direct access, avoiding the need for serial search. Serial search can take many minutes over the data volume stored on tape, but might take hours over the volume considered for optical storage. On erasable media, such as magnetic disks, when further data are appended or analyzed, the access structures are modified and rewritten. If the data are to be retrieved in a flexible manner, if the indexed data elements are small (have a fine granularity), or if information is frequently added, those access structures will require a great deal of storage, since the access structure must be updated and added to older

structures already on the disk. Even if the data are stable, as long as the access management routines expect rewrite capability for the access structures, optical storage may be inefficient for this function.

5.B.6. Data Base Machines

Data base machines combine processing and storage capabilities and as such are not separately projected in Table 5.1. The current generation of data base machines uses relatively simple internal algorithms for data storage and access. They provide a significant increase in performance and simplicity of use over software running in normal machines (2-3 for IDM-500 versus the functionally similar ORACLE software). They also remove a processing load from the main computer.

The current generation of data base machines, due to their simple design, will show few fundamental improvements in the future. They provide the greatest advantage for relatively simple files. Sequential scanning of large files is considerably faster, perhaps by a factor of 5, versus software program access.

We expect that, in the future, algorithms in data base machines will become as sophisticated as algorithms now possible in software. At this point the performance of data base machines will be limited by the performance of the attached storage devices. Since the systems will rely less on existing processor capabilities, they will be able to use these storage devices to a better extent than is possible with generalized processors. Relatively small data quantities used for ancilliary and working storage in processing will be kept by data base machines in semiconductor memory, providing an order of magnitude faster access to those data as compared to magnetic disks.

5.B.7. Communication

Communication has become an essential part of modern computation. The feasibility of work stations, time-shared computations, and access to large scientific processors, is based to great extent on increased communication capabilities. Distinct types of communication capabilities will be used in SSDMU environments. We distinguish four generic types: local networks, public telephone-based systems, dedicated telephone-type

Table 5.3 Projected Advances in Communications, Assuming Constant Cost

	Communications--Performance--values are upper limits for widely available, commercial equipment		
	1983	1986	1995
Long-distance-switched circuit full duplex	1.2 kbps	9.6 kbps	56 kbps
Local area network	1 Mbps	6 Mbps	30 Mbps
Satellite	5 Mbps	50 Mbps	200 Mbps

	Communications--costs at present time in $1000/month		
	1.5 Mbps	56 kbps	9.6 kbps
Leased landline (AT&T)			
1000 miles	20	4.8	2.0
2000 miles	56	6.2	2.4
Leased satellite channel (ASC) (includes antennas; any distance)	45		

systems, and satellite networks. Table 5.3 and Figure 5.3 summarize current and projected communications capabilities and costs.

5.B.7.1. Local Networks. Local networks are seeing rapid growth. They are typically based on cable networks within single or closely located buildings. The lines

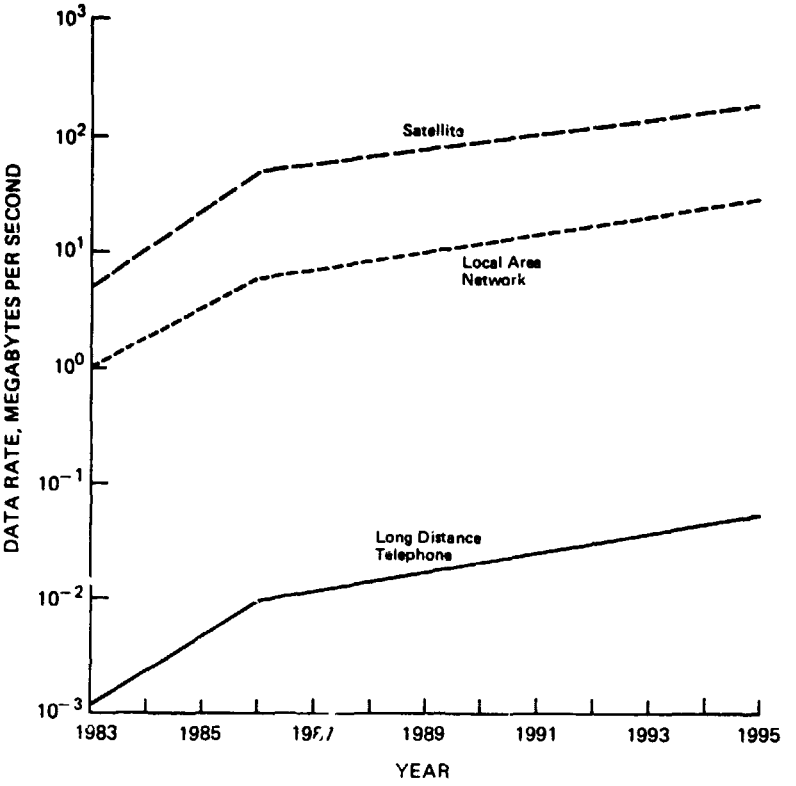


FIGURE 5.3 Projected growth in communications at constant cost.

may use dedicated cables based on cable TV network (CATV) technology, or they may share cabling for a private telephone branch exchange (PBX) network. Optical fiber technology can lead to further increases in network data rates, so that the limiting factor becomes the capability of the input-output interface.

Local networks are used to interconnect work stations and institutional processors. They provide the capability to share access to peripherals such as high-speed printers, data bases and archival storage units, which are not appropriate for single-user work stations. To interconnect local networks and other communication facilities, gateways can be provided.

These networks are economical, and the capability is largely limited by the input-output data rates of the individual processors and the number of processors that they are expected to serve. Since networks require their own control and error-correcting subsystems, the speed at which data can be acquired from devices over a network versus direct data rates will typically be a factor of 2 less than if the same devices were attached locally to the processors.

5.B.7.2. Public Telephone Access. The public telephone system provides the most flexible and, in small quantities, the most economical way to access remote computers. Typical speeds of interaction for a remote user with a simple terminal are 300, 1200, and 9600 bps. Further costs are incurred for local and long-distance tolls. We expect that charging by time used will become common even for local telephone calls, so that incremental costs will increase somewhat, and could become substantial for 8-hour/day hookups.

At the 9600 bps rate, the information transfer is so fast that for a terminal the speed is only utilized a fraction of the time. The human processing time exceeds the transmission time. For computer to computer communication, however, this speed can be quite inadequate since large files will still take many minutes to hours to be transmitted. Thus we see public telephone services mainly being effective for management and mail communication among the scientific community. In fact, use of TELEMAIL at 300 to 1200 baud in writing this report significantly decreased the response time to draft versions of both text and tables. Remote directory and catalog searches, some data browsing, and occasional data transfer might take place over these lines. Likewise, remote processing might be done.

5.B.7.3. Dedicated Telephone Networks. With dedicated networks, conventional telephone links can be acquired and driven at much higher rates. Such a network becomes effective if it can serve a sufficiently large number of users. The prime example of a dedicated telephone network is the ARPANET and its derivatives. These systems are based on lines of 56 kbps capability. The resource allocation of dedicated telephone networks is based on packet technology, which allows very effective sharing of the communication lines. Many conversations and file transfers can occur during the same time

internally on one line, subject only to the data rate (typically 50 kbps) limitation. At this speed, effective file transfer of moderately sized files is feasible as well as fully adequate remote user interaction.

In order for a node to participate in a dedicated network such as the ARPA (Advanced Research Project Agency--DOD) system, a substantial hardware investment is required. Some of this investment is due to the substantial software needed to packetize and properly control access to the network. An interface is currently budgeted at approximately $35,000 to $100,000 and requires some maintenance effort in order to keep up with changes and advances in protocols as the networks grow.

The existence of dedicated networks could tie SSDMUs and researchers using space science data much closer together than they currently are. The investment would be a major one. On the other hand, the requirements to have remote directory and catalog searches, the ability to remotely browse through data, and in some cases, the access needed to a variety of data sets, could be met rather nicely by such a system.

5.B.8. Satellite Networks

A satellite network consists of ground-based transmission stations, ground-based receiver stations, and space-based transponders. Commercial transponders typically have a data rate on the order of 1.2 Mbps. The speed obtainable via satellite networks is adequate for all types of remote interactions foreseen today. Depending on the area to be covered, one or several transponders should suffice for the needs of the space science community. Adequate transmission stations are projected to cost approximately $100,000, while receive-only stations could be considerably more economical, perhaps about $20,000. One can thus project relatively inexpensive ways of transmitting data directly to users, including those at universities, although full duplex transmission capabilities may have to be limited to large regional centers.

5.B.9. Data Display and Presentation

Space science users prefer to represent data as plots, charts, color graphics, and images. These representations

TABLE 5.4 Projected Advances in CRT Display Resolution, Assuming Constant Cost

	1983	1986	1995
Most common screen resolution (lines)	512	1024	2048

are more space efficient and often more easily understood than tables of numbers. Such representations were in the past generated by hand, or through photographic manipulation. Today scientists are using digitally stored data to a great extent, and the direct display of such data is a high priority. Thanks to the recent commercial developments of graphics for business and design/manufacturing, graphics terminals are taken for granted, and more powerful image display devices are now available at reasonable cost. Trends in data display and presentation devices are given in Table 5.4.

5.B.9.1. Plotting Devices. The growth of automated drafting for manufacturing and architecture has led to a wide choice of incremental pen plotters. These may have multiple pens with different inks or colors that are automatically selectable during the drawing of a plot. A plotter with one-thousandth-inch steps using 11" x 17" paper and four or eight selectable pens costs currently under $2,500. Where the pen plotter draws one point or line at a time, printers based on xerographic principles may plot a full row or page mixed text and graphics simultaneously at far higher speed. These faster plotting devices are commonly restricted to black images. Special, but expensive, models will plot in three colors or on mylar film. Speeds also vary. Plotters are now available for less than $5,000 that plot a full 8-1/2" x 11" page in 150 seconds, regardless of the number of points used. Linear resolution, though, is one-third that of the incremental pen plotters. We see mainly increases in speed, resolution, and commandability for plotters.

5.B.9.2. Visual Display. The common visual display device is the video monitor, brought to high development through the television industry. A monochrome monitor

capable of displaying 10^7 points currently costs under $3,000. Color video monitors are far more complex in design. A color monitor displaying 2×10^6 points currently costs under $5,000. These monitors are bit-mapped, driven by a video generator that scans a section of digital memory containing a bit-pattern corresponding to the graphic image. An integrated system with monitor and memory mapping will display an image of 3×10^5 points with 256 shades of grey or separate color hues and currently costs about $7,000. Larger memory units that will display over 10^7 points are now commercially available, but are more properly considered as components of image processing systems. We foresee significant increases in the level of processing of data within memory, leading to a greater degree of interactive display and analysis.

5.B.9.3. Color/Grey-Scale Hardcopy. Plotters, while precise and inexpensive, have limited ability for representing grey-scale or color shading. For this purpose, a modulated CRT or laser beam that exposes photosensitive material is practical. Such hardcopy devices are driven either directly from the video signal feeding a monitor, or through a digital interface. Those using a video signal are least expensive, though limited in resolution. The registration for color images is excellent, since exposures are made through automatically positioned color filters. A unit displaying up to 2×10^5 points is available for under $2,000. A similar video-driven design for up to 2×10^6 points now costs under $10,000. Units with digital interfaces have resolutions of up to 10^7 points and can be currently purchased for about $25,000, and those with over 10^8 points, depending on film size, cost about $80,000. Such units are slower in operation than the direct video-driven units and are more suitable for high-volume, production-oriented systems.

5.C. EXISTING AND PROJECTED SOFTWARE CAPABILITIES

5.C.1. Introduction

Hardware improvements, without corresponding improvements in software capabilities, will not meet the demands on data management and computation discussed in the previous chapters. To solve the problems faced in space

data processing, equal attention has to be given to software. System support software is mainly provided by the hardware vendors or specialized software groups. The applications software that is required to process science data is developed by the combination of trained people and effective tools. We discuss the human element first.

5.C.1.1. People. The formalization of programming experience over the last 25 years has enabled major improvements in education. The effect of these improvements is that recent computer science graduates are rapidly productive and show great flexibility and ingenuity in using the available tools. Unfortunately, many existing NASA installations can take only partial advantage of these developments in a direct way, since their programming population has been largely stable. Increasing the awareness of the developing gap between traditional, experienced programmers and recently trained programmers, together with providing opportunities for continuing education and retraining, can help mitigate this problem. In addition, both NASA and the space science community would benefit from more exposure to current computer science techniques.

5.C.1.2. Tools. In the remainder of this section we will comment on the tools for software development. These tools have seen continuous improvement, especially in the area of reliability. Enhancement of reliability and productivity has been aided by research into program verification and development methodologies. These are research areas that are sometimes criticized as not being directly relevant. The formalization of the needed concepts and constructs, even while they are too limited for automatic application of verification techniques, is an important contribution.

In the tool areas we consider (1) traditional languages used to write program procedures, (2) the area of nonprocedural languages, (3) the use of software packages, which we define as ready-made collections of programs, and (4) the topic of data base management. Nonprocedural languages are not well-defined, so Section 5.C.3 will include some definitions.

5.C.2. Computer Languages

The area of computer languages has been a major topic of research and development for 25 years. Our conceptual understanding of languages and the compilers to handle languages has greatly increased. The use of computer languages remains our primary tool to utilize computers. It is unfortunate that our progress in using the results of this research, namely, new languages, has been slow, although on the positive side, our ability to use the existing languages has certainly improved. We will cite some languages to support this contention.

5.C.2.1. Fortran. Fortran, the earliest practical language to be used for translating formulas into computer codes, is still the workhorse of much programming in SSDMU environments. Its widespread acceptance makes programs written in Fortran transportable with modest effort, subject to the usual considerations of good software development practices. We do not see Fortran being replaced by traditional numeric programs in the timeframe we are considering.

5.C.2.2. PASCAL. PASCAL has become one of the major languages used when teaching programming. Because of this aspect it will be seen more, and used more, in all kinds of environments. PASCAL, as originally defined, is easy to implement on both large and small machines, and this contributes to its spread. A major lack of basic PASCAL is that arrays cannot be parameterized, which limits the generality of subroutines. This restriction can be expected to be overcome in future versions of PASCAL and has been addressed in the PASCAL Standards. A successor language, MODULA II, overcomes many of these problems, and we foresee its spread in system applications.

A more serious problem to portability and compatability of PASCAL and MODULA is the limited input-output definitions provided with PASCAL. All external files are treated as one continuous stream of characters. While this concept has great generality, it limits the use of PASCAL in data processing, where often more complex data storage structures are essential.

5.C.2.3. PL/1. PL/1 is a much more comprehensive language, but the complexity of its implementation has caused the spread of the language to be quite slow. Its

distinguishing feature for data processing is that record input-output is defined within the language. Such a definition is essential to make data processing programs as portable as numeric programs are now. The fact that PL/1 is now available for Digital Equipment Corporation VAX Machines under the VMS operating system, as well as on Burroughs and IBM personal computers, may make PL/1 a more valid choice than when it was available only on major IBM equipment. We cannot predict, however, much momentum in its further development.

5.C.2.4. ADA. ADA is a language recently developed under sponsorship of DOD with a strong emphasis on real-time processing. The research efforts invested by DOD through industry and academia in the development of ADA give it a great deal of momentum. A major concern of the sponsor is continuing portability of ADA programs. An important aspect of ADA development is the recognition of the importance of a comprehensive support environment. Although this environment does not exist today, when it becomes available it will make not only programs but also programming methodology much more portable across machines than is seen now

The major weakness of ADA for NASA data processing is the lack of record input-output facilities; the capability to provide packages may overcome that. We hope that packages that define input-ouput with adequate capability to support data base management will be developed before an excessive variety of approaches introduces de facto incompatibilities into ADA.

5.C.2.5. The C Language. The C language was developed in the UNIX environment for DEC PDP/11 and later VAX computers. It appears to be capturing much of the programming market segment that was previously seen to require assembly language, programming. Since it is a relatively low level language it provides facilities for detailed control of hardware. Compilers for C are available for a great variety of machines, ranging from micros to mainframes. If care is taken, the code can be quite portable, especially for machines with common character sizes.

We do not see C as a major replacement for existing procedural codes, but expect that in its critical niche it will have a long-term future. Use of C versus assembly language can greatly promote portability of critical programs from machine to machine at a cost that is much

less than a complete recoding of these routines in assembly language. Whereas recoding a small assembly language program requires about 50 percent of the original effort, that effort in C may be about 10 percent.

5.C.2.6. LISP. LISP is the major implementation language for packages using artificial intelligence (AI) techniques. Today well-developed programs based on AI techniques have at times been recoded in other languages in order to gain increased execution efficiency, albeit at a great loss of flexibility. Future developments in LISP such as Standard LISP and LISP machines should make such recoding less frequent. We do expect to see LISP-coded packages finding use in space data processing, but we do not see this language becoming a major programming language within the community.

5.C.2.7. Other Procedural Languages. There are many other programming languages that are found to limited extents within the NASA environment. The most common language for commercial data processing is COBOL. It is relatively rarely used within scientific data processing. PROLOG, a logic programming language, may see increased utilization for artificial intelligence applications. It can be viewed both as a logic programming language, requiring a programmer who is versed in logic and the implementation rules of PROLOG, and as an artificial intelligence system using predefined resolution techniques. The former view is probably more realistic.

Other specialized languages associated with special packages will be discussed in Section 5.C.4.

5.C.3. Nonprocedural Languages

We define nonprocedural languages to be those languages where the actions to be carried out by the computer are not specified step-by-step but are automatically derived from a specification of an objective to be achieved.

The specifications are given to the language processor, and a program is generated to carry out the task. The programs depend on substantial, prewritten libraries. There are a wide variety of nonprocedural languages, although in total they only perform a small fraction of space data processing. A common feature of nonprocedural systems is that they include in their

processing programs a fair amount of application-dependent semantics. This makes these systems much less general than conventional programming languages.

A simple form of nonprocedural languages is report generators. In report generators the layout and formulas for variables to be printed as reports are given. These specifications are converted to programs that create the report. Report generators will also be sensitive to specifications of the output devices to be used, so that the same report specification will produce well-formatted output on a variety of devices. For example, MARK IV is a report generator used at the Jet Propulsion Laboratory for space data catalogs.

Simulation languages are another class of nonprocedural languages. Simulation languages permit a scientific model to be defined in terms of constraint equations and an initial state. Simulation languages exist both for discrete (SIMSCRIPT) and continuous (CSMP) models. The modeling constraints are specified to the simulator, and the programming systems find solutions that satisfy these constraints. Once a physical structure is described, a simulation program will evaluate the model through successive timesteps. It may halt when equilibrium is achieved, or when a predefined condition has been reached.

Image processing is a major issue within NASA and packages, such as UNIPS from the University of Florida provide a nonprocedural language for image processing. At this time, image processing languages have not been generalized to the extent that they are portable between systems, although many of the semantics should be independent of the machine environment.

Another class of nonprocedural languages is the symbolic expression evaluators. These operate on algebraic expressions provided in symbolic form and reduce the expressions into simpler and often computationally more feasible expressions. These systems are finding increased use in design and engineering applications. MACSgma is such a language, which is available from MIT via the ARPANET.

A special case of nonprocedural languages is found in data base management systems that include both data description languages and data manipulation languages. We will discuss their functions in Section 5.C.5.

5.C.4. Software Packages

In the near term we see the greatest improvements for SSDMUs in an increased use of portable software program packages. A software package is an integrated collection of programs designed to solve problems of some given category. Very sophisticated program packages provide facilities akin to nonprocedural languages, but are controlled using very high level procedural languages. The user of packaged systems can express problems suitable for the package concisely. On the other hand, problems outside the specialty of a package cannot be expressed, and problems on the boundary may be awkward to handle. A user who needs a variety of tools may have to know several systems, and this causes confusion and frustration.

Packages will be needed because relative software development costs are not dropping as quickly as hardware costs. Thus relatively small SSDMUs will find it increasingly difficult to write and maintain their own software.

Program packages are often developed at large institutions and then shared with other users. A major hindrance to faster spread of programming packages is lack of portability, documentation, completeness, and consistency. Documentation of the languages that the user needs to control these packages is often inadequate. While improved documentation can overcome some of these problems, the volume of the documentation required to describe these packages is often great, making it difficult to comprehend what the user should do.

Packages that have developed high-quality user interfaces are finding increased popularity. A problem with some of these interfaces is that they may be quite machine or terminal dependent, especially if the interfaces require on-line user interaction. Programs that are operated in a more traditional batch-oriented setting are often more portable.

As the user population of a package increases, feedback from the users will improve the package, if there is a group that is willing to take responsibility to update the package. Feedback works well when packages are being maintained by commercial organizations or large SSDMUs and not as well when packages are being developed and maintained by relatively small, research-oriented SSDMUs. The Transportable Applications Executive (TAE), developed and maintained by the Goddard Space Flight

Center, is an example of a reasonably successful NASA-sponsored endeavor. The effort toward developing a transportable data analysis package at the Space Telescope Science Institute, which is directed toward the Space Telescope Mission, but coordinated with other astronomy analysis system developments, is another promising area of activity. The STSCI package is meant to be available generally to the ST user community. We recommend that NASA continue to support these efforts, since it is difficult for small research institutes or academic groups to find the resources to develop or even to improve packages to levels that are beneficial to outside users.

A successful package developer must estimate the number of potential users, and attract competent programmers for the task. A developer also has to build a support organization to help with installation, sometimes adaptation, and problems encountered in use of the products. Except for the world of microcomputers, where packages are widely distributed and mass-marketed, we see that such packages will cost typically between $10,000 to $100,000 if developed and distributed commercially. It is not commercially possible for developers to provide software at lower prices if the users expect service and support.

Software can, at times, be obtained more cheaply from noncommercial sources such as from small university research groups. Such software in the past has typically not been supported and often has been poorly documented. The COSMIC Project at the University of Georgia provides a service, used by many government agencies, including NASA, for software distribution. While some programs are well-maintained, e.g., NASTRAN, the expectation for obtaining readily usable software from noncommercial sources is low. The final costs of installing apparently free software have often been quite high in the past.

If SSDMUs that develop potentially useful software are also going to provide some support service to their users, they will have to develop a means for recharging the costs incurred. We see some of this happening now; several universities distribute their software for fees that are higher than the direct distribution costs, although they are much less than the commercial prices computed by summing investment depreciation, production costs, marketing cost, and profit. We cannot estimate what the relative costs will be for laboratory- and university-developed software versus commercially

developed software if identical accounting principles were followed. This is an important issue, since NASA should have a role in software development and distribution for its community. NASA should aggressively support the development and maintenance of software packages for use by the space science community, to the extent that the packages are not being developed by commercial vendors. An example of a software package that would be of great use for analysis of imaging spectrometer data is one using an expert systems approach to extraction of spectral information pertinent to mineral chemistry, vegetation type, etc.

5.C.5. Data Base Management Systems

5.C.5.1. Status. Data base management systems formalize the handling of large quantities of data kept on external storage devices. The semantics of processing large quantities of data are well enough understood to have made it possible to create data base management systems that are applicable to a wide range of applications, including applications of use in a variety of SSDMU environments.

A data base management system includes facilities for record management, handling of multiple but related files, and a schema. The schema for a data base is a symbolic description of the relations between data base parameters. It is represented in such a way that it can be formally interpreted by programs that access the data base. The symbolic description of the data permits data to be shared by diverse users. The users are now isolated from the detailed physical storage of the data values. The symbolic definitions permit growth of the data base, changes to certain limits of the data base structure, and portability of the data to different storage devices and computers without affecting the programs using the data.

Data base management systems have been widely accepted in industry and government, but are still relatively little used for basic data collection in most SSDMU environments. One problem hindering their acceptance is the extremely large volume of space science data and to lesser degrees the intrinsic structure of the data. Certain data types, such as vectors and images, are not as strongly supported by commercial data base systems as a traditional record and field structure. We do foresee,

however, that data base systems must enter to a larger extent the scientific community. Without greater use of data base systems, the search and access requirements outlined in Chapter 4 would not be met.

5.C.5.2. Types of Data Base Management Systems: Interfaces. There are a variety of approaches to data bases, and since there is much active work in the area, a fair amount of confusion about the applicability of data base approaches for space science applications continues to exist. In data bases we can distinguish between the user interface and the underlying implementation. Through the interfaces, the users specify the manipulations to be carried out on the data base. Specialized computer languages, using nonprocedural (relational) or procedural (navigational) approaches, provide the interface. The underlying implementation type determines the effectiveness with which the actions can be carried out. We will define types for both aspects.

A relational interface provides a nonprocedural way to access and manipulate data. Limitations are that conditional and loop-type (e.g., do loops) structures are not part of the basic relational interface specifications. If such types of access specifications are needed, a procedural capability must be invoked. These procedures are described using data manipulation languages and combined into small program segments that implement transactions to be performed on the data base.

Some data base systems do not provide a relational query capability, but only provide an interface for procedural access. Data bases that implement a network structure, and require the user to follow the network to locate data, are frequently found in commercial data processing. If the network is limited in complexity to a single hierarchy, we speak of a hierarchical interface. A hierarchical interface can be very natural for a user whose understanding of the data structure coincides with the way in which the data base interface presents the data. A system that internally uses a network structure, but provides multiple distinct hierarchical interfaces, is IBM's IMS, frequently used to aid in space vehicle manufacturing.

While procedural access to a data base is powerful, it also implies that the user understands the data base structure. However, it is desirable that the program manipulation language be independent of the data base structure. This independence will avoid having data base

changes constrained by compatibility requirements related to other users of the data base.

Programs access data base through transactions. A transaction is a small program for some well-specified type of task, often invoked from a terminal, which interacts with the user and reads and writes the data base. Keeping a transaction program isolated from the actual structure of the data base by always interpreting all requests via the data base schema can reduce the efficiency of the transaction program. The ability to adjust a transaction program by recompilation with a revised structure description when the data base is changed can provide an adequate compromise.

Access to data via natural languages (e.g., English) is feasible today for special environments. An example of such access is the Moon Rock Catalog System developed for the Smithsonian Astrophysical Observatory. The ambiguities that plague general processing of natural language statements are generally avoided in the data base environment where the scope of the vocabulary is limited by the scope of the data base. The natural language system can be constructed using a modest number of verbs plus nouns that are taken from the data base schema and from some files of the data base itself.

5.C.5.3. Types of Data Base Management Systems: Implementation Structure. The choice of data base implementation has a tremendous effect on efficiency. The issues of interfaces and implementation are strongly linked today, more than they should be.

The same terms--relational, network, and hierarchy--are used for implementation and interfaces, but all combinations of interface and implementation are feasible. In a pure relational implementation, each data type is treated as a separate file. The implementation ignores any relationships between data in separate files. When the data base is interrogated, the user specifies candidate relationships in the queries. The effects of not designing connections into the data base are a simpler structure and great flexibility. In the alternative implementations--sometimes referred to as hierarchical, network, or functional systems) linkages between the files are permitted. Referential structures are common on data bases that support commerical data processing operations. The linkages, which implement cross references among related data, can provide much more rapid access to related data, but have to be

TABLE 5.5 Selected Data Base Management Systems Used for Space Research Data

Name	Type	Supplier	User	Usage
RIM	Relational	Boeing	Ocean pilot	Catalog
Oracle	Relational	ORACLE/RSI Menlo Park, CA	Pilot climate data base	Catalog
INGRES	Relational	Relational-Technology Berkeley, CA	UCLA Voyager data	Catalog
IDM 500-Omnibase	Relational data base machine	Britton-Lee	Space telescope	Astronomy catalog
IDM 500-Omnibase	Relational data base machine	Britton-Lee	JPL-SFOC	Space flight operations

carefully designed in order to avoid constraints on the generality of the data manipulation.

We project continued improvements in data base system technology, especially for general relational interfaces that will become available on higher performance implementations. As noted, SSDMUs must take advantage of data base management system capabilities to meet the data management challenges raised in Chapter 4.

5.C.5.4. Data Base Management Systems Now and in the Future. Examples of data base management systems now in use for space research data are given in Table 5.5. Many commercial systems place great emphasis on rapid access to individual records. For many scientific applications, large quantities of similar data from distinct records have to be obtained. This problem is addressed in some data bases that are oriented toward CAD-CAM applications and also in some medical systems, but not generally in commercially available systems. The point still remains, however, that data base management systems have been underutilized in SSDMU environments

Considerable attention needs to be given to the overall area of selection and use of data base management systems for the space sciences. In particular, the space

science community should be encouraged to utilize data base management systems software in their activities. A number of commercially available packages can adequately handle both directories and catalogs of existing space science data. Particular attention should be given to the need to access data by location (geographic or space coordinates), by time, and in ways that depend on predefined data attributes, such as a set of parameter values that would indicate an interesting event. For the future, coupling of artificial intelligence into data base management of scientific data is an area that NASA, together with the space science community, should certainly explore. Without highly capable data base management, the complex, high- volume data of the future will be underutilized.

5.D. MATCHES BETWEEN USER REQUIREMENTS AND TECHNOLOGY

We now compare the growth in demand for computation and data management discussed in Chapters 3 and 4, with the capability that will be provided by hardware and software. We have not placed software capabilities on a quantitative scale. Software can only make the capability of hardware accessible. It does not add by itself to the performance of the hardware.

In order to overcome limitations of the raw performance capability of serial processor hardware, such as multiuser systems, we expect that the high-speed scientific processors that are being developed will have parallel data-reduction capabilities that greatly exceed their increase in performance. Since this benefit is only obtained if the problems are suitable for parallel processing, a processor can achieve this speedup only on a fraction of the tasks that are required. In certain areas in the space sciences, however, such as image processing, that fraction may be close to one. When we deal with data at higher levels of abstraction, where more complex models of analysis are used, much of the regularity that can utilize parallel processing techniques disappears, but at these levels the quantities of data to be handled are expected to be much smaller.

As discussed in Chapter 3, space science data are growing rapidly, doubling every several years if viewed over a decade time scale. Comparison of the rates of growth of data with the rate of growth storage capacity at constant cost, shows that data growth rates will

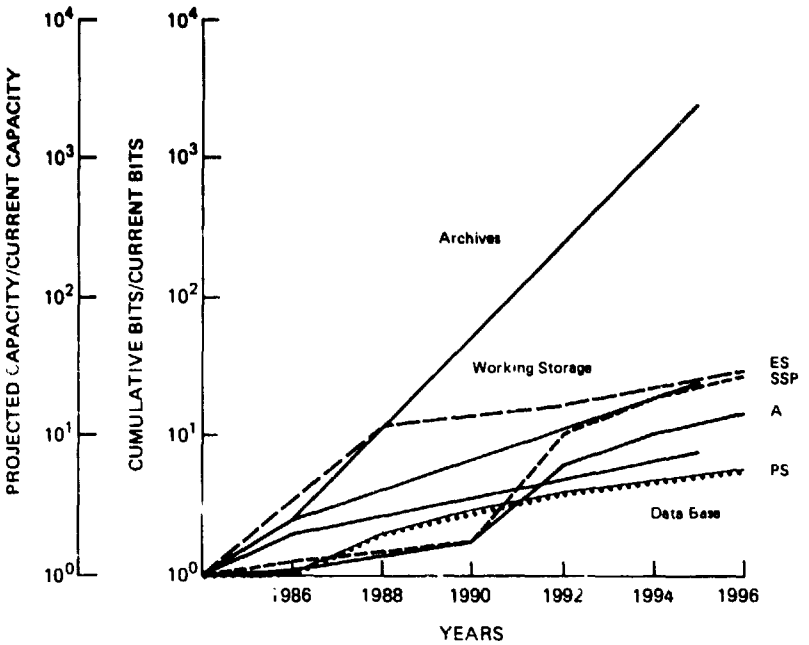


FIGURE 5.4 Comparison of rates of growth of storage demands with rates of growth of storage capacity. ES = earth sciences, SSP = solar and space physics, A = astronomy, and PS = planetary sciences. It is assumed that storage capacity and data volume are in balance at present. Thus the projections are normalized to the data and storage growth curves for 1984. Deviations in slopes between data and storage capacity growth curves thus allow one to identify when even the status quo (i.e., balance today) cannot be met. For example, data base storage will increase fast enough to meet only planetary science demands.

exceed the rate of growth of constant cost storage capacity in many cases (Figure 5.4). Thus, at constant cost, it does not seem possible to be able to store in working, repository, or archival storage even that fraction of space science data currently stored. To maintain even the status quo of data storage will require an increase in funding.

To obtain an estimate for processor requirements, we derive the processing power needed from the projected

storage demands. We again base our origin on the assumption that we need to at least maintain the present status quo, the present fraction of data that are processed to some level. This is an excessively pessimistic assumption if we consider the volume of unprocessed space data now being stored, but it does provide a minimum requirement. We derive the range of processing demands for future years from two assumptions: a lower bound based on the assumption that processor demand goes up linearly with data quantity, N; an upper bound determined by the assumption that processor demand increases by the order of N log(N) of data being stored. Both bounds can be defended based on information theory. The rate of growth of processing demands and the rate of growth of processing capabilities are overlain in Figure 5.5. Three processing capability envelopes are shown:

1. Work stations. The growth in capability for work stations. This growth is initially very rapid, and continues later at a slope that falls within the range of growth of demand.
2. Multiuser machines. The growth in capability for multiuser machines. This technology is somewhat more mature, and rises less steeply than work stations.
3. Large scientific processors. Large scientific computers, for parallel operations, show a continuing fairly steep curve, as the research investment being made in this technology pays off. This improvement, however, is restricted to computation permitting parallel processing, e.g., for image data reduction.

From Figure 5.5 we can deduce that work stations and large-scale parallel scientific calculations permit future processing of space science data at current base cost. The standard multiuser machines used by many research groups will soon fall behind the processing status quo. Unfortunately, relatively few high-powered work stations or parallel processors are now available at SSDMUs. A significant investment beyond the status quo (current funding for data management and computation) is needed if the user demands for processing are to be met. We feel that, in addition, innovative ways of combining hardware and software must be developed to meet the processing challenges. Advanced work stations attached to multiuser systems is one approach. New concepts, such as the "hypercube" multiprocessor being developed at the California Institute of Technology,

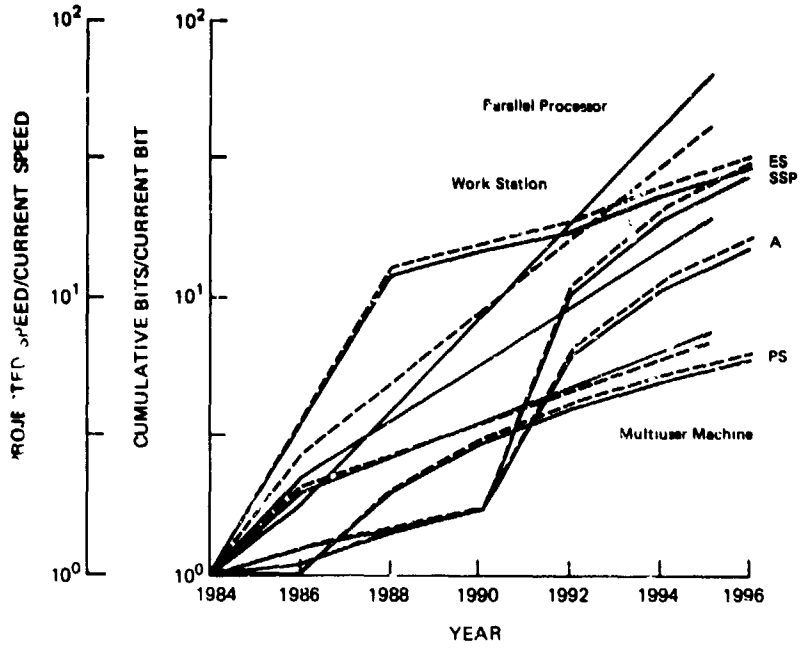


FIGURE 5.5 Comparison of rates of growth of processing demands with projected improvements in processing speed. ES = earth sciences, SSP = solar and space physics, A = astronomy, PS = planetary sciences. Upper bounds on data curves based on Nlog(N) where N = number of bits. Lower bounds based on N. As in Figure 5.4, data and processing speeds have been normalized to 1984 values. Note that multiuser systems will only meet planetary demands. Upper bound for work stations is for integer and lower bound is for floating point calculations.

provide another possible approach. Certainly, researchers will need greater access to the large-scale parallel processors, now located at major facilities, such as NASA centers.

5.E. RECOMMENDATIONS FOR TECHNOLOGY UTILIZATION AND DEVELOPMENT

The analysis performed in this chapter and supported by information gleaned from Chapters 3 and 4 le[illegible]

following technology recommendations that should be implemented to improve data management and computation within SSDMU environments.

5.E.1. Directories and Catalogs

Directories and catalogs of space science data sets, using a commercial data base management system, are feasible and are reasonable approaches to the objective of making space science data more accessible. This recommendation has already been made for the space physics community (NRC, 1984), and we recommend that the concept be extended to other space science disciplines. Accessible means available to both rapid access by existing researchers and also reasonably convenient access by researchers who do correlative and secondary data analysis. The latter may in the long range be a large fraction of the space science community.

We recommend that NASA vigorously pursue selection and implementation of directories and catalogs of existing data in a variety of SSDMU environments, including access to non-NASA data. While NASA cannot be expected to drive the construction of data sets, catalogs, and directories for other agencies or other governments, NASA can take an active leadership role in setting standards and practices for data base management, and it could encourage other bodies to participate in the process of making data accessible to the space science community. Access to non-NASA data is crucial for the solar and space physics and earth sciences communities. Electronic access is highly desirable, including the ability to search through directories and catalogs, and to browse through data subsets.

5.E.2. Standards

In order to support directory and catalog access and, more generally, integration and portability among space science groups, standards must be established and followed to a much greater extent than in the past. Standards must be developed in cooperation with the space science community to be effective. In some situations, adequate standards exist and should be followed, even if they disable some technological optimization. Examples are the communication standards for data packet transmission

(ISO X.25 etc.) and the local network standards (IEEE--802).

In other areas, nearly suitable standards exist. The office work stations are developing standards for image representation (GKS) and transmittal, as well as standards for electronic mail. These standards can often be expanded within their original concepts and will minimize the costs of reinventing, even if they still require rebuilding of software.

In areas where NASA and its research community has expertise and needs for sharing, it should become a leader in the standards area. For example, formats for raster-scan digital images as stored on disk and tape should be standardized.

5.E.3. Technology Development Efforts

Hardware and software research to support NASA objectives should be focused on those areas where commercial development will be slow. Since the total research support will be limited, it will be important to identify pressure points and provide enough support in those fields to get a critical mass. This means ignoring topics that are currently popular, although those areas should be tracked for developmental support. Review of proposed research projects has to include a mix of users and scientists knowledgeable in areas of NASA concern. As noted, areas to evaluate include applications software packages, high-speed communications networks, software for parallel processors, advanced data base management software for scientific data, and augmentation of multiuser systems with high-speed work stations or use of other innovative combinations that allow researchers to maintain at least a processing status quo.

We recommend an approach to developing applications software packages that will be of use to the space science community. In some cases, small changes to vendor-supplied packages may be needed. In other cases, major development efforts may be called for. In the latter case, the science community should be directly involved, including actual development work in appropriate situations. In any case, packages should not be developed without the continuing advice from the eventual users of the packages.

6. SPACE SCIENCE DATA MANAGEMENT UNITS THAT MEET USER REQUIREMENTS IN REASONABLE WAYS

The country needs and, unless I mistake its temper, the country demands bold, persistent experimentation. It is common sense to take a method and try it. If it fails, admit it frankly and try another. But above all, try something.

Franklin Delano Roosevelt

6.A. INTRODUCTION

In previous chapters we dealt with space science data volumes, growth rates, and uses, we summarized probable technological advances in computation and data management of relevance to SSDMU environments in the 1980s and 1990s, and we recommended a series of technology endeavors to meet user demands. However, as noted in CODMAC's previous report (NRC, 1982), technology limitations have not been the prime impediments to improved computation and data management in the past. Rather, limitations in the management approaches to data issues have been the prime impediments. Thus considerable attention needs to be devoted to how to plan, implement, and operate SSDMUs, given the need for data centers, repositories, and active data bases, and given their widespread geographic distribution. In this chapter we review the "pilot" approach that the NASA Information Systems Office (ISO) has taken to solve selected data problems, we discuss a number of SSDMU examples that we feel meet or will meet user requirements in reasonable ways, and we develop guidelines for an approach that involves distributed, but coordinated SSDMUs of varying sizes and levels of responsibilities. Basic and assumed tenets integrated throughout the discussions are the principles for successful management of scientific data that were developed by CODMAC and that are listed in Table 6.1. The primary tenet is the active involvement of the science community in the planning, implementation, and operational phases of SSDMUs.

TABLE 6.1 Principles for Successful Management of Space Science Data from First CODMAC Report (NRC, 1982)

---

1. Scientific Involvement: There should be active involvement of scientists from inception to completion of space missions, projects, and programs in order to assure production of, and access to, high-quality data sets. Scientists should be involved in planning, acquisition, processing, and archiving data. Such involvement will maximize the science-oriented and applications-oriented missions and improve the quality of applications data for application users.

2. Scientific Oversight: Oversight of scientific data-management activities should be implemented through a peer-review process that involves the user community.

3. Data Availability: Data should be made available to the scientific user community in a manner suited to scientific research needs and have the following characteristics:

(a) The data formats should strike a proper balance between flexibility and the economies of nonchanging record structure. They should be designed for ease of use by the scientist. The ability to compare diverse data sets in compatible forms may be vital to a successful research effort.

(b) Appropriate ancillary data should be supplied, as needed, with the primary data.

(c) Data should be processed and distributed to users in a timely fashion as required by the user community. This responsibility applies to principal investigators and to NASA and other agencies involved in data collection. Emphasis must be given to ensuring that data are validated.

(d) Proper documentation should accompany all data sets that have been validated and are ready for distribution or archival storage.

4. Facilities: A proper balance between cost and scientific productivity should govern the data processing and storage capabilities provided to the scientist.

5. Software: Special emphasis should be devoted to the acquisition or production of structured, transportable, and adequately documented software.

6. Scientific Data Storage: Scientific data should be suitably annotated and stored in a permanent and retrievable form. Data should be purged only when deemed no longer needed by responsible scientific overseers.

TABLE 6.1 (continued)

7. Data System Funding: Adequate financial resources should be set aside early in each project to complete data base management and computation activities; these resources should be clearly protected from loss due to overruns in costs in other parts of a given project.

6.B. PILOT PROGRAMS--LEARNING FROM EXPERIENCE

6.B.1. Introduction

In response to CODMAC's deliberations and the perceived needs of the space science community, the ISO has initiated a number of pilot computation and data management activities. The pilot programs are meant to implement, in experimental ways, prototype information systems that (1) directly involve the science community, (2) mainly utilize existing technologies, and (3) help improve computation and data management in each of the space science disciplines. The pilots are meant to be focused on data sets and driven by research projects that embody a major subset of the requirements for the entire discipline. The intent is to learn through technology and management experiments, eventually developing a prototype system that can be "handed-off" in some way for management and support by NASA's research and analysis programs. Pilots are planned for a 5-year development effort before being handed over to the relevant discipline programs.

6.B.2. Pilot Program Descriptions

In this section we describe the scope and activities within the four pilots existing or planned within the ISO and the space science community.

6.B.2.1. Pilot Ocean Data System. The Pilot Ocean Data System (PODS), begun in 1980, is the most mature pilot activity and, in fact, is in the process of being "handed over" to the Oceanic Processes Branch, Earth Science and Applications Program Office of NASA. The primary purpose of PODS is to provide access to oceanic satellite data

sets. The primary data used to evaluate the PODS approach have been the Seasat data sets. PODS was developed by the Jet Propulsion Laboratory (JPL) with help from the oceanography community. PODS consists of a central computer system at JPL with about 10^9 bytes of Seasat and other data on-line. Users access directories and catalogs of the data remotely using the RIM data base management software in interactive sessions. Browsing can be done with special, preprocessed data files, and data sets can be delivered as tables or plots. Alternatively, data can be mailed, based on user requests. In summary, PODS is an example of an SSDMU sporting centralized directory, catalog, and data services.

6.B.2.2. Pilot Climate Data System. The Pilot Climate Data Systems (PCDS), based at the Goddard Space Flight Center (GSFC) is, in many ways, parallel to JPL's PODS efforts for the oceanic sciences. The PCDS is a centralized data base housing selected climate data, with directory, catalog, and data request services. Extensive graphics capability exists within the PCDS area at the GSFC, using the Transportable Applications Executive (TAE), together with the Template graphics package. The Oracle data base management software is being used to manage the data, which include both satellite and ground meteorological measurements. The PCDS has been put together with advice from a science steering group composed of NASA and university scientists.

6.B.2.3. Pilot Planetary Data System. The Pilot Planetary Data System (PPDS) is designed to experiment with ways to improve computation and data management for planetary missions (e.g., MGCO mission) and for SSDMUs that are involved in processing and curation of planetary data. Planetary data and researchers are widely distributed, being located at federal, university, and private laboratories. Thus, the approach being used in PPDS activities is a distributed one, involving housing test data sets at five universities, at JPL, and the U.S. Geological Survey. The individual sites will contribute to a directory and catalog of data probably to be housed in a central data base machine at JPL. Individual data sets, on the other hand, will be housed with and under the control of the various groups involved. The sites will be electronically linked by 1200- and 9600-baud "dial-up" modems to allow users to do directory and catalog searches and to then be directed to the

appropriate sites for access to the data sets proper. The experiments are centered in part on understanding the best ways of implementing such a distributed approach, including both management and technology lessons.

6.B.2.4. Pilot Land Data System. In the earth sciences, effective use of satellite remote sensing data has been consistently handicapped by inadequate information systems. The goal of the Pilot Land Data System (PLDS), which is still in the planning stage, is to establish a limited-scale information system to explore scientific, technical, and management approaches to satisfying the needs of that part of the earth sciences community concerned with the space-borne observations of the land surface (Estes et al., 1984) Because the research community and the data sets of concern are located at a number of institutions, the approach taken is to develop a prototype distributed information system. The PLDS is being specifically structured to serve the needs of NASA and NASA-related land science users in universities, private industry, and other federal and state governmental agencies. Development of the PLDS represents a significant challenge, due to the number and size of relevant data acquisition, networking, processing and analysis systems, and the need to connect scientists at a number of institutions across the country who are currently employing a variety of hardware and software systems. As such, the PLDS is conceived of as a proof-of-concept tool.

PLDS implementation will proceed in stages to involve, in system management and operation, researchers with a long-term commitment to the use of the data and to sharing their data with others for the purpose of conducting science research. System development will proceed to link key research groups conducting land science research with key data archive, depositories, and suppliers. PLDS will strive to improve the ability to do science and to minimize the time currently spent by scientists performing library, communications, and image processing functions. PLDS will proceed through building on existing systems, with the integration of and testing of available, well-understood ("low-risk") technology. Using a science scenario approach employing ongoing research to drive pilot planning and implementation, PLDS is expected to form the basis of a full-scale land data system. Thus PLDS can in turn serve as an information system prototype for the observations from Earth Observation System (EOS), a suite of Earth-observing

spacecraft proposed for the Space Station era (Butler et al., 1984).

6.B.3. Guidelines for Pilots

Pilot programs provide logical ways of learning through experience about the practical problems of developing technology tools in a variety of SSDMU environments. We applaud the strong involvement of the science community in the pilot activities. However, the pilot programs are much more than technology evaluation efforts. The pilots can be used to gain experience in ways to manage data-intensive activities. We suggest that the management experiences gained may be just as important as the technology experiments. We recommend that the pilot programs be structured in ways to ensure that those experiences are recorded and used as guidelines for operational systems.

The pilots should be designed toward specific long-range objectives, such as developing management philosophies and technologies for mission repositories or archives, or developing methods for managing distributed active data base sites. It is mandatory that pilots be developed with close cooperation between ISO and the discipline areas within NASA and with clear directions as to what system or SSDMU environment is envisioned at the completion of the pilot. The discipline offices, if they are to inherit maintenance costs for the operational equivalents of the pilots, should also clearly understand what the financial burdens will be. Thus far, the eventual design goals for the prototype systems, the manner of "handing off" to discipline programs, and the recognition of continuing costs beyond the 5-year pilot periods have not always been clear.

We recommend that the pilots move toward developing management approaches and technology methods that are directed toward realization of a distributed SSDMU approach involving data centers, repositories, and active data bases that are linked by an information network. We also recommend that pilots be initiated for other disciplines and that the pilots focus toward developing information network capabilities to meet geographically distributed systems. The information network development efforts should be concentrated on linking together the three major SSDMU types, with an emphasis on remote

access to directories, catalogs, browse data files, and to data proper.

Selection of participating organizations for pilots should be open to the community by "Dear Colleague" or other informal, but open solicitation routes. Institutions should be selected for quality, diversity, existing capabilities and experience, potential future applications of the developed experiences, and unique or special areas of research, data base possession, or other important attributes.

6.C. EXAMPLES OF EXISTING AND PLANNED SPACE SCIENCE DATA MANAGEMENT UNITS

A description of several existing or planned SSDMUs that we consider to be examples of reasonable ways to meet user requirements is now given. The intent is to provide the reader, through examples, with attributes of reasonable SSDMUs. Examples are chosen from several scientific fields. Some are new developments, and all involve large data sets. In some cases, data repositories are involved; active data bases are involved in others; and in the following case the institution is an archive, repository, and active data base site.

6.C.1. Space Telescope Science Institute

The Space Telescope Science Institute (STScI) represents an important firststep in the implementation of CODMAC's 1982 recommendations. The Space Telescope (ST), as the first more or less permanent observatory in space, has its scientific management assigned to an independent institute, run by astronomers. The STScI is operated by the Association of Universities for Research in Astronomy, Inc., under contract to NASA. The STScI has as its primary responsibility the conduct of the science program of ST, following policy guidelines established by NASA.

As such, the STScI's responsibilities include selecting, funding, and providing technical support to observers and archival researchers; planning, scheduling, and implementing observations; processing, archiving, and distributing data; evaluating performance and advising NASA; and ensuring wide use of ST data.

The STScI will carry out observing proposal solicitation, including educating the community about observing

opportunities, and will set up and implement peer review of proposals and allocation of available observing time. After proposals are approved, the STScI will carry out long-term planning and detailed scheduling of observations, including command sequence generation, guide star selection, calibration activities, etc. It will also carry out the observations, monitoring the instruments and supporting astronomers in real-time activities (target acquisition, real-time data evaluation, instrument parameter selection) at a science control center at the STScI. The science data stream will be sent to the STScI, where it will be edited, calibrated, and archived.

The basic philosophy of user interaction involves astronomers proposing observations, coming to the STScI to carry out the observations with support from the staff, being provided with edited and/or "pipeline" calibrated data, carrying out some amount of interactive data analysis using STScI-supplied software and hardware, and then taking data, intermediate results and possibly software home for further analysis. Perhaps more significantly, the concept of archival research is very much a part of the ST program. It is planned that astronomers can submit proposals to do archival research (data are nonproprietary after a year), and be supported to do this work on the same basis as observers. Thus the existence of a permanent and adequate archive at the STScI is a given. Finally, the STScI has been assigned the traditional NSSDC data curation responsibilities for ST data and must answer public requests for data. It is clear that for ST, the STScI operates in all three modes of data management systems: a data repository, an active data base, and a data center.

The baseline capabilities in the area of data management originally planned (and funded) for the STScI include a pipeline data processing system, a host computer environment, and a tape archive catalogued by a hardware data base management system. These systems were commercially developed via an independent contract. In addition, a set of basic data analysis programs were developed by the STScI itself. Although significant progress has been made toward a documented, transportable, and "user-friendly" and "programmer-friendly" data analysis system, much of the baseline system still consists of moderately machine-dependent, classical software. Although requirements for a state-of-the-art archive system (on-line catalog and possibly data, remote access browse facility) have been generally agreed upon

in principle by NASA, adequacy of funding and dev[illegible] methodology remain in question at this time.

The charter of the STScI is commendable in it dedication to CODMAC recommendations, including insistence on a staff that includes active researchers. The actual implementation of the ST support facilities has, however, suffered to some extent from a residue of the same problems that have always plagued data management systems. For example, although extensive scientific involvement is planned for the operational era (see Table 6.1, 1), the STScI was not established prior to the specification of the bulk of the ground system. Significant problems thus developed due to lack of scientific involvement in areas such as the planning and scheduling system, and the command language for data analysis. Although data formats have evolved toward ease of scientific use (Table 6.1, 3a), ancillary data (3b) remains a problem. Little attention was paid to transportability of software (5). Finally, financial resources for operations and data management activities were often threatened by overruns in other parts of the project (Table 6.1); resources for the archival system, remote access, and data system modifications still need protection.

6.C.2. Space Physics Analysis Network

This system is an effort developed by the Space Plasma Physics Branch of NASA's Earth Sciences and Applications Program Office. It was established by a Data Systems Users Working Group (DSUWG) of that branch, in order to respond to user needs for a space physics analysis network. The Space Physics Analysis Network (SPAN), which is managed by the Marshall Space Flight Center (MSFC), consists of a communications network in a star configuration. SPAN provides computer-to-computer communication, distributive processing, data archiving at the MSFC central node, and standardization of the file structures. Data rates vary from 300 baud to 56 kbps. Through nodes other than the central node, this net is interfaced with other networks, such as ARPANET and TELENET. Researchers are finding that the network greatly enhances the correlative output of the network institutions and has promoted the sharing of software developments. One hundred and six space physics users are currently involved, and plans call for a large expansion in the number of nodes. SPAN is part of the

Data Systems Technology Program (DSTP) and is playing a role in the SPACELAB program. SPAN already provides access to a number of large data sets and with the addition of new nodes will come more and diverse data sets. Clearly, SPAN is a step in the right direction in terms of linking directories, catalogs, data, and researchers together.

6.C.3. Galileo NIMS Experiments

This Galileo Mission includes a Near Infrared Mapping Spectrometer (NIMS), an experiment involving the first use of an imaging spectrometer in deep space. NIMS is on the Galileo Orbiter, which is scheduled to begin observations in 1989 of the Galilean satellites and Jupiter over a 20-month period.

The NIMS experiment will produce images of the satellite surfaces and of the Jupiter cloud-tops at 208 spectral bands from 0.7 to 5 μm. The raw images will have a size of 20 x n spatial pixels with n = about 100,600. The total expected data volume is about a terabit, data useful to a variety of planetary sciences users, including atmospheric science, geology, volcanology, and geophysics.

Three locations in the United States, and one British and one French site, are involved in major way data processing for the mission. These locations nt foci of expertise and (in some cases) techno ince data set users and associated technique deve , along with co-investigators, are concentrated in th centers. The U.S. centers will not only act directly in support of the experiment and the project (Public Information, Mission Operations, Science Data Analysis), but also will be foci for further science efforts associated with these data sets and in support of future missions. Thus these centers have long-term and developmental involvement with the data bases, and are logical centers for management of library and active data base management systems.

Each U.S. center is assigned an area of emphasis: Jet Propulsion Laboratory for reformatting of the data, first-look, archiving; U.S. Geological Survey for geometric calibration and mapping; University of Hawaii for spectral and radiometric calibration and spectral ("image cube") analysis. The European centers will supervise atmospheric studies.

The U.S. centers are to be connected by electronic communication links, and data products will be shared between the JPL Center and other centers as well as between centers. The computing hardware and operating systems are similar but not identical. A tentative decision has been made to have complete copies of the basic data set at each center and pass between centers only the algorithm and parameters needed to recalculate derived data products.

Along with the data reduction directly associated with the Galileo-NIMS experiment, each center will continue science data analysis as part of other research programs. Thus there will be active data sets at the centers as well as repository data sets. These data sets will be available to other users as funds and technology allow. It is not clear yet if and when an archival data set will be generated and where it will reside on a long-term basis.

6.C.4. Planetary Data System

Since the concepts for the Planetary Data System (PDS) have been quite well developed (Kieffer et al., 1984), it is useful to describe them in some detail. PDS is a plan for an aggregation of SSDMUs to archive, distribute, and analyze planetary data. The Pilot Planetary Data System (PPDS) is being used as a means to gain experience in ways to structure a distributed system and to provide a prototype system that could grow into a PDS. The PDS concept is based on a lead SSDMU or SSDMUs, linked to a set of active data base sites, a concept originally developed in a CODMAC summer study (in 1983) that led to this report (see Chapter 4, Table 4.2).

The specific functions of the lead SSDMU or SSDMUs would be largely those of what we term the data center or centers. The functions would include the following:

1. To manage and control active data base sites.
2. To maintain and distribute directories and catalogs.
3. To maintain primary archive for raw data and redundant archives for the active data bases.
4. To interface with individual planetary missions and sometimes allow the PDS to be a data repository for a given mission.

5. To distribute mission data to active data base site.
6. To provide a leadership role in developing and enforcing data format standards.
7. To take responsibility for "standard" software of interest to a wide set of users. PDS should encourage the development of such software in transportable and easily used code.
8. To provide access to accurate supplemental observational data describing the viewing geometry, with the capability to update the data as improved navigational data analysis becomes available.

The specific functions of the SSDMUs that would be locations of active data bases would be as follows:

1. To develop specialized processed data sets to meet the specific research needs at that site.
2. To maintain a catalog of such data sets and a replica of the master catalog from the lead site.
3. To maintain documentation of the processing steps involved in generating the data sets.
4. To provide limited specialized data processing hardware and software to remote users.
5. To provide more involved data processing in a user work station environment.

Both lead SSDMUs (data centers) and active data base site SSDMUs should involve scientists actively engaged in research utilizing the data bases. PDS would have a peer group review panel to provide advice and review of the functioning of the system and to establish criteria for the addition of documentation and data to the system. Security provision for the data bases is clearly an integral part of the system.

The PDS, which was developed by the user community through a series of workshops, is clearly a step in the right direction. We recommend that NASA include funding in the future for phasing-in of the PDS as the Pilot Planetary Data system matures and the appropriate experiences and technologies can be utilized in the operational (i.e., PDS) environment. Further, we recommend that NASA explore incoporation of deep space mission operations and data repositories into the PDS concept.

6.C.5. Earth Observation Data System (EODS)

As noted in Chapter 4, there is a trend within NASA and the space science community interested Earth observations to ask questions that are multidisciplinary in nature and global in scale. At the present time, the land, oceans, atmosphere, and climate research communities are characterized by geographically dispersed users with varying levels of technical sophistication, operating in a more or less independent manner. Satellite remote sensing offers the community interested in Earth observations a unique tool, one that can supply these scientists with large volumes of data of a consistency and scale previously unattainable. Yet, the effective use of this tool has constantly been hindered by the lack of adequate information systems.

The overall goal of the planned Earth Observation Data System is to provide a powerful and responsive system to support earth science research (Butler et al., 1984). EODS would support research that will facilitate understanding of the complex interactions that characterize our planet, through mapping, inventory, monitoring, predicting, and modeling. EODS will provide a mechanism for improving science access to and the sharing of Earth observation data sets, both NASA and non-NASA, and advance processing capabilities and analysis techniques. Characteristics that will be required in an Earth observation include the following:

- An intelligent user-friendly interface that facilitates the ability to use the EODS with a minimum of training and/or understanding of the total system;
- Data management tools that will allow researchers to rapidly review and select relevant science data sets from a variety of geographically dispersed archive sites;
- Systematic archiving and maintenance of space, ground, ancillary, and correlative data under NASA control;
- Access to directories and catalogs of relevant non-NASA data, e.g., data from operational satellites;
- Mechanisms that facilitate rapid access to archived data necessary to conduct Earth observation research;
- Provision of the history of origin, calibration information, quality assessment, and processing that has occurred for all data;

- The ability to have data registered, calibrated, projected, and otherwise modified as a service with minimal scientist interaction;
- The capability to modify, correct, or change data into a format compatible with that employed by the scientist user of the EODS;
- The ability to transfer scientific and technical data among user nodes of the system rapidly and routinely;
- Access to remote computers and peripherals for scientific analysis; and,
- The ability to access software tools that may be resident on a variety of hardware running under different operating systems from other nodes in support of scientific research that would then be accomplished in a local computing environment.

Successful implementation of an EODS with these characteristics can significantly enhance our ability to accomplish Earth observation research. As currently envisioned, EODS would be an information system with a distributed architecture, intelligent attributes, and value-added services. In concept, EODS would support the most technically demanding computer operations with minimal user knowledge of, or experience on, the system. A major goal of the system would be to reduce the information processing burden on scientists without compromising their ability to conduct scientific investigations. EODS would, in fact, be a logical follow on to the Pilot Land Data System.

6.D. SUMMARY OF TRENDS AND GUIDELINES FOR FUTURE SPACE SCIENCE DATA MANAGEMENT UNITS

Based on the data volumes, rates of growth, probable uses of data, and the trends in SDMUs that we deem to be reasonable, we can envision a set of generic functions and responsibilities that should be assigned to various types of SSDMUs. We recommend that the appropriate way to meet the computation and data management challenges in the 1980s and 1990s is by carefully defining the functions of data centers, repositories, and active data bases, and linking them together with an appropriate information network. Figure 6.1 is meant to give a general description of the different functions and responsibilities and interrelationships for SSDMU aggregates that meet user requirements in reasonable ways. The figure is not meant

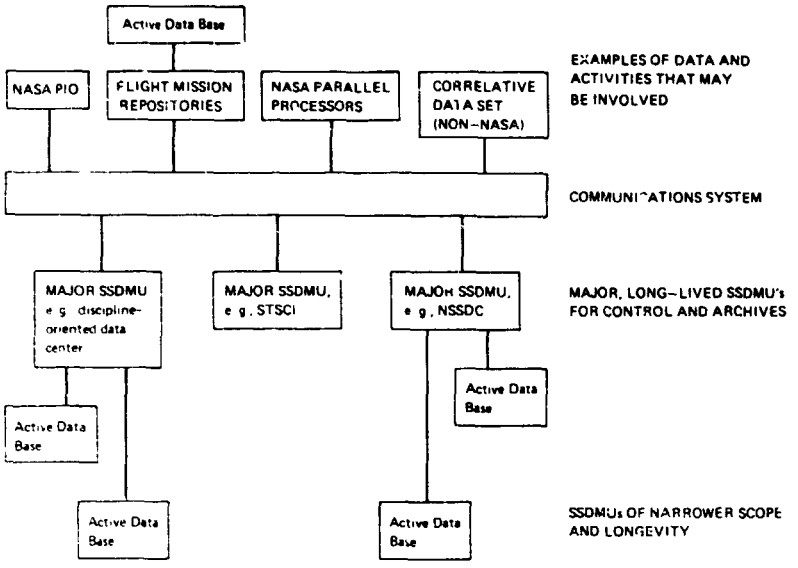


FIGURE 6.1 Functional overview of a distributed information system for the space sciences. Active data base sites are shown as example configurations only as are major SSDMUs. Key to the success is a communications system that is transparent to a science user. Management and control will also be major issues.

to show electronic or communication pathways, although these obviously relate to the functional map. The figure is similar in concept to the functions described in the data analysis network for solar and space physics as outlined in the NRC (1984) report. The levels represent different degrees of responsibility and generalization, permanence, and size. The figure, together with the discussion below, is meant to provide guidelines for implementation of future SSDMUs to meet the significant computation and data management challenges of the 1980s and 1990s.

6.D.1. Data Centers

A connected set of data centers, active data base sites, and data repositories, in the aggregate, form a computation and data management system. In the sections

that follow we offer suggestions as to responsibilities within the system in the spirit of providing a "road-map" to be used in planning, implementing, and operating the distributed computation and data management system for the space sciences. The primary management responsibility should lie in the data centers, which could be structured around major discipline subsets of NASA space science activity, e.g., planetary science or astronomical science. These data centers should have a high degree of permanence of leadership and funding and thus should reside at NASA centers (e.g., NSSDC), JPL, or major institutes, such as the Space Telescope Institute. These data centers should also take a leadership role in terms of arranging for access to other, non-NASA data sets. This access will be crucial, for example, for the Earth Observing Data System discussed in the previous section.

These long-lived sites should have the overall responsibility for the system. They should receive policy guidance from an advisory group composed of representatives of the user community. The management should report to an executive committee that includes the user community, including active data base representatives, and an appropriate NASA headquarters representative who would provide coordination across major discipline areas. A member of this executive committee should represent the system on a NASA standing data management advisory group (see Chapter 7).

The data centers could be responsible for the following:

1. Directories of catalogs. Standards for data catalogs. In addition it should maintain a high-level, low-detail catalog of other relevant systems.
2. All the relevant data sets.
3. Negotiating with the flight projects from the inception of a project for the design of data repositories and transfer of data sets to the archival SSDMU.
4. Setting the standards and qualifying new data sets from PIs and other sources that represent new acquisitions to the data base. These could include nonspace data and catalogs.
5. Determining and negotiating data transfers with non-NASA governmental, private, and foreign institutes in its major discipline area. In addition, tney should determine the usage level and fee structure for such

non-NASA usage, keeping in mind that the primary purpose of the SDMU is to serve the community of NASA PIs.

6. Establishing the suitable level of free access to catalogs and data for the broad NASA and non-NASA community of individual national and international scientists. The level of such usage might be different for different groups of users. For example, message and bulletin board service might be limited to NASA PIs.

7. Serving the needs of the NASA Public Information Office, through which the general public has access to the catalogs, data bases, and data products.

8. Developing and managing an information network that provides remote access, with the necessary bandwidth, to directories, catalogs, browse files, data, and special purpose facilities such as the major NASA parallel processing computers, and selected mission repositories and active data base sites.

9. Providing electronic message and bulletin board service to its user community.

10. Preparing an annual budget to carry out these responsibilities. This budget would not include facilities, hardware, line-charge, and other direct support for PIs. Cost for PIs and the utilization of the SSDMUs beyond the free access level would be determined by the established procedures for proposals. Appropriate units of usage, (e.g., CPU time, file space, hard copy delivery) will be established for use in proposals.

6.D.2. Active Data Base Sites

Active data base sites should be SSDMUs where researchers are actively utilizing a subset of data from a repository or an archive. The sites should be contracted for fixed periods of time for active data base locations so that the number of sites does not necessarily increase with time.

6.D.3. Philosophy of Operation of the Distributed Information System

The data centers and active data base sites, together with the links to flight missions (e.g., repositories) and to specialized NASA computations facilities, form a high-bandwidth systemnet, at least intellectually, and eventually, also electronically. Each site should have

host computers of a size appropriate to their needs, which include providing directories for NASA and authorized non-NASA users in any discipline area and guest accounts for non-NASA no-charge users.

The system should provide access for users by (1) U.S. mail, (2) telephone lines, (3) commercial nets, and, by special arrangments, (4) high-speed lines as appropriate. The communication costs and any special hardware costs should be borne by the users. Since it will be possible for users to enter the system net from the geographically dispersed sites, the individual PI communications costs should be minimized.

A general model for our distribution information system approach is that functions should be pushed to the deepest level (i.e., the most specialized subset of a discipline) so that the nets are most responsive to individual science needs. At the same time, since the trend in some areas of space science is toward multi-disciplinary studies, depending on the interactions of scientists from more than one discipline or subdiscipline, general policies and standards must be set by the data centers, with the guidance of the science data advisory groups, so that the system units can function as an integrated whole.

The funding and management of the data centers and the information network should be separate from missions or science program offices since these facilities transcend any given mission or program. A significant part of the budget should be for overall operations, acquisition of information network general purpose hardware, the development and maintenance of system-level software, and application tools. Much of this tool development might be delegated to specific discipline or subdiscipline units. The net should provide state-of-the-art data base management systems and information processing, but it is not its purpose to do research in these areas.

While it is not the responsibility of the information network to design or manage flight data systems, such design and management should be coordinated within the system from the conception to the completion of the flight project. Then, the proper connections between active data bases, repositories, and archives would be more probable. In fact, this approach ensures that the data centers are leaders in the areas of computation and data management, rather than being the last places that data are placed. This approach should result in more coordination among various aspects of the "data chain" and result in placement of higher quality data in the centers.

7. NASA ROLES IN COMPUTATION AND DATA MANAGEMENT

7.A. INTRODUCTION

Implementing the management and technology recommendations generated in this document should put the space science community in a better position to meet the data management and computation challenges posed by existing and future data sets. However, it is not clear to us that either NASA or the space science community is currently postured in such a way to efficiently implement geographically distributed information systems involving data centers, repositories, and active data base sites. Thus we list the following broad "calls to action" in the spirit of moving to meet the challenges posed by space science data and associated science objectives.

7.B. NASA ROLES IN COMPUTATION AND DATA MANAGEMENT

NASA has a fundamental role in planning and managing the space science and applications research programs on a broad, long-term, disciplinary and interdisciplinary basis. The individual flight missions are not sufficient to achieve the goals outlined in this report. NASA has the obligation to ensure that space science data are collected, safeguarded, and made accessible, and that appropriate uses are made of those data. These uses include both the immediate investigations arising from specific missions, and the broader, longer term uses that result in the development of a cohesive understanding of the state of our universe and the physical processes involved.

7.B.1. Rationale

The data produced as a result of the nation's space missions represent a valuable and often unique resource, and the expenditures of human and fiscal resources to acquire them are large. The analysis of these data is a complex and lengthy process, also requiring the commitment of major resources if the full benefits of the programs are to be achieved. Many of the uses of the data cannot be foreseen in advance. Frequently, new ideas for uses of the data emerge long after the data are acquired, as a result of the continuously evolving understanding of the physical processes under study. This process is both multidisciplinary and interdisciplinary in nature and involves the use of data from multiple sources, acquired over an extended period of time. Even though some of the acquired data may never be fully utilized, it is often not possible to decide in advance which data will be of critical value in gaining future new scientific understanding. NASA has a responsibility for the overall program success, including not only the mission flight phase, but also this long-term creative research process.

7.B.2. Recommendations For Improvement

1. NASA should establish a budget that will provide balanced support, not only for the instrument development, flight support, and immediate post-launch data handling, but also for the information processing, exchange, analysis, archiving, and other related activities required for the longer term purposeful extraction of the important research information content. Experience indicates that the information extraction resources, including data centers, repositories, and active data bases, will need to be generally commensurate with those invested in the instrument preparation and flight support.
2. In its planning, NASA should provide for the maintenance of research, both within NASA and in the universities, which will provide for the continuity of support and stability required to assure the long-term viability of the research programs, including the training of the future space scientists. This requires facilities and funding to provide access to the space science data with SSDMUs and for data processing and analysis.

3. As noted in Chapter 5, NASA should develop and implement specific plans for establishing the technical capabilities required for the efficient and effective maintenance and use of the space science data. The selection of technical approaches should be a matter for active collaboration by the space science community and information system professionals, to assure that the systems are appropriate and adequate for the task, but not beyond the needs.

7.C. THE NEED FOR NASA LEADERSHIP

NASA should exercise strong management leadership in establishing a disciplinary and interdisciplinary approach to space science research that balances the resources among the various components of the activity, with the objective of achieving the greatest return from its investment in the space sciences.

7.C.1. Rationale

Our previous CODMAC document states that NASA's approach to space science data management in the past has been less than fully successful. It is essential that this important area receive more management attention than is currently devoted, particularly in view of the rapidly growing data volumes, the complex user needs, and advances in relevant technology.

7.C.2. Recommendations

1. There should be an explicit, clearly understood assignment within NASA of responsibilities for computation and data management functions to specific offices and individuals. Since the overall responsibility for the effectiveness and productivity of the science and applications programs rests with the Associate Administrator for Space Science and Applications, that individual should take the lead in ensuring that the various functions are clearly defined, and that responsibilities are unambiguously assigned as necessary to accomplish the tasks. Responsibilities shared with or carried by the other associate administrators should be explicitly agreed upon and formalized.

2. The Information Systems Office of the OSSA should have responsibility for activities that bear on the effectiveness of use of space science data. These activities should include the development or acquisition of hardware and software systems; archival, repository, and active data base activities; development of standards; and budgeting and resource control processes as they pertain to computation and data management. That office should manage the data centers and the information networks that connect data centers, repositories, and active data bases. To do these tasks requires an increase in both staff and funding within the ISO.

3. NASA should reemphasize the individual responsibilities of its principal investigators, team leaders, and program and project managers and scientists for their roles in data management, including the depositing of appropriately documented research data in the repositories and data centers. NASA should establish new requirements for the immediate notification of the central data directory of the existence of new data sets resulting from the analysis process. When combined with data centers that are actively involved in the distributed information system involving the centers, repositories, and active data bases, the result should be retention of higher quality, better documented data for use by the broad space science community.

4. NASA should establish requirements for projects to plan for early and adequate funding for the data analysis and archiving functions, including data system, algorithm, software, and hardware development, and should follow up to assure that those requirements are met. It should establish mechanisms (such as associate administrator approval and NASA Advisory Committee oversight, for example) to assist in improved protection of funds allocated for prelaunch development of mission data processing systems, postlaunch data analysis, the archiving of data and related information, and the development of general-purpose or discipline-oriented information systems, against reprogramming as a result of hardware overruns, mission stretch-outs, and other similar competing factors.

7.D. SCIENCE COMMUNITY INVOLVEMENT

NASA and the scientific community need to work together to achieve the common goal--to maximize the scientific return from space science data.

7.D.1. Rationale

Since space science data are a valuable national resource and are acquired at public expense, it is important that they be maintained in a manner suitable for use by the general scientific community. The data management problems described in our previous document (NRC, 1982) are due partly to lack of adequate scientific involvement. NASA has an obligation to properly manage space science data, including providing opportunities for participation by the science community. In turn, the science community has the obligation to follow the rules and procedures established for managing the data, and the willingness to devote the time and energy required to assist actively in the process.

7.D.2. Recommendations

1. NASA should establish a standing data advisory group composed of experienced space scientists (data users), as well as experts in the relevant technologies, possibly as a subgroup within the NASA Advisory Committee structure or as a subgroup to the Space Science and Applications Advisory Committee. This group should advise the administrator's office on matters of data policy, together with computation and data management practices. The range of advice should include data systems planning, operational and institutional arrangements, the collection, storage, and distribution of data, coordinating activities with other organizations involved in the collection or distribution of data, and maintaining the appropriate technologies for efficient data management. It is necessary that this advisory group have access to the most senior level of agency management in order to effectively coordinate the efforts of the diverse activities of NASA, including scientific research, applications development, systems engineering, tracking and communications technology, and computer science.

2. Scientists must accept responsibility for delivering data to the repositories and data centers in a documented, useable form and in a timely manner. Although this has usually been required by NASA contracts, it has not always been done. NASA should be more specific in its contractural requirements in this area, allocate adequate funds for its accomplishment, and take steps to ensure compliance with these provisions. Implementing

our distribution information system approach should help in this key problem area.

7.E. CALL FOR COOPERATION WITH OTHER AGENCIES

Information about non-NASA data bases, and a means for access to those data, should be available to NASA researchers. In addition, NASA's data should be available to support the research programs of other agencies.

7.E.1. Rationale

NASA is not the only source of data and other relevant supporting information for space research. By facilitating the exchange of data with other organizations, NASA improves its capability for multidisciplinary research and improves the ability of the other organizations to carry out their research.

7.E.2. Recommendations

1. As noted in our report, data center directories and catalogs for space science data should include references to related data from other agencies, such as NOAA, NSF, USGS, DOE, DOD, etc.
2. Agencies should coordinate data archive holdings and make data accessible to each other.
3. Agencies should develop coherent cost policies. At present, some agencies attempt to recover some portions of their costs for supplying such data, while others allow free access to the data.
4. Ultimately, agencies should combine their directories and catalogs into a common system, and provide for the smooth exchange or transfer of data and other research products. The overall goal should be to develop a structure that would make the process of locating and acquiring data independent of source.

7.F. RECOGNITION OF DISTRIBUTED COMPUTATION AND DATA MANAGEMENT APPROACH

NASA has the responsibility to ensure that space science data are adequately captured, preserved, and made

accessible, both for the immediate scientific investigations related to specific missions, and for longer term investigations. Innovative approaches must be developed to make these longer term investigations possible. As discussed in numerous places throughout this document, we feel that recognition of the functions of data centers, repositories, and active data bases, together with linking them through an information network, is a necessary first step. Given that recognition, we offer the following recommendations:

1. NASA should make adequate provisions for placing space science and related supporting data into suitable, controlled data centers for long-term retention. All original information should be retained initially until the value of the data can be ascertained. Data at various levels of reduction, combination, and analysis should also be retained in those cases where such data sets may have a general utility. A key aspect is to explore the future role of the National Space Science Data Center (NSSDC) as a data center and a leader of the information system. Should it be a universal data center for space science data, or should there be a series of discipline-oriented archival facilities? Each mode offers certain advantages. In the former case, NSSDC could exert strong leadership over the whole space science computation and data management arena, while in the second case, the archives would presumably be closer to the science community. The particular mode of implementation needs to be explored in some depth, including weighing both costs and scientific benefits associated with different options. As a first step, we recommend that NSSDC develop and maintain space science directories and catalogs, that they be involved in information network implementation, and that they maintain solar and space physics data.

2. In general, active data bases should be established with, or in close association with, active space science research groups. These arrangements should be designed to benefit the researchers by their close association with the holdings. The researcher's direct involvement will assist in keeping the data dynamic.

3. Having determined data availability by use of the directory and catalogs, it should be possible for users to obtain their data from the systems on a time scale and in a form that is reasonably compatible with the nature of the data, available technologies, and the user's needs.

4. Provisions should be made for retaining some portion of the data for an indefinitely long period at data centers. Such provisions should include the copying of data when necessary to preserve them, or when new media offer technical or economic advantages.

5. Acquisition, review, and elimination of data from the data center should be by an explicit, formal process involving participation by the scientists best qualified to judge the future value of the data.

6. NASA should seek congressional authorization to retain funds collected through data sales. This should serve as an added incentive to establish a sound charging policy.

REFERENCES

Butler, D., et al., 1984, Earth Observing System--Science and Mission Requirements Working Group Report, NASA TM 86129.

Estes, J., et al., 1984, Pilot Land Data System, NASA TM 86250, 170 pp.

Kieffer, H., et al., 1984, The Planetary Data System, NASA TM, in press.

NASA, 1984, NASA Space Systems Technology Model, Executive Summary, Code PS, NASA, 278 pp.

NRC, 1984, Solar Terrestrial Data Access, Distribution, and Archiving, Joint Data Panel of the Committee on Solar and Space Physics, Space Science Board, and the Committee on Solar-Terrestrial Research, Board on Atmospheric Science and Climate, National Academy Press, 31 pp.

NRC, 1982, Data Management and Computation, Volume 1: Issues and Recommendations, CODMAC, Space Science Board, National Research Council, National Academy Press, 167 pp.