Data Management and Computation--Volume I: Issues and Recommendations

DETAILS

185 pages | 8.5 x 11 | null
ISBN null | DOI 10.17226/12366

AUTHORS

Committee on Data Management and Computation, Space Science Board, Assembly of Mathematical and Physical Sciences, National Research Council

Data Management and Computation

Volume 1: Issues and Recommendations

Committee on Data Management and Computation
Space Science Board
Assembly of Mathematical and Physical Sciences
National Research Council

NATIONAL ACADEMY PRESS
Washington, D.C. 1982

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.

Available from
Space Science Board
2101 Constitution Avenue
Washington, D.C. 20418

Space Science Board

A. G. W. Cameron, *Chairman*
Shelton S. Alexander
Ralph Bernstein
Daniel B. Botkin
Neal S. Bricker
Richard C. Canfield
Francis S. Johnson
Robert Kretsinger
Louis J. Lanzerotti
Eugene H. Levy
John Lewis
Lynn Margulis
Edward P. Ney
Juan Oró
Laurence E. Peterson
David Pines
Louis Rancitelli
Sean C. Solomon
Stephen Strom
Kip S. Thorne
James A. Van Allen
Robert M. Walker
Carl I. Wunsch

Ex officio

James W. Plummer
Laurence A. Soderblom
Robert W. Rummel

Bruce N. Gregory, *Executive Secretary*

Committee on Data Management and Computation

Ralph Bernstein, International Business Machines Corporation, *Chairman*
Shelton S. Alexander, The Pennsylvania State University
Raymond Arvidson, Washington University
Olen Ely, Teledyne Brown Engineering
Clyde Goad, National Oceanic and Atmospheric Administration/National Ocean Survey
Edward J. Groth, Princeton University
David A. Landgrebe, Purdue University
Richard Legeckis, National Oceanic and Atmospheric Administration/National Environmental Satellite Service
Robert L. McPherron, University of California, Los Angeles
Harvey Tananbaum, Harvard-Smithsonian Center for Astrophysics
Donald L. Turcotte, Cornell University
Thomas H. Vonder Haar, Colorado State University

Ex officio

A. G. W. Cameron, Harvard University
Michael A. Chinnery, Massachusetts Institute of Technology

Richard C. Hart, *Executive Secretary*

Panel on Technology Trends

Olen Ely, Teledyne Brown Engineering, *Chairman*
Mel Harvey, Harris Corporation
Pei Hsia, University of Alabama
George Hodges, Teledyne Brown Engineering
A. S. Hoagland, International Business Machines Corporation
Rein Turn, TRW, Inc.

Foreword

This document is one of a series prepared by committees of the Space Science Board (SSB) that develop strategies for space science over the period of a decade. Several reports in this series have been completed: *Report on Space Science 1975* (Part II, Report of the Committee on Planetary and Lunar Exploration, which covers the outer planets); *Strategy for Exploration of the Inner Planets: 1977-1987* (1978); *Strategy for Space Astronomy and Astrophysics for the 1980's* (1979); *Life beyond the Earth's Environment: The Biology of Living Organisms in Space* (1979); *Solar-System Space Physics in the 1980's: A Research Strategy* (1980); *Strategy for the Exploration of Primitive Solar-System Bodies–Asteroids, Comets, and Meteoroids: 1980-1990* (1980); *Strategy for Space Research in Gravitational Physics in the 1980's*, and *Origin and Evolution of Life–Implications for the Planets: A Scientific Strategy for the 1980's*. Other reports are in preparation.

These strategy reports set scientific goals and priorities intended to maximize the scientific return on the nation's investment in space science. But it is not sufficient to be concerned only with the "what" of space science; the maintenance of a healthy and productive space-science program also requires that we scrutinize and make recommendations about the "how" of space science. The SSB is concerned about many such issues, data analysis, theory, and instrument development, which are essential for the proper pursuit of scientific knowledge.

The present report on data management and computation was prepared in response to our perception that data problems were pervasive throughout the space sciences. The data chain from satellite to ground to preprocessing to principal investigator to reduction and analysis and archiving is central to all

of space-science results. Yet it has suffered from inefficiencies all along the line, ranging from inadequate funding and application of advanced technologies to indifference on the part of management and scientist alike. The present report of the SSB Committee on Data Management and Computation (CODMAC) systematically addresses these issues and makes recommendations for improved treatment all along the data chain.

The report of CODMAC was debated by the SSB at a series of meetings and received general approval by the Board on October 24, 1980. The SSB hopes that the recommendations made herein will be found useful in other areas of science as well.

A. G. W. Cameron, *Chairman*
Space Science Board

Preface

Mankind has entered a second industrial revolution. The first multiplied man's strength and energy by the utilization of machinery that has had a profound and, to some extent, an unanticipated effect on his life and environment. The second industrial revolution is multiplying his mental powers, enlarging his memory, and expanding his control. This is occurring through the use of computers to implement functions that were impossible, impractical, or too expensive only two decades ago.

The influence of computers—on industry, business, science, medicine, and other areas—has made our lives easier and the conduct of our affairs more productive. The influence and utilization of computers in science has greatly enhanced our ability to solve difficult and complicated questions about the nature of our universe. In the area of space science, however, computers have radically changed the very way in which science is done. Space science is wholly dependent on computers because the data acquired from instruments on spacecraft are not only complicated in form but also voluminous. The ability of computers to handle large quantities of data has also given us a major problem: since large amounts of data can be obtained, they are obtained. Torrents of data bits descend upon us from our instruments in space. How do we process the data, store them, retrieve them for scientists to use? This is our theme—how to obtain information and understanding from all of these data.

Charge

A committee of the Space Science Board (SSB) was formed in the summer of 1978 and called the Committee on Data Management and Computation (CODMAC). This committee consisted of scientists from various disciplines with extensive experience and knowledge in space-science data processing and management and in computer science. The scope of the CODMAC considerations encompassed all space-acquired data, whether derived from scientific or applications missions. The SSB provided the following charge to CODMAC to guide it in its task:

The Committee is requested to examine the management of existing and future data acquired from spacecraft and associated computations in the areas of the space and earth sciences and to make recommendations for improvements from the point of view of the scientific user. Specifically, the committee should examine the following topics:

1. Data System Planning: The extent to which the complete data management and analysis associated with a mission is studied at the planning stage and its implementation is tied to hardware development.

2. Preprocessing: Consideration of the degree to which this may take place on board the spacecraft or at various locations on the ground.

3. Distribution of Data: The extent to which data should be examined by "quick-look" techniques (or even interactively) and the methods for improving the times required to supply principal investigators with their complete data sets.

4. Data Standardization and Fidelity: Recommendations for standardization of data formats, data compaction, error correction procedures, and other methods for assuring data quality.

5. Software Development: Recommendations for timeliness in software development, standardization of software procedures, and languages to maximize portability and inheritance in future projects and the improvement of programmer productivity.

6. Distribution of Computational Capabilities: The degree to which scientific users, depending on their computational requirements, can optimally utilize shared or dedicated computing facilities. This question will require an analysis of systems of different scale and should take into account projected trends in costs and capabilities of various systems.

7. Mass Data Storage and Retrieval: Recommendations for improvements in archiving, cataloging, and retrieval of data. This question should examine current practices and should make recommendations for standardization of data-base management systems, for the institutional siting of archives for the scientific user, and for desired technology developments to facilitate these three functions. Attention should be given to the problems of making catalogs useful and to convenient guides to data bases for both occasional and frequent users. Consideration should be given to the expected trends in data storage and retrieval capabilities over the next one to two decades and their impact on future requirements.

8. Interactive Processing: An examination of the variety of ways in which human decision-making may be desirably introduced at various points in data-processing procedures.

LIMITATIONS

In attempting to address the elements of this charge, we have investigated many areas of data systems and computational systems. Although we have tried to be as comprehensive as time and resources allowed, there are several areas of importance that we have not been able to assess adequately. Most notable of these are communications systems and data-base management systems.

Acknowledgments

This report deals with space-science data management and computation. It is the product of the efforts of a number of capable and motivated scientists and engineers. The members of the Committee on Data Management and Computation (CODMAC) generously gave of their time and talents to discuss and evaluate past methods of data handling and computation, the current status of scientific computation, and potential for and definition of future methods and organizational structures for managing scientific data. Credit is also due to the universities, agencies, and industries who lent us the substantial talents of their professional people in this effort.

A great deal of credit is due to Alastair G. W. Cameron, the Chairman of the NRC Space Science Board. Not only did he initially establish the need for this study and define our charge, but also he attended most of the meetings, contributed substantially to the discussions, and provided us with his wisdom and insight.

The interactions with other SSB committees was most helpful in providing CODMAC with data user experience and additional scientific insight and in contributing to a critical review of the concepts and material in this report.

A panel dealing with technology was formed to assemble data and provide guidance on future technology trends. This group prepared a great deal of material and provided the committee with excellent information. Their support is also gratefully acknowledged.

Thanks also are due to various agencies of the U.S. Government, universities, and industries that provided us with data and information to help us in our studies. Worthy of mention is the National Aeronautics and Space Administration, for whom the study is primarily intended and who recognized

the need for more order and organization in the management of space-science data. Individuals who provided substantial amounts of help were John McElroy, Janet Heuser, Erwin Schmerling, Michael J. Wiskerchen, Adrienne Timothy, Ichtiaque Rasool, James Vette, and many others who share our interest and concern with this problem.

Michael Chinnery of the Massachusetts Institute of Technology is chairing a companion study on this problem by the NRC Geophysics Research Board. His liaison, support, and views have provided helpful guidance to CODMAC. The NRC Space Applications Board also provided helpful contributions. The report has benefited from reviews of early drafts by George Ludwig, Richard Marsten, and James Head.

The National Academy of Sciences has provided the environment and resources that was necessary to conduct this study. Richard C. Hart, Executive Secretary for CODMAC, provided us with organizational and administrative support and the benefit of his scientific experience and insight in our deliberations. He had the difficult task of assembling thoughts, notes, reports, and sometimes subliminal ideas into a coherent document, and his support was invaluable. Carmela F. Jackson, SSB Secretary, provided the essential support services and was a most cordial and cooperative person to work with.

CODMAC held two meetings at sites other than at the NAS. Thanks are due to Harwood G. Kolsky and Horace P. Flatt, of the IBM Palo Alto Scientific Center, and to Bruce Murray, of the Jet Propulsion Laboratory in Pasadena, who provided meeting space, personnel, and support at meetings held at their facilities.

Ralph Bernstein, *Chairman*
Committee on Data Management
and Computation

Contents

1. **EXECUTIVE SUMMARY** 1
   - I. Summary of Problems Related to Acquisition, Analysis, and Distribution of Space-Science Data 2
   - II. Findings 5
   - III. Principles for Successful Scientific Data Management 6
   - IV. Recommendations 7
     - Policy Recommendations 7
     - Technology Recommendations 8
     - General Recommendations 9

2. **INTRODUCTION** 12

3. **RECENT EXPERIENCES WITH SCIENCE DATA ACQUISITION AND MANAGEMENT** 15
   - I. NASA Approaches 15
   - II. Science Missions 16
     - The International Sun-Earth Explorer Program 17
     - The High Energy Astrophysical Observatory-2 Mission 20
     - The Viking Mission 24
     - Space Telescope 28
     - Atmosphere Explorer 33
   - III. Applications and Operational Missions 37
     - The Landsat Program 38

	The Geostationary Operational Environmental Satellite Program	41
	The Seasat Mission	43
IV.	Present Data Archives	47
	National Space Science Data Center (NSSDC), Greenbelt, Maryland	47
	EROS Data Center (EDC), Sioux Falls, South Dakota	48
	National Geophysical and Solar-Terrestrial Data Center (NGSDC), Boulder, Colorado	51
	Satellite Data Service Division (SDSD), Washington, D.C.	51
V.	Case Studies of Organizations Involved in Science Data Management	52
	The Small Astronomy Satellite-1: An Organizational Approach with a Single Principal Investigator	52
	University of California at Los Angeles (UCLA) Space Sciences Group	54
	HEAO-2/Einstein Observatory Scientific Consortium	56
	The Space Telescope Science Institute (ST ScI)	57
	The Lunar Consortium	58
	Regional Planetary Image Facilities	59
	A Spectral Data Base for Earth Observational Research	62
	Coordinated Data Analysis Workshops	65
VI.	Conclusions	68
4.	**TECHNOLOGY DIRECTIONS FOR SCIENCE DATA MANAGEMENT**	71
I.	Technology Drivers	72
II.	Technology Perspective	73
III.	Technology Summary	74
	Sensors	75
	Space Data Processing	75
	Space Data Storage	76
	Space Data Handling	77
	Space-to-Ground Communication	78
	Computers (Ground Based)	80
	Ground Data Storage	82
	Data-Base Management	83
	Communications Networks and Distributed Processing	84
	Interactive Processing	86
	Software	87
IV.	Criteria for Technological Implementation	90

5. **RELEVANT TECHNOLOGY PROGRAMS** 92
I. Overview 92
II. Landsat Assessment System (LAS) 93
Introduction 93
Landsat-D System Overview 93
Landsat-D Ground Segment 94
Landsat-D Assessment System Tasks 94
LAS System Design 95
LAS Software 96
Implications for Science 97
III. Satellite Communications 97
Introduction 97
General Description 98
Implications for Science 101
IV. Applications Data Service (ADS) 102
Introduction 102
General Description 102
Implications for Science 103
V. NASA End-to-End Data System (NEEDS) 104
Introduction 104
General Description 105
Implications for Science 106
VI. Space Science Data Service (SSDS) 108
Introduction 108
General Description 108
Implications for Science 109
VII. Jet Propulsion Laboratory Planetary End-to-End Information System 112
Introduction 112
General Description 113
Implications for Science 113
VIII. Conclusions 115

6. **TRENDS IN COMPUTING FACILITIES** 117
I. Introduction 117
II. Decentralized Facilities 119
Pros and Cons of Decentralized Facilities 121
The Future of Decentralized Facilities 123
III. Centralized Computing Facilities 124
Examples 124
Pros and Cons of Centralized Facilities 129
The Future of Centralized Systems 130

IV. Distributed Computing Facilities 131
Examples 131
Pros and Cons of Distributed Computing Facilities 134
Future of Distributed Computing Facilities 134

7. PRINCIPLES FOR SUCCESSFUL SCIENCE DATA MANAGEMENT 135
I. Scientific Involvement 135
II. Scientific Oversight 137
III. Data Availability 137
IV. Facilities 138
V. Software 138
VI. Science Data Storage 139
VII. Data-System Funding 139

8. TYPES OF SCIENTIFIC DATA-MANAGEMENT UNITS 140
I. Introduction 140
II. Principal-Investigator Data-Management Activities 141
III. The Project Data-Management Unit 142
IV. Data-Management Units of Broader Scope 143
V. The Discipline Scientific Data-Management Unit 144

APPENDIX: Present and Expected Data Volumes 148
Summary 164

ABBREVIATIONS USED IN TEXT 165

1
Executive Summary

Since the first satellites had orbited, almost fifty years earlier, trillions and quadrillions of pulses of information had been pouring down from space, to be stored against the day when they might contribute to the advance of knowledge. Only a minute fraction of all this raw material would ever be processed; but there was no way of telling what observation some scientist might wish to consult, ten or fifty, or a hundred years from now. So everything had to be kept on file, stacked in endless air-conditioned galleries, triplicated at the three centers against the possibility of accidental loss. It was part of the real treasure of mankind, more valuable than all the gold locked uselessly away in bank vaults.

ARTHUR C. CLARKE, *2001*

The Space Science Board (SSB) of the National Research Council has had a continuing concern with questions relating to data management and computer utilization in space science. There is concern over problems with early planning of systems for scientific data acquisition, reduction, and distribution; the quality, timeliness, and accuracy of sensor data; the allocation of processing functions on the spacecraft and the ground; programmer productivity; software compatibility and portability; and the cost to the scientific user of acquiring the data. The large amount of data that have been acquired in the past, currently being acquired, and planned to be acquired in the next decade presents a challenge that will require the establishment of principles and organizational, technical, and scientific solutions.

Although future science data management will be strongly influenced by

advances in technology, from both cost and performance viewpoints, this committee believes that the majority of the current data problems are not due to technological barriers. Furthermore, projected data problems can also be solved through employing projected advances in technology, providing that the management of data operations is properly organized.

I. SUMMARY OF PROBLEMS RELATED TO ACQUISITION, ANALYSIS, AND DISTRIBUTION OF SPACE-SCIENCE DATA

In the course of its deliberations, CODMAC identified a number of data-management and computation problems associated with space-derived data. These problems, organized in sequence according to the CODMAC charge, are listed below. It must be emphasized that these problems do not apply uniformly across all missions or all disciplines. Different missions, data centers, and disciplines have had varying degrees of success in approaches to data-management and computation problems. In later sections of this report, we identify specific approaches and attempt to determine the factors that lead to the success or failure of the approach. These determinations form the basis of our recommendations, summarized in Section IV below.

1. In the area of data-system planning, three problems have been identified:

(a) There is commonly a lack of scientific involvement in data-system planning during early mission planning and during the system development phase. Typically, the interdisciplinary nature of data is not fully recognized, and, therefore, data systems are frequently not properly implemented for their actual use.

(b) Generally, data-system and data-analysis activities are not adequately funded. Underfunding results from at least three related causes: when there is insufficient planning in the early mission phases, the required funding will often be underestimated; overruns that occur during mission system development may absorb the funds allocated for data handling and analysis; and because of imperfections in the flight and ground hardware and software, the data processing may be more extensive than originally estimated.

(c) Often, a responsible scientific group for data management during and/or after missions is not clearly identified.

2. Identified problems related to preprocessing (the processing that converts data as received to the form required by the user) are the result of the huge volume of data collected by NASA missions. During 1978, almost 10^{15} bits were returned from spacecraft, and this number is expected to increase further during the 1980's. The problems are as follows:

(a) Preprocessing of all the data is possible with current computer technology, but inadequate planning and funding have prevented this from happening in most cases.

(b) Current capabilities for on-board preprocessing are insufficiently understood and on-board processors are insufficiently developed to reduce significantly the amount of data that must be transmitted from spacecraft.

3. In the area of data distribution, many problems have been identified:

(a) Commonly, there are long delays between the receipt of data on the ground and the delivery of preprocessed data to the user.

(b) Costs of data for some disciplines are frequently so high that the "small" science user cannot afford to acquire all data needed. Unfortunately, it appears that costs of some data (e.g., from Landsat) may increase by a factor of 5 as the program becomes operational.

(c) In many cases, Principal Investigators (PI's) who have been supplied with raw data that they are obliged contractually to return in corrected or reduced form to a data center after some specified period of time do not do so. In those cases where data are returned to the data center, the documentation is often incomplete, resulting in the processed data being unusable for other investigators.

(d) The user community has great difficulty in determining what data are available. The contents of some data centers (whose mission is to provide a data resource to the user community) are not widely publicized and not widely known.

(e) As a result of problems (b), (c), and (d) above, some data centers have become unable to respond to requests for data in a timely fashion.

(f) As a result of the lack of standardization of data formats (discussed below), it is difficult for users to correlate interdisciplinary data and data obtained from multiple missions or sources.

4. Problems related to data standardization and fidelity affect the capability of the scientist to use the data once they have been located and acquired. Many of the needed data are in widely distributed locations and are difficult to access. The identified problems are as follows:

(a) A wide variety of formats are employed. Users typically must devote considerable effort to understanding and/or modifying the formats of data received from data archives. This problem is particularly serious when a scientist needs to correlate data from multiple sources.

(b) Data archives generally contain insufficient or inaccurate information concerning the quality and limitations of the archived data.

(c) Data archives generally contain insufficient ancillary data such as time, attitude, orbit, or sensor calibration data.

(d) Some data are only resident at a PI location and are difficult to obtain.

5. The identified problems related to software development are as follows:

(a) Software is frequently not adequately documented.

(b) Software is not transportable as a general rule. Lack of transportability arises in part from inadequate documentation, but also because insufficient attention is given to transportability during software development.

(c) Current software development methods are costly. Lack of software transportability contributes to this because software is sometimes independently developed several times.

(d) All too often the software development is incomplete at the time of launch of a mission.

6. A problem has been identified with respect to the distribution of computational capabilities:

(a) Currently many scientists must travel to remote locations in order to obtain adequate computational resources to perform their analyses. This is an inefficient system, which reduces scientific productivity because of the inability of scientists to have access to working data files. Scientists perform best in their own environment and must have the capability to perform needed computations locally through either the use of their own computer, a distributed network, or a combination of their own computer with a distributed network.

7. In the area of mass data storage and retrieval, several problems have been identified:

(a) Often data must be purged from the archives in order to make room for current data. In many cases, data have been purged without adequate consultation with the scientific community.

(b) Catalogs associated with data archives frequently do not provide enough information for the interested user to determine whether the archived data will be useful for a particular research project. Also, catalogs are not widely available in many instances.

(c) Usually, data archives do not include an adequate browse capability. Such a facility would allow the interested user, at his home institution, to locate and inspect data sets rapidly and to select those that will be useful for further analysis.

(d) Current mass storage technology is inadequate to store at sufficiently low cost all the data returned by NASA missions. In addition, magnetic tape, the storage medium for the vast majority of the science data in archives currently, has a serious deterioration problem with time, and many of the newer technologies either have known deterioration problems or have not been available long enough to permit an assessment of their potential for storing science data.

8. With respect to interactive processing, several problems have been identified:

(a) There are a wide variety of man-machine interfaces with little standardization in hardware, operations, languages, and algorithm definition. Each time a user employs a new system, considerable effort must be expended in learning the characteristics of that system.

(b) Interactive terminals for software development, program execution, and scientific data analysis are not widely employed because of the continued use of old computer technologies.

(c) Little thought has been given to a dynamic man-machine interface with regard to scientific real-time and interactive control of flight experiments.

(d) The use of artificial intelligence and robotics for Earth-orbit and deep-space missions has not been fully exploited.

9. Finally, two problems have been identified that are applicable to several elements of the charge:

(a) Scientific users of space-acquired data frequently need the same data sets as do operational or commercial users. In many instances this results in high costs to scientists and long delays in obtaining the data.

(b) In many cases current technology is not exploited or implemented in present data systems.

(c) A number of NASA-sponsored programs (NEEDS, ADS, SSDS, JPEEIS, for example) are currently under way that are designed to alleviate many of the problems discussed above. Centralization of data systems seems to be a common theme of these programs. Such centralization has the potential to reduce active involvement by the scientific community significantly. Furthermore, these programs seem to be uncoordinated within the agency, and they seem to be proceeding without regard to developments in the industrial community.

II. FINDINGS

Since the overall objectives of this report are to identify problems associated with the management and manipulation of space-acquired data and to provide technological, programmatic, and organizational recommendations that will result in more scientific return from the data, we offer the following general conclusions. Our recommendations are given in Section IV, below.

1. There are problems with the way data are currently managed. The distribution, storage, and communication of data currently limit the efficient extraction of scientific results from space missions.

2. Technological barriers are not the major impediment to improved data handling. While certain areas of technology will need continued development (notably, on-board spacecraft systems), most of the technology required for

successful science data management either exists at present or will be available in the near future. Nevertheless, although economic factors will continue to impose technical limitations on data management, the current problems are due mainly to the structures and limitations of our institutions and management operations.

3. Data-handling problems can be significantly reduced by restructuring the data chain (from acquisition to analysis) to adhere to principles for successful science data management, as discussed in this report.

III. PRINCIPLES FOR SUCCESSFUL SCIENTIFIC DATA MANAGEMENT

In this section we state several principles on which successful scientific data management must be based. These principles were derived from the experiences described in the case studies of Chapter 3.

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 of data. Such involvement will maximize the science return on both 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.

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.

IV. RECOMMENDATIONS

Recommendations are presented under three categories: policy, technology, and general.

Policy Recommendations

1. Principles for successful data management have been defined above; they address scientific involvement and oversight, the availability of data, suitable processing facilities, software procedures, data storage, and funding. We believe that adherence to these principles will significantly improve the extraction of scientific information from space-acquired data. *We recommend that these principles become the foundation for the management of scientific data.*

2. The most successful cases of extraction of information from space-science activities have had the vigorous and continuing involvement of scientists in planning and implementing the acquisition, processing, archiving, and distribution of data. *We recommend that such active involvement be strongly encouraged and supported in the future.*

3. We define a Scientific Data Management Unit as a group of active scientists and support staff with suitable computational resources at a particular institution. These units can be implemented via a variety of organizational structures—the Principal Investigator (PI) unit, the project unit, or other interdisciplinary units that may transcend either the PI or project units in scope. *We recommend that these units be organized in accordance with the principles and guidelines presented herein. The requirements of individual disciplines, however, must be the prime concern in organizing such units.*

4. *Data-analysis funds should be adequate and should be protected against*

reprogramming as the result of such occurrences as hardware overruns and mission time delays.

Technology Recommendations

5. *We recommend that NASA have an ongoing technology management activity encompassing all areas of data systems.* The activity should be independent of any specific program and should take an overview of technology in order to establish whether NASA's needs for space data systems are being developed adequately by industry, universities, or other government agencies. The program should formulate research and development efforts in those areas where NASA and science or applications users would benefit from new developments. In the course of this determination, possibilities for technology transfer and utilization from and by industry, universities, and other government agencies should be explored to the fullest extent feasible. This program must be broader in scope than current NASA programs and must involve scientific users in order to determine the real requirements.

6. *NASA's approach to technology developments at the component level through the systems level should emphasize the capability to implement new technologies rapidly as these technologies evolve.* Modular architectures with standardized interfaces offer one approach that enables such implementation in response to requirements for growth and flexibility. One specific area that deserves special NASA attention and perhaps funding is the potential for flight use of commercial processors that have been hardened for space and military applications.

7. As new technologies emerge and/or evolve in the areas of processors, memories, computers, communications, and data handling, *NASA should have a continuously ongoing program to test, adapt, and qualify for space application those elements that will advance data management for science and applications.*

8. *Specific implementations of technology that are needed in support of science and applications data management are*:

Asynchronous data-handling systems capable of priority-controlled data acquisition, large buffer capacity, packetized data management, and retransmission capability on demand. (The NEEDS program is addressing this issue. CODMAC concurs that this is a worthy effort.)

High-capacity, all-electronic, on-board data storage with storage capabilities of up to 10^8 bits available to individual PI's, and 10^{11} bits available as a centralized mass storage device by 1985. The capacity available in each category should be capable of increasing by two orders of magnitude by 1995.

Data compression algorithms with adequate documentation that permit scientists to select and use them for special applications.

Centralized on-line mass storage devices with capacities to 10^{14} bits by 1985 for data storage within NASA.

A standard archival storage system that is compatible with NASA, NOAA, and DOI archival data requirements.

Generalized data-base-management software that emphasizes NASA scientific and applications data-base requirements.

9. *NASA should become more active in the area of satellite communications technology* in order to enable the acquisition and distribution of wideband scientific and applications data. Particular emphasis must be given to low-cost, two-way communications, to the handling of multiple wideband satellites in multiple access modes, and to low-cost receiving stations for individual scientists and applications users. As a part of this effort, *NASA should set a goal of increasing uplink capabilities to satellites by an order of magnitude during the 1980's.*

10. *Electronic transfer of data to the investigators should be implemented where economically feasible.*

11. *NASA should begin studies of man-machine interactions, requirements, and needed developments for all phases of scientific and applications data management*, beginning with real-time data acquisition and proceeding to final analysis and interpretation of data. Specific emphasis must be given to the user interface for communication with computers, including voice interaction; to data presentations, both optical and nonoptical; and to analysis and interpretation.

12. *NASA should examine its current technology development efforts to determine whether they are duplicating developments now under way in industry, DOD, or other government agencies.* This activity can be accomplished in conjunction with Recommendation 5.

General Recommendations

13. *We recommend that greater emphasis be given to documentation of space-science and application data to make them interpretable and useful to scientists not directly associated with initial acquisition of data.* Included in such documentation should be information and software to extract physical units from the raw data. Those who gather the data should also be responsible for assessing their validity as part of the documentation.

14. In some cases, data acquired from past missions have not been properly archived. *We recommend that under scientific overseers NASA vigorously pursue the archiving and preservation of such space-science data that should be permanently stored, and data no longer required for future scientific uses should be purged.*

15. *We recommend that more emphasis be given to production of user-oriented catalogs and browse files for space-science data.*

16. Software and related issues are a continuing source of problems to NASA and the science community. Consistent standards for documentation, development methodologies, languages, protocols, libraries, and portability do not exist. *NASA should establish a software organization, possibly within a structure with broader data-management activities, with responsibilities to create software policies and guidelines, to generate technical standards, to monitor enforcement of policies and standards, and to assure the availability of information related to existing software programs.* The organization should address the software issues as a joint effort among scientists within NASA, government agencies, industry, and universities. Also, the resultant standards must be compatible with the activities of international standards organizations. Specific activities within this organization would include:

(a) *The development of software acquisition management guidelines.*

(b) *The establishment of a unified software library* to minimize multiple developments of standard software.

(c) *Concentrated research on software metrics.*

(d) *The establishment of a practice of software discipline* for software developed by NASA and scientists, including, but not limited to, such practices as structured programming requirements and design languages.

17. Space-science processing requires a variety of computational capabilities. Current systems in use consist of both centralized and decentralized facilities, with centralized facilities using primarily large computers and decentralized facilities consisting primarily of minicomputers. At the same time, the sophistication of scientific models are requiring ever-increasing processing power, and the interdisciplinary nature of scientific processing is expanding the needs for access to multiple, remote data sets. Fortunately, technology advances in computers and distributed computing networks are compatible with these requirements. *We recommend that NASA work closely with the scientific community to assure access to adequate computational capabilities, communication facilities and protocols, information directories, and software and format capabilities.*

The Applications Data Service (ADS) and Space Science Data Service (SSDS) programs offer some of the required capabilities. At the same time, these programs may duplicate existing commercial capabilities. The scientific community should participate in the definition and development of these programs.

18. Computing facilities at a number of NASA Centers are a decade behind state-of-the-art systems. *We recommend that computational hardware, software, and interactive terminals be updated more frequently in order to*

keep up with the space-science data loads and developments in computer technology.

19. On-board processing of data will be important in future planetary missions, where telemetry rates constrain the total amount of data that can be returned. It will also be important in future Earth observation missions, where the amount of data collected will be large. *We recommend that greater consideration be given to preprocessing and data compaction schemes and to artificial intelligence and robotics. We further recommend that some degree of on-board preprocessing should also be experimentally implemented and evaluated in selected applications missions, but raw data should be accessible whenever possible.*

20. *We recommend that scientific investigators have access to the raw data from scientific, applications, and operational missions.*

21. *We recommend that NASA evaluate and develop the concept of "electronic browse" capability, allowing users to explore data files via communications links. We recommend that "quick-look" low-resolution data be included for use in electronic data browsing.*

2
Introduction

The purpose of computing
is insight,
not numbers.

R. W. HAMMING (1962)

The purpose of computing
numbers is not yet
in sight.

R. W. HAMMING (1970)

The principal objective of this report is to provide recommendations on data management and computation that will maximize the utilization of scientific data from science missions and the extraction of scientific data from application missions. This requires that data be made available to the scientific community on a timely basis, in a usable format, and at a reasonable cost.

The evolution of scientific disciplines, space technology, and computer technology over the past two decades has led to the current status in which impressive capabilities exist to collect, store, and analyze digital data. There is often, however, a considerable gap between actual practice and a well-conceived and implemented data-management approach based on current technology. Advances in sensor technology, communications, and digital computation capabilities have been rapid. Many more bits of data are currently acquired per year in the NASA program than are utilized. Much of the remote

sensing of the Earth and other planets has evolved from a film-based, manual interpretation technology to a digital, multispectral, multisensor, and multitemporal technology utilizing the full electromagnetic spectrum, with significant computer processing for data correction, enhancement, information extraction, data-base management, and modeling. This transition has not occurred without growing pains. Future programs involving higher resolution and wider spectral range sensors will increase the data acquisition rates by at least one order of magnitude. Significant problems exist in nearly all space programs in information extraction, processing, storage, retrieval, and dissemination. Existing and projected advances in digital processing and data-base technologies suggest that reliable technological solutions exist that will allow an increase in scientific returns from space-derived data.

During the preparation of this report a number of mission case histories have been examined in order to document the successes and failures of data management in the past and to develop recommendations for data management for the near future. Also, development trends over the next decade have been examined for electronic data-distribution systems and for computer hardware and software technology. It became clear during the course of these activities that a maximum scientific return could be achieved by merging modern capabilities for computing and data communications with a vigorous and end-to-end involvement of the scientific community.

We have structured this report so that the essential results of the study are collected in an Executive Summary to make it easier for the reader to obtain an overview of the problems and proposed solutions.

In the following sections we first outline selected case histories of missions, programs, and institutions in order to illustrate the wide range of methods for dealing with space-science data. This is not an exhaustive discussion of all past and present data-management approaches but rather an illustrative set of examples that, we believe, demonstrate the strength and weaknesses of those approaches.

We next discuss the technologies on which data management and computation are based. We have attempted to identify the existing technologies, the ways they are implemented (or not, as the case may be), and the trends for the future.

We reviewed several NASA technology programs that will influence the utilization of space-acquired data for scientific investigation and assess their potential impact on data management.

Three types of computer systems—decentralized, centralized, and distributed—are considered from the viewpoint of space-science data processing. Advantages and disadvantages of each approach are enumerated with specific case studies providing illustrative examples.

Based on the range of experiences with data management and the level of

technology available or soon to be available, we then abstract a set of general principles for successful science data management.

Finally, we end the report with illustrations of how the principles of science data management can be applied in a variety of situations, ranging from the Principal Investigator Data Management Unit to the Project Data Management Unit to a structure that we have termed the Discipline Data Management Unit. In each case, the production of high-quality, usable science data can be shown to be directly related to end-to-end scientific involvement that applies the principles that we have annunciated.

We also note that the Space Science Board has always insisted that detailed internal management issues are properly NASA's responsibility and not an appropriate area for Board recommendations. This report suggests, in broad terms, the need for management attention and emphasis on data-management systems.

3
Recent Experiences with Science Data Acquisition and Management

The greatest minds are capable of the greatest vices as well as of the greatest virtues.

RENÉ DESCARTES, *Le Discours de la Methode* (1637)

In this chapter we review the data-handling procedures for a few selected NASA missions, the operation of several data archives that handle data obtained from space missions, and several examples of consortia and workshops that have been dedicated to processing and analyzing space data. The mission reviews are formulated in terms of the charge to the committee. The missions, archives, consortia, and workshops have been chosen to illustrate the diverse ways in which data are collected, processed, distributed, and archived.

1. NASA APPROACHES

During the past 20 years, NASA has employed a variety of approaches to provide scientists with data and the capacity to analyze them. These approaches have been notable for their differences rather than their similarities, as they have been tailored to the requirements of each mission.

Many NASA projects operate with the Principal Investigator (PI) format. In its simplest form, a scientist provides an instrument that is flown on a satellite. The interaction between NASA and the PI concerns mainly the compatibility of the instrument and its data communication system with the spacecraft. The data from the instrument are provided directly to the PI for

processing and interpretation. Since the scientist has a strong incentive to publish, results usually reach the literature in a timely manner. The scientist has less incentive, however, to archive the raw and partially processed data for other users. In many cases data are not forwarded to data archival centers, or they are forwarded with insufficient documentation.

Some instruments now flown on space vehicles are more complex, and it is difficult for a single PI to develop such instrumentation. Thus, NASA began developing its own spacecraft instrumentation. Often this has been done in collaboration with teams of interested scientists. No standard protocol has developed for the interaction of these teams with the spacecraft project. In some cases, the teams actively participate in the design of the spacecraft system and/or have responsibility for data processing and interpretation.

A major problem with applications missions is that they are often treated as if they require little or no scientific involvement in data processing and interpretation. An example is Landsat, in which a primary output in the past was "photographs" of the Earth's surface. Despite the assumed lack of requirements for scientific participation in the processing and interpretation aspects of applications missions, such participation has occurred on an informal basis and has been beneficial to both the missions and the participating scientists. Improved sensor specifications, development of new uses for the data, and extraction of scientific information are some of the benefits that have occurred.

Many space missions are carried out jointly with other federal agencies. Meteorological experiments on Satellites began in 1958, and since the mid-1960's NASA and NOAA have collaborated in the launch and operation of several generations of operational weather satellites (the ESSA, GOES, NOAA, and TIROS series). The atmospheric science and oceanographic community has found the experiments on these operational satellites to be plentiful sources of scientific data. Firm plans have been made to continue and even increase the use of operational weather satellites for scientific purposes.

We now illustrate the various approaches to scientific data utilization by specific examples from several missions. While these examples do not represent an exhaustive discussion of either project or data-management approaches, we believe that they are a representative sample from which lessons can be learned and conclusions drawn. Both science and applications missions are considered. In one case the mission has not yet been flown.

II. SCIENCE MISSIONS

The science missions described in this section have all had as their primary goal the derivation of scientific information. Their success should therefore be judged on the extent to which they facilitated the derivation of such infor-

mation. In fact, all of these missions are seen as major successes. In the areas of data management and computation, however, each mission could have been considerably improved.

The International Sun-Earth Explorer Program

The International Sun-Earth Explorer (ISEE) program is typical of the traditional NASA approach to space missions. This program consists of three satellites designed to study the solar wind and its interaction with the Earth's magnetic field. One satellite (ISEE-3) is continuously in the solar wind monitoring the input to the magnetosphere, the other two (ISEE-1 and -2) are close together and pass through the magnetosphere on an eccentric orbit. Instruments for these two spacecraft and the separation strategy were chosen with the goal of performing detailed studies of the responses of magnetospheric boundaries to changes in the solar wind.

The ISEE program is typical in the sense that each experiment was provided by a PI who had complete control of his instrument. It is also typical in the manner in which data are handled. Data are transmitted to the ground, decommutated by NASA, and shipped to the PI. At a later time, attitude/orbit tapes are shipped as well. On receipt of the data, the PI processes them in combination with attitude/orbit information to produce various files and displays of calibrated data. The PI and his co-investigators and colleagues then engage in studies of interest to them. If these involve use of data from other ISEE spacecraft, or from spacecraft outside the ISEE program or from ground experiments, special arrangement with other PI's are required. The ISEE project does not formally support this activity.*

DATA SYSTEM PLANNING

The ISEE data system was not planned but inherited from earlier programs with the ground rule that no changes could be made in the system. This led to several data-handling problems.

One problem was the lack of simultaneous data from all three spacecraft. Since the ISEE mission is designed to study magnetospheric responses to solar-wind variations, this is indeed a serious limitation. A second problem results from gaps in the data stream. Data are transmitted in messages of about 2-h duration with approximately 1-min gaps in between. These gaps greatly complicate data-processing programs that use recursive filtering to track spacecraft spin period, spin phase, and instrument temperature, for example.

*An exception is the support of "pool tape" generation discussed in the following Section.

PREPROCESSING

On-board preprocessing is limited to various modes of filtering and averaging. These modes are selected by ground command to optimize the use of the available telemetry bandwidth for specific observations.

On the ground, data are received at the Information Processing Division (IPD) of the Goddard Space Flight Center. Most of the preprocessing performed at the IPD involves decommutation of the data streams from individual instruments and preparation of raw data tapes for shipment to the individual PI's. The management structure at the IPD does not provide incentives for the generation of a complete and ordered data file from a given project; rather, maximizing the number of tapes processed appears to be the primary goal. As a result, processing of ISEE telemetry tapes has fallen behind, and tapes that are delivered often contain large gaps (in addition to the 1-min gaps discussed previously).

The IPD also performs preprocessing of data for the generation of "Pool Tapes," an innovation made by the ISEE project. These tapes contain low-resolution parameters and indices that summarize the observations made by ISEE-1 in the solar wind and ISEE-3 in the magnetosphere. Along with the pool tapes, the IPD produces microfilm plots of the data contained on the tapes. The pool data are made available to interested researchers and serve as a guide to interesting events that may then be studied in detail with high-resolution data obtained from the PI's.

Although the pool data are not of sufficiently high quality or resolution to be used as the sole source of data for a research project, experience with pool data has shown them to be quite useful and demonstrates that some ground preprocessing can be extremely valuable. On the other hand, the fact that data processed by PI's is generally needed to conduct detailed studies indicates that a close interaction between knowledgeable scientists and the raw data is a fundamental requirement of an effective data-processing system.

DISTRIBUTION OF DATA

The distribution of data is accomplished by mailing tapes to the PI's, and, as noted earlier, there is a large backlog of data to be processed and shipped; data that are shipped often contain substantial gaps. One consequence of this traditional distribution system is the loss of data that occurs when an instrument fails in a specific mode and the failure itself is not recognized until later. No on-line, quick-look capability was ever developed by the project.

DATA STANDARDIZATION AND FIDELITY

The ISEE project has not attempted to define standard data formats other than those required to provide investigators with raw data, attitude/orbit in-

formation, and pool data. The project has not imposed constraints on the format of data processed by the individual investigators. Since these investigators are located in a variety of institutions in different countries, there are a large number of data formats currently being produced. This makes it difficult to conduct multiexperiment studies of single events.

SOFTWARE DEVELOPMENT

The ISEE project has not considered the problem of software development at the level of individual experiments. No attempt was made by the project to specify languages or programming documentation standards or to provide software support. Program development prior to spacecraft launch was not supported either by adequate funding or by the provision of simulated spacecraft data. Software is not considered a deliverable item and remains the property of the PI's.

Software for the production of pool data was developed by the project from PI-supplied algorithms. Development of this software took an excessively long time because the development effort was underfunded.

DISTRIBUTION OF COMPUTING CAPABILITY

Computational capability has been made available to the ISEE PI's through provision of funds that have been used to purchase time on "mainframe" computers or in some cases to purchase dedicated minicomputer systems. The latter has been quite cost-effective since the dedicated systems can (and have been) tailored to the requirements of the individual experiments.

MASS DATA STORAGE AND RETRIEVAL

The major data archive for which the ISEE project is responsible is a tape library containing copies of the raw data tapes sent to the PI's. This archive provides the capability to recreate a raw data tape should the tape be lost or destroyed. The calibration software is developed by the PI's on their own computers and is not delivered to the project. Thus, it is difficult for meaningful information to be extracted from this archive by anyone other than a PI.

The ISEE project also maintains an archive of pool data. As discussed above, these data are useful for identifying interesting events but cannot be used for detailed studies of the events.

INTERACTIVE PROCESSING

Interactive processing is not a major element in the ISEE project and is not formally supported. Most ISEE investigators utilize mainframe computers and

The rapid development of high-quality data-processing programs depends on several factors. Most important is quick turnaround. This means both rapid input and output from the computer and short incidence times of jobs within the computer. Also important is access to graphics display devices that enable the investigators to see and interact with their data conveniently.

The High Energy Astrophysical Observatory-2 Mission

The High Energy Astrophysical Observatory-2 (HEAO-2, since renamed the Einstein Observatory) satellite contains the first imaging x-ray telescope to be placed in Earth orbit.

For this mission, investigators responding to the NASA Announcement of Opportunity (AO) formed a team involving scientists at American Science and Engineering [this group later relocated to the Smithsonian Astrophysical Observatory (SAO)], the Massachusetts Institute of Technology, Columbia University, and Goddard Space Flight Center. A single PI was named in the proposal to head the team, and principal scientists were named at each institution. The primary scientific instrumentation developed included a grazing-incidence focusing x-ray telescope, two focal-plane imaging detectors, and two focal-plane spectrometers. The experimenter team was responsible for technical overview of the overall optical system (above plus optical bench, focal-plane transport assembly, three star trackers, fiducial light system, and thermal control system) as well as instrument integration and calibration.

Scientific involvement in the mission planning and prelaunch software development was through the consortium of investigators, with the lead role taken by the SAO group. Initial definition for these activities was included at the time of negotiation of the development (phase C/D) contracts. As described below, the actual scientific activities in the data area were substantially the work of the consortium team and were based on these initial definitions. Since it was recognized by the experimenters and NASA that the mission presented a substantial capability and a unique resource, it was agreed to waive the traditional PI rights and operate the observatory as a National Facility with a substantial guest observer program.

DATA SYSTEM PLANNING

Data system planning was recognized at a relatively low level in phase C/D plans and budgets. The data system development involved an individual scientist who was also responsible for the scientific direction of prelaunch mission operations activities—planning the observing program, commanding of the satellite, real-time and near-real-time monitoring, and contingency planning.

There was some early definition of quick-look and production data flow, an identification of a data processing/computer hardware budget (about $200,000), and the conceptual design of major blocks in the data reduction system 24 to 18 months before launch. These led to the concept of data flow through SAO; the centralization of certain software development and data reduction such as preprocessing, aspect analysis, and basic image reduction; and the wide distribution of detailed scientific analysis. Hardware concerns and mission operations planning completely dominated prelaunch activities, but some prelaunch software development was planned and funds were identified (about $650,000) and protected. As discussed under Software Development below these funds were insufficient to carry out the required developments fully.

PREPROCESSING

On-board preprocessing is minimal, and the overall data rate (6.4 kbps) is relatively low. Individual events are counted and transmitted, with some dedicated hardware to scale rates to handle brighter sources. On the ground, GSFC preprocessing of data emphasizes synchronization checks, timing, decommutation, and concatenation of recorder dumps. SAO generates a standard output file containing the results of scientific image preprocessing, which includes the aspect solution, the detector electrical-to-sky transformation using the aspect data and calibration light data, and the generation of a standard output file. For the MIT spectrometer, a similar preprocessing takes place at SAO, while preprocessing of the Goddard spectrometer data is done by the Goddard scientific group using a small dedicated facility.

DISTRIBUTION OF DATA

The quick-look data represent 40 percent of the total data and are provided to experimenters within 24 to 48 h. The data are used to supplement detailed monitoring of the hardware status and for a preliminary scientific analysis, which is essential for planning follow-up observations within the allowed 30-day visibility window. The data are also used to refine planned exposure times for upcoming sources of similar types. The importance of quick-look data is due to the delay time in the delivery of production data, 4 to 8 weeks at present. The data are delivered from ground stations to GSFC within a few days and require a few days to process at GSFC, but the current backlog, or queue, at the GSFC Information Processing Division determines the 4- to 8-week delay in delivery to the experimenters. Preprocessed images are generally available within 2 weeks of the delivery of the production data to SAO, although delivery of processed data to outside observers has proceeded much

more slowly. This is, in part, because of a desire to obtain a complete data set for outside observers before delivery and in part because of lack of completeness of the software processing system (see Software Development, below).

DATA STANDARDIZATION AND FIDELITY

HEAO-2 has four unique formats, one for each focal-plane instrument. There is little on-board data compaction or error correction, but with the imaging data this is not a serious problem since individual photon detections are counted. During preprocessing at SAO, data compaction does occur in the construction of an image file consisting of individual detections and photon maps, with zero event intervals removed from the file.

SOFTWARE DEVELOPMENT

Planning started about 2 years before launch with one individual involved part-time. Several scientists and programmers became involved in the period 12 to 18 months before launch and formed the basic software development team. Some of these people were involved with the hardware development (and appreciated particular nuances associated with the instrument performance); others were primarily software experts. The software planning and development involved coordinated scientific specifications, informal weekly reviews, formal design reviews, formal schedules and progress monitoring, actual coding of modules, testing with simulated and realistic calibration data, overall system planning, and integration of the individual modules. The overall concepts were sound, but time and manpower were insufficient to complete the software development before launch. At the time of launch the basic processing system and software existed, but software to handle variations and contingencies did not. This was due to underestimation of requirements (particularly software development time and costs by the experimenters) and to inadequate funding of science activities after earlier forced cost reductions in the HEAO program.

The remaining software development proceeded postlaunch, competing with operations, observing program planning, and science activities. This resulted in less orderly development (lower programmer productivity), less testing, and much less documentation. Much of the software is in Fortran and could be used on future programs in x-ray astronomy (although many subroutines are based on the scientific library developed for the Eclipse computers). Documentation is only partially complete, and a more detailed description is being developed.

Planning and funding for the postlaunch phase did not sufficiently delineate requirements for operations, for planning the observing program, for data

reduction and analysis (at SAO and in support of guest investigators), and for modifying and developing new software. Inadequate software development and science budgets resulted from underestimates by the scientific team and pressure in the early phases of the program to minimize costs in these areas. These pressures arose from the desire to maximize the amount of money available for spacecraft hardware development. Postlaunch conflicts among the areas delineated above found NASA less appreciative of science requirements as compared with operations requirements, with which they had greater familiarity and responsibility.

From this experience, it is clear that (1) experimenters need to develop more accurate means to estimate realistic costs associated with software and science activities, (2) science budgets should be established early in a program and protected from spacecraft hardware costs, and (3) NASA should develop a balanced appreciation of both science and operations requirements.

DISTRIBUTION OF COMPUTATIONAL CAPABILITIES

The HEAO-2 program has centralized aspect and image reduction capabilities at SAO based on dedicated minicomputers. The other major members of the investigator team (MIT, GSFC, and Columbia University), each responsible for 15 percent or more of the data, also have dedicated minicomputers. The MIT and Columbia computers utilize hardware almost completely compatible with that at SAO, while the GSFC scientific group does not. The GSFC group, therefore, requires either software conversion or greater processing of their imaging data at SAO. The guest program involved 25 percent of the data and over 400 users in the first 2 years. Guest observers came to SAO to work with their data for 1 to 2 weeks and completed their analyses at their home institution. Facilities are provided at SAO to analyze structure, spectra, and time variability; some programming support is available to develop specialized analysis routines.

A study before launch indicated that dedicated minicomputers (including their maintenance and operating costs) were more cost effective than shared mainframes, partly because they could be run 24 h/day with many of the costs fixed. The HEAO-2 group at SAO has two systems: one for standard reduction and one for user scientific analysis and advanced software development. A side benefit of this configuration is the redundancy provided by the two nearly identical systems and peripherals. An additional important factor with dedicated systems is the ability to determine priorities internally, thereby allowing spacecraft contingencies to be dealt with on an immediate basis. Also, the use of the basic computers, peripherals, and terminals can be planned to meet project priorities, avoiding conflicts with outside computer users.

MASS DATA STORAGE AND RETRIEVAL

Data are stored at SAO on standard computer-compatible tapes. The data rate is such that about four tapes per day are required. Images are maintained on disk (for about 3 months) and backed up on tape. A computer-based logging system provides for record keeping and facilitates data retrieval as well as the observation scheduling activity. There are contractual requirements and agreements to send appropriate processed data to the National Space Science Data Center (NSSDC) at various stages. Data sets will be of manageable sizes, especially compared with earth resources (Landsat) data sets. No specific plans yet exist for describing the data to be sent to NSSDC, for providing software that can be used to access and display the data, or for providing for general community awareness of the existence of the data at the NSSDC. For the period during which SAO is under contract, access to the data will invariably be through SAO and not through NSSDC.

INTERACTIVE PROCESSING

The use of dedicated minicomputers with terminals allows for flexibility in the processing system. The processing can run automatically, thereby minimizing human intervention in routine data reduction, but options allow a variety of aspect processing and science processing to be executed by overriding the automatic mode and interacting directly with the data. The computer also tracks the observing program and monitors the data processing to determine when merging of image segments is feasible. An automatic source-detection capability is applied after data have been merged. Individual scientists can then decide on appropriate further processing. Possibilities include various image displays with manipulation and permanent copy capability, extended source analysis, and spectral or temporal analysis. The approach is to automate as many functions as possible but at the same time to provide for scientist intervention whenever appropriate.

The Viking Mission

The Viking mission to Mars consisted of four spacecraft: two Orbiters and two Landers. The Orbiters carried identical science packages, consisting of a vidicon imaging system, a spectrometer designed to map atmospheric water-vapor abundances, and an infrared thermal mapper designed to measure atmospheric and surface temperatures and, thereby, to determine surface thermal inertias. The primary goal of these instruments was to select a warm, wet location that would maximize the probability that the Lander experiments would detect evidence for life. The Landers also carried identical sci-

ence packages, consisting of two facsimile cameras (stereo), three biology experiments, a mass chromatograph–mass spectrometer, an x-ray fluorescence instrument, a meteorology instrument, and a seismometer. Measurements were also obtained within the upper atmosphere by use of mass spectrometers within the aeroshield.

During the mission, data acquired by the Lander were primarily recorded on tape and then relayed at 16 kbps to an Orbiter for transmission to Earth. The Landers also had a capability of transmitting directly to Earth at about 1 kbps. The two Orbiters and one of the Landers are now inoperative. The Viking 1 Lander is in an automatic mode in which small amounts of data will be transmitted back to Earth about every 9 days. Viking has returned more data than all previous planetary missions combined. As such, the mode of planning for and handling the data load is of interest.

DATA SYSTEM PLANNING

Scientific participation in the Viking mission operated as follows: In response to a NASA Announcement of Opportunity (AO), teams or individuals proposed to participate in development of, and data analysis from, a given experiment. Team leaders were chosen for each experiment. The team leaders, together with the Project Scientist, composed the Science Steering Group (SSG). Through the course of the project, the goal of the SSG was to maintain a direct science involvement in the mission.

In the early phases of the mission, data were received and processed at JPL. As the mission progressed, more and more partially reduced data were shipped to investigators' home institutions for final processing. Problems existed with both situations. For example, Lander image data were initially processed at a "quick-look" facility, the First Order Viking Lander Image Processing (FOVLIP), where digital tapes were also generated. The tapes were hand carried to the JPL Image Processing Laboratory, where final products were produced. The image formats were almost totally incompatible, so that investigators receiving the quick-look and final copies of the images or digital tapes sometimes had a difficult time understanding the differences between the versions.

The prime problem with the data system handling, however, was that sufficient funds were not initially included in project estimates for extended mission operations and data-analysis costs. As a consequence, the Viking Project, on the urging of the SSG, kept returning to NASA Headquarters for an extension of the mission lifetime, along with funds to cover the associated cost. Toward the end of the mission, delays of 6 to 8 months in receiving data products were common because of the minimal amount of money available for producing reduced data sets.

PREPROCESSING

The only significant on-board preprocessing conducted during the Viking mission involved programming the on-board computers as to what instrument data sets to record on tape. For instance, the Lander cameras could be programmed to image in six wavelengths, covering 0.4 to 1.1 μm or to image in a monochromatic mode with four focus settings. The channels used, along with the camera pointing angles, were selectable. As another example, the Lander seismometer could be placed in a "threshold" mode, whereby data were recorded only when a signal above a critical magnitude was received.

On-the-ground preprocessing for image data consisted mainly of noise removal for generation of tapes, together with filtering and contrast enhancement for hard copies of the images. Most of the early image analysis was conducted with hard-copy data rather than with digital data, so these steps were important. The extent of ground preprocessing varied widely for the other instruments.

DISTRIBUTION OF DATA

The mode of Viking data distribution varied with the kind of data. Each investigator team, except the imaging teams, had its own group for reducing data, even during the early phases of the mission. Image data, when reduced, were deposited at the Viking data library. The library then distributed the data to Viking investigators and to the NSSDC, which did not receive digital tapes but only hard copies together with available documentation. Selected Viking investigators did receive tapes, and the U.S. Geological Survey (USGS) in Flagstaff, Arizona, has a complete collection. However, there is currently *no* archival facility for storing image tapes for Viking or for any other planetary mission. When the Viking library became full, older tapes were simply transferred to a warehouse.

Severe problems with the interface between the Viking Mission and the NSSDC were encountered during the first 2 years of the mission. The typical mode of operation was for NSSDC to be sent boxes of image products, Viking Orbiter image mosaics for example, without any documentation. As a consequence, the NSSDC was extremely hard pressed to fulfill its obligations as a depository for these data. The situation became more tolerable after intervention of discipline chiefs at NASA Headquarters, whose programs depended on use of the data. Catalogs (including cross references to ancillary data) currently exist for Lander images; similar catalogs exist for Orbiter image products. Some data for Viking experiments conducted during the early phases of the mission still have not been deposited with the NSSDC.

DATA STANDARDIZATION AND FIDELITY

There are three modes of producing Viking Lander images, each of which has a different format of the data block and different algorithms for processing the data. In addition, each nonimaging experiment team essentially defined its own standards for producing and archiving data. Basically, there was little standardization of Viking products once the data were removed from the telemetry data stream.

SOFTWARE DEVELOPMENT

Much of the software for Viking was written specifically to process Viking data sets. This was especially true of nonimaging data. Much of the software for processing image data was inherited from previous missions. In addition, some specialized image-processing routines were developed to process the Lander stereo images.

DISTRIBUTION OF COMPUTATIONAL CAPABILITIES

At the beginning of the mission, the vast majority of data analysis was conducted at the JPL computing facilities. As the mission progressed, more computation began to be conducted at individual investigators' home institutions. In several instances arguments were made to the Project and to NASA Headquarters that computation costs could be decreased by up to an order of magnitude with dedicated minicomputers rather than JPL mainframe systems. As an example, a dedicated minicomputer system is in operation at Washington University. The cost of the system was about $126,000. Amortized over the 5 years of the Mars Data Analysis Program (a program primarily for Viking data analysis), and including programming and maintenance costs, typical image-processing operations cost an order of magnitude less than at the Image Processing Laboratory (IPL) of JPL.

MASS DATA STORAGE AND RETRIEVAL

The Viking mission continued much longer than expected. As a consequence, the mission was chronically understaffed to archive and maintain the data files. As mentioned earlier, older planetary image digital tapes are stored in warehouses, with little regard to preserving them in an archival fashion. Documentation has severely lagged behind production of reduced data, often severely compromising NSSDC's ability to act as a repository for Viking data sets. No truly mass storage system was used during Viking.

The development of Regional Planetary Image Facilities has alleviated the problem of data retrieval or, more generally, user access to Viking image data sets. Six of these facilities exist around the country, serving as regional repositories for the data sets. Users can examine data firsthand and then order their own subsets from the NSSDC. In addition, engineering data documenting such as image locations and resolutions are computer searchable via interactive terminals. Image displays can be used to examine the images that fulfill search constraints by use of automated microfiche readers. Currently, the readers are being replaced by videodisk players.

INTERACTIVE PROCESSING

The principal example of interactive processing during the Viking Mission was the "quick looks" of Lander images produced immediately after receipt of the data. The quick looks were generated on a software/hardware system, now dismantled, which was called FOVLIP. The intent was to use the image data to search for optimum sites for the surface sampler to acquire soil samples. In practice, FOVLIP was also used to find the sampler arm when it "hung up" in various positions. Such an interactive capability was a crucial element in the successful acquisition of soil samples. In addition, an interactive software/hardware system at JPL was used to process Lander stereo images quickly for the same purpose, i.e., where is the sampler arm, and how should it be commanded to acquire a soil sample?

Unfortunately, other interactive processing took a backseat to these two tasks. Funds were generally unavailable for such processing at IPL. An example of the inadequacies is that it took nearly 2 years at IPL to produce computer-generated mosaics from the Lander images. On the other hand, interactive processing on a minicomputer-based system is the dominant mode for processing Viking data at an individual investigator's home institution.

Space Telescope

In this section, we describe a future mission that is in active development at present. It is an important example because it shows to what extent the management of scientific data has been altered as a result of previous experience.

The Space Telescope (ST) is to be launched in the mid-1980's and will be the first large (2.4-m aperture) astronomical telescope to be placed above the Earth's atmosphere. The ST has many advantages over ground-based telescopes: it will be able to conduct observations not only at optical (visible) wavelengths but also at ultraviolet and infrared wavelengths, where absorption by the atmosphere prohibits or greatly interferes with ground-based observations; background light at all wavelengths will be reduced, and the lack

of atmospheric fluctuations will permit temporal studies in the 1-sec to 1-msec range; the telescope will be above the atmospheric turbulence and will be diffraction limited, thereby providing sharper images and allowing the detection of objects that are 50 to 100 times fainter than can be detected with the largest ground-based telescope.

The Space Telescope Project is a major departure from previous NASA astronomical satellite missions in at least four ways: (1) It is an optical telescope--previous missions have been directed toward wavelengths where the atmosphere is opaque. (2) It has a large aperture–previous astronomical satellites have had apertures of the order of 0.6 m. (3) It will have a long lifetime with Shuttle maintenance or refurbishment as necessary. Current planning for the Space Telescope extends its life through the year 2000. (4) It will be primarily a guest observer facility–previous astronomical satellites have been operated in the PI mode with some (or none in some cases) guest observer participation.

The first two departures will affect the scientific problems to be dealt with by the ST and will ensure the involvement of the traditional (optical) astronomical community in the ST program. The latter two departures affect the manner in which the ST is operated and its scientific data are managed. Because the ST will be operated as a guest observer facility and will have a long lifetime, NASA is following the recommendations of the "Hornig Committee" (*Institutional Arrangements for the Space Telescope*, National Academy of Sciences, Washington, D.C., 1976) in establishing a Space Telescope Science Institute (ST ScI) that will be operated under contract to NASA and that will have responsibility for the scientific operations of the ST. This responsibility includes advising NASA on the need for new instruments or instrument refurbishment; funding of guest observers; solicitation and evaluation of observing proposals; planning and scheduling of observations; execution of science quick-look functions during the performance of observations; science data processing and calibration; science data archiving, cataloging, and retrieval; delivery of data products to users; and limited support for science data analysis. The ScI will interact with the ST through the Space Telescope Operations and Control Center (STOCC), located at GSFC, which will be responsible for spacecraft scheduling, command generation, engineering data management, and overall ST health and safety.

The ST ScI was placed under contract in 1981, about 5 years before launch, and since it will be a new organization, it will probably not become effective in its advisory and planning roles until about 2 years before launch. This has had several consequences for the development of the ST program. First, the hardware and software necessary for the ScI to carry out its functions are being procured under other contracts and are planned to be installed as a turnkey system. Second, the development phase scientific participation

in the ST project has been obtained through the AO process. The ST Project Scientist chairs the Science Working Group (SWG), which consists of NASA scientists and scientists responding to the AO. About half of these scientists are PI's, who, together with other scientists on an Investigation Definition Team (IDT), are responsible for developing one of the five scientific instruments (SI's) and carrying out an investigation with it. One of the six PI's heads the Astrometry Team, which is responsible for advising on the development of the Fine Guidance System (being built as part of the telescope) and carrying out an astrometric program. Another PI represents the European Space Agency, which is funding and developing one of the five instruments and the solar arrays. The remainder of these scientists include Telescope Scientists, responsible for advising on the development of the telescope, the Data and Operation Team Leader, responsible for advising on data management and science operations, and Interdisciplinary Scientists, who provide general advice on scientific matters.

DATA SYSTEM PLANNING AND SOFTWARE DEVELOPMENT

The Science Institute is the natural organization for presenting the needs of the astronomical community to the ST project. However, the late date at which the ScI will be placed under contract precludes it from performing this function, and NASA has used the advice of the development-phase scientists in its data system planning. These scientists have tried to be as representative of the astronomical community as possible, but it cannot be expected that they would take as broad a view as the Science Institute would.

The ST project has been receptive to the requirements of the scientists. A problem arose, however, because the ground-system hardware and software was not being procured by the ST project (which is funded by the Office of Space Science) but by the Missions and Data Operations Directorate (MDOD, which is funded by the Office of Space Tracking and Data Acquisition) at GSFC. The communication between the project and MDOD is via a Project Operations Requirements Document, which contains functional and performance requirements but no implementation requirements. The requirements document is prepared by the project, and then the requirements are implemented by MDOD in any manner they desire as long as the requirements are met and the costs remain within budget. Communication through the requirements document, even when supplemented by face-to-face communication, is difficult at best. For example, the requirements that the ground system be streamlined, use modern technology, and be able to take advantage of technology improvements are implementation requirements, not functional requirements, and as such cannot be included in the requirements document. The ST project is having a difficult time getting MDOD to agree to implemen-

tation requirements such as these. Another problem has to do with minimization of operating costs. Inadequate attention is being paid by MDOD, which is responsible for development costs, while the project will be responsible for operating costs (at least for the science operations portion of mission operations). A third problem has to do with interpretation of requirements. A list of requirements is, in fact, a sterile document and requires considerable interpretation before an implementation can be implemented. MDOD personnel have consistently "overinterpreted" the requirements and arrived at large cost estimates. This led to pressure for wholesale removal of functional requirements from the requirements document. The project resisted this by allowing performance (but not functional) requirements to be descoped to the point where the ST ground system is just barely adequate, and in some cases inadequate, for the tasks it must perform. This strategy was based on the fact that since implementation of the requirements is really not expensive, they can be implemented by the ScI either after it comes under contract or after launch. However, even with the descoping of requirements, the MDOD cost estimates were still prohibitively high. This led to a restructuring of the ground-system procurement, with MDOD procuring only that portion that deals directly with the spacecraft and the ST project procuring that portion that deals with science operations and science data management—the Science Operations Ground System (SOGS). The SOGS consists almost entirely of the hardware and software to be operated by the ScI. As such, it would seem natural that the ScI develop the SOGS. This would lead to more scientific involvement and would probably reduce costs. The project has resisted this idea, and it remains to be seen how the ScI and the SOGS will turn out.

PREPROCESSING

Limited preprocessing functions will be implemented on-board. These will include on-board target acquisition (which involves obtaining a target acquisition image, on-board processing of the image to determine the location of the target, and then repointing the telescope so the target is centered in the aperture of the instrument), averaging of successive detector readouts (with bad readout rejection), synchronous averaging of pulsar signals, exposure-meter control (in which the cumulative exposure must reach a preset level before the exposure is terminated and the next exposure is begun), and error correction encoding (which inserts check bits in the data stream so that the data may be recovered in the event of telemetry errors).

All ground preprocessing functions (with the possible exception of data capture) will be performed at the ScI. These functions will convert the raw data to a standard, calibrated format. Essentially all data will undergo photometric correction, and some images will require geometric correction.

In addition, the ScI will keep copies of the raw data so that observers with specialized requirements can perform unique calibrations or corrections.

DISTRIBUTION OF DATA

All science data generated by the ST will be sent to the ScI, where they will undergo the preprocessing described above. When the preprocessing functions are complete, the data will be available to the guest observer who proposed the observation. In addition, the data will be kept in the ScI archives and, after approximately 1 year, will be available to any interested party. Note that this scheme avoids the problem of the ScI having to persuade the observers to return the data to the archives.

DATA STANDARDIZATION AND FIDELITY

As noted above, all ST data will be processed to a standard level of calibration. At present, it is not known what this level will be nor how successful the processing activity will be.

Some data format standardization has occurred. The project has defined a standard telemetry packet architecture and has defined the contents of some of the ancillary fields in the packet format. In addition, a standard header packet has been defined. There can be up to two instrument-unique formats for the data fields in the packets. Formats to be used during the ground data processing have not yet been defined.

DISTRIBUTION OF COMPUTATIONAL CAPABILITIES

The project has made funds available to the PI's for the acquisition of computational resources. These funds have been used to purchase computing time, shared minicomputer systems, and dedicated minicomputer systems, depending on the requirements of the individual IDT's. These computer resources are being used for software development and will probably be used to support data analysis by the IDT's after launch.

The data-processing system at the ScI is being planned around a network of minicomputers. The system is being sized to handle planning and scheduling, observation support, data reception and calibration, data archiving and retrieval, and limited amounts of data analysis. It is expected that most ST observers will have access to data-analysis facilities at their home institutions. Therefore, the analysis capability is being sized to support only the following: analyses performed by ScI staff; preliminary, quick-look analysis by guest observers; and analyses by those guest observers who do not have adequate resources at their home institutions.

There are no plans to provide remote access to the ScI facilities as recom-

mended by the Hornig Committee. Implementation of such access at a later date, however, is not ruled out.

MASS DATA STORAGE AND RETRIEVAL

The ScI will be responsible for data archival, cataloging, and retrieval. Present plans are that the bulk data sets will be stored off-line on computer compatible tapes. Indices to this archive will be maintained on-line. Thus, it will be possible to search the catalog rapidly, but considerable delays will be introduced before the data can actually be retrieved from the archive.

The archive will not be needed until the mid-1980's at the earliest. It is expected that significant advances in mass storage technology (allowing, for example, rapid on-line access to 10^{13} bits of data—10 years worth of ST operation) will have occurred by this time, yet the project is taking no notice of such advances and is proceeding with planning for the tape-based archive.

INTERACTIVE PROCESSING

The ST project recognizes the value of, and supports the implementation of, interactive processing. It is envisaged that the calibration operations will be performed in an automatic "pipeline" mode with provision for occasional scientist intervention when required. Quick-look operations will be performed interactively, in some cases, in near real time. Analysis will be performed either interactively or noninteractively at the user's option.

Atmosphere Explorer

The Atmosphere Explorer (AE) mission was an innovative step forward in the analysis of satellite data. For the first time considerable thought was put into the design of the data system and a wide variety of data services provided by the project. The mission consisted of three spacecraft designed to investigate the photochemical and energy-transport processes accompanying the absorption of solar ultraviolet radiation in the Earth's thermosphere. The measurement package included 14 complementary experiments measuring atmospheric composition, temperature, flow, and radiation. The spacecraft were placed in low-altitude, eccentric polar orbits, which made possible measurements of the vertical distribution of various atmospheric properties as a function of latitude and local time. This is an example of a project successfully pursued in a (multiple) PI mode.

DATA SYSTEM PLANNING

The AE mission adopted a new philosophy in pursuing its science goals and a new approach to the handling, processing, and analysis of satellite data.

The basic tenet of this philosophy was the belief that all data acquired by the spacecraft must be available to all experimenters for cooperative study. The new approach to data processing utilized a dedicated computer to manage all experimenter data files, to provide computational resources, and to service a network of interactive terminals.

The AE data system was planned in advance and provided by the project. It included a central computer, a data-management facility, and a network of leased lines and remote terminals. The data-management facility included a number of specific types of files and programs for manipulating them, both interactively and through batch processing.

PREPROCESSING

The AE system made somewhat more extensive use of preprocessing than other satellite programs. Data acquired by the experiments were stored in on-board tape recorders and were transmitted to the ground when the spacecraft was over the receiving stations. The data were recorded on tape for backup purposes and relayed by communication lines to the processor at GSFC. Normally, all telemetry data were transmitted from the processor directly to the Operations Control Center computer and the input processor of the AE data system (see Figure 3.1). As data were received, the input processor created a backup raw telemetry data tape. It then extracted attitude and orbit information, which were sent to the attitude/orbit computer for detailed processing and also written on backup tape. In addition, it extracted data critical for spacecraft operation and sent them to the Operation Control Center computer. It next edited, filtered, reversed, time annotated, and time smoothed the raw telemetry data. In the next step the time-smoothed telemetry data were sent to the central processor of the AE data system, as simultaneously a master data tape of time-smoothed telemetry data was created. As data were received by the central processor, they were stored in the on-line data base for future processing. Finally, attitude/orbit data prepared by the attitude/orbit computer and magnetic solar activity correlative data were also entered into the central processor (through tapes and cards) and written in special files.

DISTRIBUTION OF DATA

In the AE data system, data were not distributed to individual investigators but instead resided in the central archive of the system. Limited amounts of data could be transmitted over the 3600-Baud leased lines for processing in minicomputer systems. Usually, however, the data were processed by the central processor and written into Geophysical Units (GU) files and into the Unified Abstract (UA) files. The UA file was a fixed-format file containing data

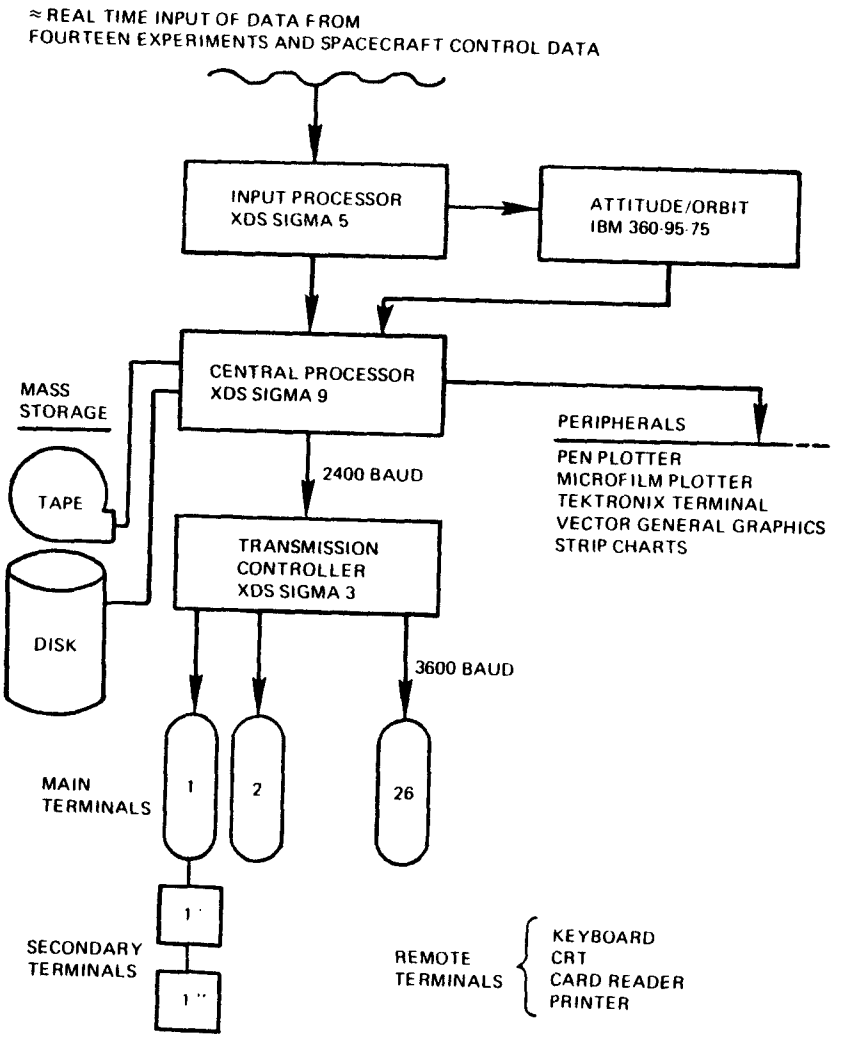

FIGURE 3.1 NASA Atmospheric Explorer system.

from all experiments at a resolution of 15 sec. The file contained, on-line, all data acquired during the mission. The GU file consisted of a number of specialized files designed by individual investigators containing their calibrated data at any time resolution. The GU files could be processed by graphics software to produce microfilm plots of the investigator's data, which were then mailed to him.

For many experimenters, correlative data are needed at much higher time

resolutions than provided by the UA file. To obtain these detailed data, the investigator must use the access subroutines or main programs created by his fellow experimenters. In some cases these are extremely difficult to use. In a few cases, built-in security codes deny access to the data even when access would otherwise be simple.

DATA STANDARDIZATION AND FIDELITY

The AE system made use of a number of specialized files. Most of these were defined by the designers of the data-management facility. These included the raw telemetry data, UA files, spacecraft attitude files, spacecraft orbit files, and magnetic solar activity (MSA) files. Files defined by individual investigator were the GU files. Validity of the data in the telemetry, attitude/orbit, and MSA files was the responsibility of project programmers. Validity of the GU and UA data files was the responsibility of each investigator creating or contributing to a file.

SOFTWARE DEVELOPMENT

The AE project provided most of the software necessary for management of all files except the GU files. The software required to create and manipulate these files was generated by the individual investigators. However, considerable software supporting these files was also provided in the form of subroutines for opening, closing, storing, retrieving, deleting, and merging files, as well as maintaining directory information. No formal arrangements were made to share software between investigators. Software was generally written in Fortran. One limitation was the inadequate support of time-series analysis: the system does not have a general time-series format. As a result, it does not provide access to subroutines that make it simple to obtain any finite interval of data at any time resolution. Also, there are no simple display programs. Instead, it provides general-purpose graphic packages so that each user must continually create his own plot programs.

DISTRIBUTION OF COMPUTATIONAL CAPABILITY

The AE system philosophy required that calibrated data reside in the central data base. To further this end it was required that processing of data be carried out by the central processor. It was also thought desirable to carry out detailed analysis with the central processor as well. To allow investigators to use previously developed software, however, as well as to reduce the load on the central processor, some investigators acquired minicomputer facilities. These facilities have been used in a variety of ways to process and analyze data extracted from the AE system via the leased data lines.

A serious limitation of the AE system is that it utilizes obsolete computers.

Because of this, equipment maintenance is expensive, and upgrading of the system is virtually impossible. Also, no provision was made for eventual transfer of the data base to another computer. Since the data-management facility is machine dependent, it cannot be transferred easily without extensive reprogramming, a task for which no money was made available at the end of the project.

MASS DATA STORAGE AND RETRIEVAL

The AE data system was designed to use magnetic disks as the primary storage and medium for all files. A total of 14 disks, each with 86 million bytes of storage, were provided. Since this memory was not large enough to hold all data from the mission, an additional archive of 1600-bpi computer tapes was required. The data-management facility provided programs for checking a file catalog to determine whether a file was on-line. If not, the system automatically located the tape and issued commands for an operator to mount the proper tape and promote it to disk. Programs were also provided to demote files from disk to tape on request.

A limitation of the AE system is the way in which data sets and disk packs are managed. Because the system is used for so many tasks, the CPU and disks are often saturated. As a result, it has been necessary to institute a policy of automatic demotion of the raw data and GU files each day. Thus, the files of interest to researchers must be promoted to disk each day. Because the access subroutines are based on discrete files rather than time intervals, entire files must be promoted. However, since so many files must be promoted each day, the tape drives are frequently occupied. As a consequence, a researcher must sit at his remote terminal and continuously request a tape drive. Once the tape drive becomes available, the terminal is automatically locked for a time of 3 to 30 min while the tape file is promoted to disk.

INTERACTIVE PROCESSING

The AE system made extensive use of interactive processing in software development and file management. The central feature of this system was the data-base inquiry system that enabled any user to determine the location and status of all data files. Analysis of experiment data, however, was carried out by batch processing.

III. APPLICATIONS AND OPERATIONAL MISSIONS

The programs to be described in this section are examples of applications or operational missions. The primary objective of such missions is not to obtain scientific information but rather to obtain data that can be used on a routine

basis for such applications as weather forecasting or crop-yield predictions. Nevertheless, the vast quantities of data resulting from such missions constitute a valuable scientific resource and should be made available, to the maximum extent practical, to interested scientists. In addition, scientific participation in such missions often improves the effectiveness of the missions in such ways as development of improved algorithms, discovery of new uses for the data, and improvement of sensor designs.

The Landsat Program

The Landsat Program has so far involved three satellites: Landsat-1, -2, and -3 were launched in 1972, 1975, and 1978, respectively; in addition, two more, Landsat-D and -D′ are scheduled to be launched in 1982 and 1983. Landsat-1, -2, and -3 have been first-generation developmental missions. Landsat-D and -D′ are second-generation spacecraft that will be used both for further research and for phasing into an operational Earth observation system.

The first three Landsats employ return-beam vidicon (RBV) televisionlike systems and multispectral scanners (MSS) to generate Earth surface imagery or quantitative data in several spectral bands. Landsat-D will also be equipped with an MSS; in addition, Landsat-D or -D′ will carry an enhanced multispectral scanner known as the thematic mapper (TM).

Landsat data are used in a wide variety of Earth resources disciplines: agriculture (crop, forest and range census, crop yield, vegetation diseases indentification, soil mapping, land-use inventory), environmental tasks, tasks in oceanography (fish production, ship routing, sea conditions), hydrology (water-resources inventory, flood monitoring, pollution monitoring), geology (tectonic feature identification, geologic mapping, mineral exploration, earthquake area studies, glacier and volcano temporal studies), and geography (thematic land-use maps, physical geography).

Landsat-1 data were analyzed by approximately 400 PI's. Subsequent Landsats have used the team concept to conduct various investigations. In addition, data from all three satellites were made available to the general scientific and applications communities.

An important feature of the Landsat program is the large quantity of data generated. The Landsat program currently generates data at the rate of 10^{14} bits per year. These data rates are extremely large by comparison with data rates from typical scientific missions. Such high data volumes place great strain on all aspects of data management and computation associated with the program. It is highly desirable from a scientific standpoint that the same degree of data accessibility and utility be achieved with Landsat data as that required with data from a typical scientific mission.

DATA SYSTEM PLANNING

There was little or no involvement of the scientific community in data system planning for Landsat-1, -2, and -3. The sensor characteristics and ground system were specified without adequate requirements definition or performance specification. The situation is improving with Landsat-D: the scientific community has been involved in sensor specification activities, and a number of studies have been conducted to verify and validate planned sensor parameters. However, involvement of the scientific community in data system planning is still limited.

PREPROCESSING

Electro-optical image-processing technology was used to implement preprocessing functions for Landsat-1 and -2 data. Users could obtain only corrected film products—no corrected digital products were available. Many users therefore implemented their own digital image-processing systems, leading to delay in obtaining results and significant duplication of effort. Digital image-processing technology was implemented for Landsat-3 data, but progress was delayed by multiagency, multicompany implementation problems and incompatibilities.

DISTRIBUTION OF DATA

Initially, Landsat-1 data were distributed directly to the PI's by NASA—a mechanism that worked well as the data were provided within a short period of time and at no cost to the user. Later, the Department of the Interior's Earth Resource Observation System (EROS) Data Center (EDC) assumed responsibility for data distribution. Since that time there have been delays of many months between user requests for digital data products and the delivery of the products. Also, digital data products are now expensive for many users—a single Landsat scene on computer-compatible tapes now costs $200, and this cost will increase significantly when the Landsat program becomes operational. These problems are due in part to the fact that the EDC was not ready to assume data distribution responsibility when it did—the necessary trained staff, hardware, software, and image-processing technology were not in place when the transfer of responsibility took place. Significant delays have also resulted from poor planning of the NASA portion of the ground system. The high cost of digital data is due in part to the large volume of data stored within the EDC and the large variety of data look-up and data product services that the EDC attempts to provide. An interesting phenomenon resulting from the high cost of EDC-supplied data is the growth of an "underground" data distribution system in which investigators who have acquired Landsat

scenes trade copies of the digital image tapes. Data costs are expected to increase by a factor of 5 when the Landsat becomes operational—a serious detriment to the user.

DATA STANDARDIZATION AND FIDELITY

Landsat data formats have changed several times, resulting in user confusion and expensive reprogramming. The significant technological advances to be incorporated into the TM will necessarily result in new standards for preprocessing these data. It is important that these changes be well planned if the value of these new data is not to be compromised.

For some time, the radiometric calibrations were unstandardized. This was due, in part, to instrumentation problems and, in part, to ground processing problems. Only recently have accurate radiometrically and geometrically corrected digital data been available to users. Greater involvement by the scientific community in the early phases of the program might have produced an improved multispectral scanner design and improved processing algorithms, thereby avoiding the delays in reaching standardized radiometric calibrations.

SOFTWARE DEVELOPMENT

Because no projectwide software standards and procedures have existed and because the Landsat program has not funded centralized development of image-analysis software, the software developed for Landsat data analysis has generally been nontransportable. As a result, each user has had to develop or obtain his or her own analysis software. In each case, this software is largely image-processing software, and a large and expensive duplication of effort has occurred.

DISTRIBUTION OF COMPUTATIONAL CAPABILITIES

Each investigator (scientific or application) team used its own resources to process the data. There was no central facility with a scientific or application set of programs that could be used.

One positive beneficial aspect has been that the large amount of data that is being processed is distributed over many locations and facilities. However, only 1-5 percent of the data that have been acquired has been digitally processed.

MASS DATA STORAGE AND RETRIEVAL

As discussed above, Earth resources data, including Landsat data, are stored in the EDC. The EDC provides convenient on-site data look-up facilities, e.g.,

a user can quickly determine what data are available for a region of interest and receive a printout, but there are problems with data product delivery and cost.

INTERACTIVE PROCESSING

There is such a large amount of information contained in an image that extraction of relevant information almost always involves human involvement with the data. Thus, interactive processing has become quite extensive in Landsat data-analysis efforts.

The Geostationary Operational Environmental Satellite Program

The use of data provided by the Geostationary Operational Environmental Satellite (GOES) program is an example of science utilization of data acquired by an operational satellite system.

NOAA's National Environmental Satellite Service (NESS) uses weather satellite data as it flows through an *operational* system. Operations are oriented toward the real-time use of the data and data products in weather forecasting and weather briefing operations, the primary mission of NESS. The 15 years' experience at NESS has also led to a moderately efficient data-archiving system. The sheer volume of digital weather satellite data has been, and is, a problem (more than 2×10^6 bps from U.S. satellites alone).

DATA SYSTEM PLANNING

The data system for the GOES satellite was initially planned at NOAA to satisfy the real-time user requirements. Scientists who participate in projects to improve or to implement operational NOAA projects have considerable involvement in the data system planning. For example, NESS maintains a 24-h current archive of geographically located GOES data that can be accessed from a temporary disk. However, scientists outside of NOAA depend on the digital archive distributed by the Environmental Data and Information Service (EDIS). Therefore, the data system is *not* designed to allow rapid, real-time access to the GOES data for research applications outside of NOAA.

PREPROCESSING

Little or no preprocessing is done onboard the weather satellites. Preprocessing (calibration, duty-cycle expansion, first-cut Earth location, for example) is done at the central satellite ground station in real time, followed by immediate retransmission of the preprocessed image through the "communications" side of the geostationary satellites. The user then receives a preprocessed

image or data set broadcast from a satellite. Both scientific and commercial user response to this NESS activity has been generally favorable.

DISTRIBUTION OF DATA

Because of the need to use weather satellite data instantaneously for forecasting (and to direct scientific research programs in the field) the weather satellites have used a vhf direct broadcast mode since the mid-1960's. Today, this practice continues and is used by many research groups to acquire small batches of weather satellite data as the satellites pass over their local areas. In 1974, with the launch of the first operational geostationary satellite (SMS-1) the communications capability from orbit was used to broadcast weather satellite data at both vhf and uhf (after the preprocessing noted above). Scientific users have been enthusiastic about this method of weather satellite data distribution. Several research groups from DOD, two universities, and three government research laboratories have acquired 5-10-m antennas needed to receive the full-resolution, digital GOES/SMS data. Data from these Direct Readout Satellite Ground Stations (DRSGS), in turn, serve the needs of scientific collaborators at other universities and research groups.

Use of DRSGS to receive data has been shown to be far more cost effective than extracting the same data set from a central archive. These cost savings are used to recover the costs of the antenna, receiver, maintenance, and operations.

In the other problem areas of data distribution (such as the cost of obtaining data from archives, difficulty of correlating two satellite data sets, and long delays) weather satellite users share frustrations similar to those of the users of the EDC.

DATA STANDARDIZATION AND FIDELITY

Since many of the operational satellite data are used quickly for forecasting, their detailed calibration and verification are often not up to scientific standards. Thus, scientific users develop their own special calibrations and quality assurance checks. Some standardization of their methodologies has occurred through sharing of programs, but more coordination in this area would be useful and could help to reduce research costs by eliminating some duplication of effort.

SOFTWARE DEVELOPMENT

As in the area of data standardization and fidelity, little coordination has occurred in the area of software development. Although some transportability is introduced by the common use of Fortran, the design of many data-

processing programs depends heavily on the available system hardware and software, so only limited sharing of software has occurred. Greater coordination in this area could help to reduce duplication of effort and possibly reduce research costs.

DISTRIBUTION OF COMPUTATIONAL CAPABILITIES

No organized or sponsored effort to put more computing capability into the hands of scientific users has existed. Individual groups have begun to acquire and use large and small minicomputers with success. These machines are especially useful in connection with the reception of the broadcast satellite data and with interactive processing.

MASS DATA STORAGE AND RETRIEVAL

Two technological experiments have been successful through their preliminary stages. NOAA/NESS uses an Ampex terabit memory system to store NOAA satellite data. The University of Wisconsin has modified a Sony video recorder for mass storage of digital GOES data and finds great economics over tape storage. NESS, Colorado State University, and others have purchased the Wisconsin device.

NOAA publishes catalogs for the data it archives and keeps all data that enter the archive in digital form (only a sample of the GOES data are included) and as photographic images.

INTERACTIVE PROCESSING

Several groups have pioneered the use of interactive processing of weather satellite image data. Although many advantages are recognized (for example, satellite data can be mixed with other weather and radio data for analysis), there has been some reluctance on the part of funding agencies to support these activities.

The Seasat Mission

Seasat was a "proof-of-concept" mission that grew out of interest in the application of satellite altimetry to ocean circulation studies. A 1969 conference developed a program of Earth and ocean physics that was based on the measurement of distance (or ranges). This program split into the present geodynamics program and Seasat. From its inception in the early 1970's, the Seasat mission focused on the application of satellite altimetry.

By the spring of 1978, a number of individuals had banded together as a "Seasat User Working Group." Their interest extended beyond altimetry and

included wind and wave fields, imaging radar, and sea-surface temperatures. These individuals came from NOAA, the Navy, NASA, the Defense Mapping Agency, and the academic community and desired Seasat data for uses ranging from daily, routine production of wind and wave fields for operational use to the satisfaction of scientific curiosity about the interaction between the air and the sea.

To serve these interests, the planned set of instruments included an altimeter, a scatterometer (SASS), a passive microwave radiometer (SMMR), an infrared radiometer (IR), a synthetic-aperture radar (SAR), and, finally, global-positioning-system transponders (which were later deleted).

DATA SYSTEM PLANNING

The investigators for Seasat were not selected in the usual NASA manner (i.e., via Announcement of Opportunity). Rather, the Seasat project recruited the participation of individuals of known expertise. These individuals were organized into a user working group, concerned with scientific or operational applications of the data, and instrument working groups, concerned with the evaluation of the instruments.

One of the project ground rules was that users were expected to pay for most of the data analysis with funds obtained from other (i.e., non-Seasat project) sources. Despite this lack of financial support from the project, the working groups were generally enthusiastic and gave much support to the project both before and after launch. The working groups, however, had no real authority and functioned only as advisory bodies.

Separately from the Seasat project, a joint NASA/NOAA Announcement of Opportunity solicited the participation of individuals in scientific investigation using the Seasat data. About $4 million, over 3 years, was to be provided to nonfederal investigations to use Seasat data, and some 30 investigators were funded. These scientists were promised Seasat data by 60 to 90 days after launch, which was much earlier than the project had ever hoped to deliver data. This situation was exacerbated by problems in both algorithm development and production data processing that delayed delivery of the data products for about 9 months beyond the original project schedule.

PREPROCESSING

The only on-board preprocessing that was planned was the extraction of wave heights from the altimeter data. As it turns out, this capability was never used, and all preprocessing was done on the ground.

The ground data system for Seasat was really two separate systems–one for the SAR, the other for the remaining four instruments. Because of the high data rate, SAR operation was scheduled while the spacecraft was within

sight of a ground station. The resulting data were relayed directly to the ground station, where they were recorded on high-density tape recorders. These tapes were then shipped to JPL for processing. JPL had planned to process some 2600 minutes of data (a few hundred images) during the first year after launch. Equipment problems reduced this to about 500 minutes—most of which was done in the last 3 months of the year after launch.

The system at JPL for SAR processing was an optical correlator originally built for processing aircraft-acquired instrument data. Substantial modifications were required because of the different geometry and characteristics of the satellite system. The original $1.1 million requested for this modification was finally reduced to $0.6 million, and work was delayed because there had been hope that much more substantial funding might be available for the development of more expensive but more accurate digital processing. By a year after launch there were still no funds available for such a development.

The data system for the other instruments used on-board tape recorders and the new capability for centralized processing at the Information Processing Division (IPD) at GSFC. Data were recorded on board, dumped to a ground station, and relayed to GSFC for data capture. The IPD produced two kinds of tapes: "a project master data file" and an "altitude and orbit file." The data in the two files were shipped to JPL and combined to give an Earth-located "sensor data file," and geophysical data were calculated and stored in a "geophysical data file." The last two steps were to be initiated as algorithms improved, eventually leading to the production of a complete geophysical data file generated with the final algorithms.

The plan was for IPD to produce tapes within 24 h of receiving the telemetry data and for JPL to produce sensor data files within 48 h of receiving the IPD tapes. Problems with the IPD systems delayed all of this, and delivery of the last of the tapes to JPL occurred 6 months after the satellite failed. JPL produced sensor data files within 6 weeks of that delivery.

When the Seasat satellite failed in October 1978, after 3 months of operation, the project was reorganized into a "Seasat Data Utilization Project." The scope of the new project was the development of the final algorithms, the production of a complete set of geophysical data (if appropriate), and the comparison of Seasat data with several major surface oceanographic investigations. The project was also to host several workshops related to the surface investigations and make useful geophysical data available through EDIS as rapidly as feasible even if in preliminary form (i.e., with less than perfect algorithms).

DATA DISTRIBUTION

Distribution of Seasat data to the scientific users was planned through the EDIS of NOAA. The first data arrived at EDIS in May 1979, almost a year

after launch. By 14 months after launch, only the following data were available at EDIS: (1) a complete sensor data file from the altimeter, (2) twelve continuous days of altimeter interim geophysical data, (3) selected areas of altimeter geophysical data, (4) two complete days of SASS interim geophysical data, (5) selected areas [the same as in item (3)] of SASS interim geophysical data, and (6) 130 passes (2-10 min each) of SAR data. Because of continuing problems with algorithm development, no SMMR data were available 14 months after launch.

User access to Seasat data through EDIS has not yet been tested, although it is believed that such access will be relatively straightforward and convenient.

DATA STANDARDIZATION AND FIDELITY

Seasat data will eventually be of established quality, through comparison of the data with surface truth.

No data format standards have been applied, and it is not known to what extent this lack of format standardization will affect the eventual use of the data for scientific investigations.

SOFTWARE DEVELOPMENT

As can be seen from the above discussion, development of preprocessing software has taken much longer than originally planned, primarily because this activity was not adequately funded until the satellite failed. A few individuals from the working groups did aid JPL in the development of algorithm software.

DISTRIBUTION OF COMPUTATIONAL CAPABILITIES

JPL made available time on its general-purpose computer (Univac 1108) to allow algorithm development at the investigator's expense. This off-site processing capability was, however, rarely used.

MASS DATA STORAGE AND RETRIEVAL

All data were stored on computer compatible tape at JPL. In a few instances, individuals were able to access Seasat data for algorithm development via remote terminals.

INTERACTIVE PROCESSING

As just stated, except for a few individuals within the working groups who used the JPL central computer for algorithm development, no interactive processing was performed.

IV. PRESENT DATA ARCHIVES

The National Space Science Data Center (NSSDC) of NASA and the EROS Data Center (EDC) of USGS are the primary archival centers for the scientific data collected from spacecraft. Such archives ensure that the data sets are preserved for future use. Equally important goals of the data centers are to catalog and distribute data for use in the near term by both scientists and other users (e.g., operational agencies, commercial vendors, and educational groups).

The NSSDC and EDC, along with a variety of smaller, specialized data archive centers have struggled in recent years under the threefold burden of increasing amounts of space needed and Earth science data, limited budgets, the need to convert to new technology.

The burden has been partially alleviated by NOAA, which has taken on archival and distribution responsibility for a major portion of satellite data pertaining to the atmosphere, the oceans, and other Earth-physics data through the National Geophysical and Solar-Terrestrial Data Center and the Satellite Data Services Division of the EDIS.

National Space Science Data Center (NSSDC), Greenbelt, Maryland

The National Space Science Data Center (NSSDC) was established by NASA to provide data and information from space-science experiments in order to support additional studies beyond those performed by PI's. NSSDC produces catalogs, users guides, and reports on active and planned spacecraft and experiments and also maintains a staff to interact with and support users.

NSSDC's goals are to further the widest practical use of reduced data obtained from space-science investigations and to provide investigators with an active repository for such data. NSSDC is responsible for the collection, organization, storage, announcement, retrieval, dissemination, and exchange of data received from satellite experiments, sounding-rocket probes, and high-altitude aeronautical and balloon investigations. In addition, NSSDC collects some correlative data, such as magnetograms and ionograms, from ground-based observatories and stations for NASA investigators and for on-site use at NSSDC in the analysis and evaluation of space-science experimental results.

Users can obtain data from NSSDC by a letter request, telephone request, or on-site visit. The user specifies the NSSDC identification number, the name of the satellite/experiment, the form of the data product (film, hardcopy, color, size, or other) and the time or location desired. The user also specifies why the data are needed, the subject of his work, his affiliation, and any government contracts he may have for performing his study.

NSSDC provides a wide variety of data catalogs and documentation to assist users:

1. Data Catalog of Satellite Experiments--divided into eight discipline categories--astronomy, geodesy and gravimetry, ionospheric physics, meteorology, particles and fields, planetary atmospheres, planetology, and solar physics.
2. *Report on Active and Planned Spacecraft and Experiments.*
3. Lunar and Planetary Photography Catalogs and Users Guides.
4. Meteorological Data Catalogs and Users Guides.
5. *Handbook of Correlative Data.*
6. Spacecraft Program Bibliographies.
7. Reports on Models of the Trapped Radiation Environment.
8. World Data Center-A for Rockets and Satellites Catalogs of Data
9. WDC-A-R&S Sounding Rocket Launching (SRL) Reports.
10. SPACEWARN Bulletins.

NSSDC provides facilities for reproduction of data and for on-site data use. The Data Center staff will assist users with data searches and with use of the microfilm reader and light table.

The NSSDC also uses a minicomputer system for internal functions, such as logging and tracking user requests, providing accounting statistics, duplicating digital tapes, generating documentation listings, and supporting special projects.

NSSDC provides data products in the following basic forms: hardcopy (books, bound volumes or pages), digital magnetic tape (reels), microfilm reels (35 mm, 16 mm, and other sizes), microfiche cards, photographic film (color or black and white negatives and positives, and slides), strip or brush charts (rolls in 35 mm or other sizes). Unfortunately, magnetic tapes of lunar and planetary image data are not available from NSSDC.

NSSDC provides data products free of charge as long as the request is moderate in size. Large data-set users are asked to fund these requests at cost.

Reactions of science users to NSSDC are mixed. The cataloging, search methods, and distribution of some data received general praise. Delays in receiving data, however, were often encountered, and sometimes these were an impediment to science analysis (this was sometimes because of the reduced data not being submitted by the PI). Probably the most serious complaint about NSSDC is that a significant fraction of the sensor data is not adequately documented. When data are made available to an inquiring scientist, they are not always in a usable format. In some cases the problem lies with an inadequate staff at NSSDC; but in most cases the data provided to NSSDC do not have the necessary documented software to calibrate the data and to provide necessary ancillary information.

EROS Data Center (EDC), Sioux Falls, South Dakota

The EROS Data Center (EDC) is operated by the Department of the Interior (DOI) to provide access primarily to NASA's Landsat data, aerial photog-

raphy acquired by the DOI, and photography and imagery acquired by NASA from research aircraft and from Skylab, Apollo, and Gemini spacecraft. EDC provides data archiving and generation of film products and computer-compatible tapes (CCT) for users. EDC has a central computer complex that controls a data base of over 6 million images and photographs of the Earth's surface features, performs searches of data on geographic areas of interest, and serves as a R&D tool for image data reproduction processes. The computerized data storage and retrieval system is based on a geographic system of latitude and longitude, supplemented by information about image quality, cloud cover, and type of data.

Primary input media for EDC are high-density tapes (HDT's) generated by NASA GSFC. These tapes contain two types of data: those that have been radiometrically and geometrically corrected with resampling of the data to fit a known map projection and those that have been radiometrically corrected but have not been geometrically corrected or resampled.

In addition to mailed HDT's, a Domsat link has been established at EDC. The NASA Landsat ground stations at Goldstone and Fairbanks transmit raw Landsat data to GSFC via Domsat. GSFC then preprocesses and reformats the data (HDT) and retransmits these via Domsat to EDC, where they are further preprocessed.

The Image Processing Facility at GSFC is the first stage in the preprocessing of Landsat data. The HDT's are classified by two major image data categories based on sensor type and on whether the image data have been geometrically corrected. All HDT's contain data that have been radiometrically corrected.

Data received at EDC are registered by the Inquiry, Order and Accounting System (INORAC), which provides quality assessment and initial cataloging of the data. INORAC also generates "work-order cards," which specify desired image-enhancement parameters (haze removal, contrast stretch, or edge enhancement) and type of output desired (film, CCT, or special order).

Based on these work-order cards, the EROS Digital Image Processing System (EDIPS) extracts digital image data from the HDT's, enhances the imagery, and records the resulting processed data on high-resolution film or CCT. Standard format processing consists of the following steps performed under automated control:

1. Initiate and select the tape readout using TRIG time on the tape.
2. Generate histogram data.
3. Generate annotation data.
4. Provide automatic haze removal.
5. Provide automatic contrast stretch.
6. Generate film product.
7. Generate inputs to the EDC main image file.
8. Generate processing status data file.

Special orders need individual work-order cards for each scene to be processed. Special format processing includes generation of products in which the processing steps are modified, rearranged, or deleted.

The EDC also maintains an R&D Data Analysis Laboratory (DAL), designed to provide both digital and analog multispectral/multitemporal image-analysis capabilities in support of all technology-transfer programs, with prime emphasis on federal government agencies. DAL capabilities include development of digital classification methods and interactive thematic extraction techniques. Additional R&D support is provided by the USGS Flagstaff facility.

EDC provides standard and special-order processing. Standard "pipeline" processing turns HDT inputs into 241-mm film products. Special-order processing transforms correction and uncorrected HDT's into 241-mm film and CCT's. Duplication of HDT's and CCT's is also provided.

EDC also provides training and conducts workshops on remote sensing. The scientific teaching staff offers discipline-oriented courses in agriculture, forestry, geology, and hydrology. Assistance is provided to users in the operation of equipment such as densitometers, additive color viewers, zoom transfer scopes, stereo viewers, and in the use of computerized multispectral systems to classify specific phenomena.

User comments relative to EDC operation range from very positive (the search system) to very negative (delays in availability of data and its cost of $200/scene for CCT data). Because of its primary role as a data center for application-oriented users, EDC has been required to recover its costs of services. Thus, science users may be limited in the number of data sets that they can order because the volume of digital data that they require could become extremely expensive.

Many of the shortcomings of data availability via EDC result from the fact that there are two different agencies at two different sites involved in preprocessing. Long delays in the data stream have occurred at various times in the past at each of the two sites. Not only must direct functions such as hardware operation and software development be successfully pursued at both sites simultaneously but also indirect functions such as planning, securing funding, adoption of standards, formats, and policies.

A number of the difficulties of delay and cost associated with Landsat data distribution may be related back to the early decision to produce image products (photographic) primarily, rather than digital products (quantitative data). The initial selection of an electro-optical preprocessing approach for Landsat-1 and -2 rather than a digitally based system at NASA/GSFC is an example of this. A similar orientation at EDC has also existed. This has resulted in a difficult, slow, and expensive effort, particularly now that the need for a more digitally based orientation has become apparent.

National Geophysical and Solar-Terrestrial Data Center (NGSDC), Boulder, Colorado

The Environmental Data and Information Service (EDIS) in Boulder, Colorado, is a branch of NOAA and has a center called the National Geophysical and Solar-Terrestrial Data Center (NGSDC), which conducts a national and international data and data information service in all scientific and technical areas involving solid-earth geophysics, marine geology and geophysics, the upper atmosphere, the space environment, and solar activity. These services are provided for scientific, technical, and lay users in governmental agencies, universities, and the private sector. NGSDC handles over 7000 requests for data annually. NGSDC hosts visiting scientists, who are paid to come to NGSDC to study the data in its data base and to perform data validation, "cleaning," and documentation.

Two services provided by EDIS are the Environmental Data Index (ENDEX) and the Oceanic and Atmospheric Scientific Information System (OASIS), which provide to users rapid, computerized referral to some available environmental data files (ENDEX) and published literature (OASIS) in the environmental sciences and marine and coastal resources. ENDEX data bases are computer-searchable, interdisciplinary files of environmental data that can be searched by geographic area, by the parameter measured, or by the institution holding the data, for example. ENDEX has three major components: descriptions of data-collection efforts; detailed inventories of large, commonly used files; and descriptions of data files. ENDEX data bases are updated every 2 years. Most of the data bases currently deal with oceanographic data files.

One of the major problems with the effectiveness of the NGSDC is that the timeliness of its data is determined by the efficiency and willingness of the diverse groups that contribute to it. The result is that there can be many months or even years of delay in the receipt of data by NGSDC, and this prevents the service from being a near-real-time data service at present. Hence, many researchers bypass it and go directly to the appropriate PI's or agencies who are producing data that they require.

Satellite Data Services Division (SDSD), Washington, D.C.

The Satellite Data Services Division (SDSD) of the EDIS National Climatic Center is the U.S. archive for environmental satellite data. While primarily intended for meteorological data, SDSD provides data of value to hydrologists, agronomists, oceanographers, and geologists.

The types and quantities of data held are too numerous to describe completely. However, the files contain photographic images, digital tapes, and derived products from all environmental satellites. The photographic products

include 35-mm, 70-mm, and 25-cm negatives from both the polar orbiters and the geostationary satellites.

In addition to maintaining satellite data and reproducing these data for users, SDSD employs data-processing specialists, meteorologists, and oceanographers to assist users in selecting the correct data for their specific investigation, produces special products when required, and assists in analysis of the data.

Costs for each type of satellite data vary according to product type, size, and quantity. Generally, the cost for a 25-cm (10-in. x 10-in.) black-and-white contact print of satellite imagery is approximately $3.50 per copy. Digital-tape products generally cost about $60 per tape.

V. CASE STUDIES OF ORGANIZATIONS INVOLVED IN SCIENCE DATA MANAGEMENT

From the mission case studies presented in Sections II and III of this chapter, it is clear that the most successful scientific use of space-acquired data occurs when interested scientists are actively involved in all the elements of the data chain: planning, collection, processing, archiving, distribution, analysis, and publication. The examples presented, however, demonstrate a wide variety of structures for obtaining the necessary scientific involvement in the data stream. In this section we present several case studies covering a range of organizations (some taken from the mission case studies and some new examples) in which data are managed in a fashion that promotes high-quality scientific results. We also attempt to identify problems that have arisen in the course of the efforts. The key element in such organizations is the commitment to science of those who manage them. Specifically, scientific excellence seems to be fostered in those places where data are maintained by scientists seriously interested in using the data themselves and sharing the data with others for the purpose of doing science.

The Small Astronomy Satellite-1: An Organizational Approach with a Single Principal Investigator

As an example of a scientific data-management experience with a single Principal Investigator (PI), we consider the Small Astronomy Satellite-1 (renamed Uhuru after launch). This satellite was designed and developed in the 1960's and operated in orbit from 1970 to 1973. As the first satellite specifically devoted to x-ray astronomy observations, it provided an opportunity and a challenge for the development of an effective scientific data-management approach.

The PI and a small group of colocated, coinvestigators were actively involved in the prelaunch data-system planning, mission operations planning, and software development. This active involvement resulted in the delivery to the PI group of quick-look data within 24 h after acquisition of the data on-orbit. This established a precedent and standard for many subsequent scientific missions. The mission operations planning also resulted in a highly successful system that allowed reliable control of the satellite spin-axis direction and satisfactory operation of the satellite as a whole. The effectiveness of the system enabled the investigators to respond in a few days to the discovery of several transient x-ray sources. The timely delivery and analysis of quick-look data also permitted the scientific team to respond with detailed follow-up observations on a time scale of weeks to a variety of discoveries such as x-ray pulsars and extended emission from clusters of galaxies.

This ability to extract important scientific information from the quick-look data was critical for the scientific success of the mission, particularly in light of an inadequate capability to deal with the full production data set. Serious problems arose in both the computer hardware and software areas. In spite of the active scientific involvement by the PI group before launch, inadequate financial resources were set aside for the development of software and for the procurement of computer hardware or processing time. In the time frame of the late 1960's when this activity took place, the requirements for rapid delivery and scientific analysis of the data were almost revolutionary, so that both the GSFC Project Office and the PI group were often breaking new ground in the data-management area.

The problems encountered in the software area included insufficient development of contingency programs before launch (particularly in the aspect area), inadequate documentation of software, and a questionable choice of programming language. The software was written in PL/1, which provided several new features particularly in the areas of manipulating data bit strings and accessing data files on disk. Since the computer available to the PI group was sufficient only to analyze the quick-look fraction of the data, additional computing capability was required to analyze the production data. The requirement to use the software written in PL/1 and the absence of substantial funding to purchase computing time led to the situation of processing the production data on the GSFC 360/65, 360/75, and 360/95 computers. Since the Uhuru project was an "outsider" and since it required the mounting of data tapes and the use of large disk files, processing was restricted to weekend nights. The combination of new software development postlaunch, restricted computer availability, and poor match of computer hardware to scientific requirements resulted in the delay of the completion of the all-sky source survey and catalog until 1976, more than 3 years after the satellite ceased operation.

To remedy this situation, in part, two preliminary catalogs were published by the PI group, as were several dozen scientific papers on individual x-ray sources. The net effect was that the dissemination of important scientific results was reasonably rapid, but the availability of the overall data base to the NSSDC and the scientific community as a whole was delayed several years beyond that intended. Several outside scientists were able to obtain relevant data sets from the PI group in order to carry out independent research programs during the time before the data were available at the NSSDC. The reduced data package provided to the NSSDC was also used by a significant number of scientists who requested data, so that the archiving of the data was reasonably successful, although delivery was very late.

Some of the difficulties encountered in this program relate to the time period in which it was carried out. For example, the option of dedicated minicomputers for production processing of data did not exist in the late 1960's. In the same context, the establishment of adequate funding for software development and data processing was quite difficult with the "state-of-the-art" nature of the data delivery and processing requirements. Other problems as well as the overall scientific success of the program can be viewed in the context of the PI nature of the scientific data management. Items to emphasize here are the rapid publication of scientific results (a success) and the late delivery of documented data sets to the NSSDC and to the scientific community as a whole (a limitation).

University of California at Los Angeles (UCLA) Space Sciences Group

The UCLA Space Sciences Group is an informal association of scientists, students, and support staff encompassing parts of the Departments of Earth and Space Sciences, Physics, Atmospheric Science, and Astronomy and the Institute of Geophysics and Planetary Physics. The group includes individuals working in theoretical studies, numerical simulation, empirical modeling, data analysis, instrument development, and ground and satellite measurements of magnetic fields. One central theme that ties the group together is its interest in the solar-wind interaction with planetary magnetospheres, in particular the Earth's magnetosphere, and the many manifestations of this interaction seen during intervals of enhanced geomagnetic activity.

Included in the research staff of the space-sciences group are university faculty members, permanent research staff, visiting scholars, postdoctoral associates, graduate students, and undergraduates. The group provides office and laboratory space, secretarial support, computer programming, and computer time for many individuals. In addition, the facilities are used to support outside researchers who request data and analysis of data held within the group's archive.

Interactions between members of the group are promoted by a variety of means, including teaching programs, joint seminars, journal clubs, informal and formal meetings, as well as the usual social activities. Resources are used to bring visitors for intervals of a fraction of a day up to a year. Collaborations between members of the group and individuals outside are encouraged in every way possible.

The major resource of this group is its archive of magnetic-field observations made in the solar wind, in the magnetospheres of the Earth and other planets, and on the Earth's surface. In cases where a UCLA scientist was the PI of an experiment, the archive contains the primary data base for the experiment. Otherwise, where data were needed for correlative purposes, the archive contains a secondary data base acquired from a data center or from a PI at another institution.

Another characteristic of the group is the availability of facilities for accessing and manipulating the data within the archive. At the lowest level, this includes a data laboratory with work space; storage for hardcopy records; storage, display, and copying for microfilm; and terminals to interact with the group's minicomputer and the university mainframe computer. In adjacent laboratories, additional facilities are available for plotting digital data on graphics terminals, drum plotters, and electrostatic printer/plotters. Drafting facilities and services are also provided.

A major factor contributing to the growth was the development of a time-series data-base management system. This system began with the definition of a standard data format sufficiently general to hold most of the data used by the group in its research activities. Development of reading and writing routines, and a variety of application programs, made this format of general utility to a wide range of users. The addition of a data-set directory, file catalog, subroutine library, and data archive completed the system. At present, this system is implemented within the batch processing environment of the University's IBM 360/91. Development of a new data-management system using the group minicomputer is currently in progress.

Another important factor in the evolution of the scientific data-management activity of the group has been the presence of a support staff of professional programmers funded by a variety of agencies and programs. The staff has developed a set of general programs, has provided consultation to scientific programmers, and, in many cases, has performed the desired processing and analysis under the direction of the scientific staff. As part of its responsibilities, this staff acquires, organizes, catalogs, and displays data from a variety of sources.

An essential element of this unit is an engineering staff actively working on the development of magnetic-field instrumentation. This staff advises the research group on the characteristics of instruments and the possible causes of

artifacts in data. They are a source of information on techniques of analysis of digital data. They are responsible for assembly, minor maintenance, and interfacing of the group minicomputer facility. They are currently assisting in the development of communication links between the group minicomputer and a variety of computers at other institutions.

A primary requirement of successful scientific data-management activity is a research staff activity using the data within the archive. In the UCLA group this research staff is responsible for directing the acquisition, reformatting, processing, display, archiving, and purging of data. Through continual interactions with the data in their research programs, the research staff identifies problems with instruments and data processing and devises methods of correcting these problems when possible.

The group provides support for its permanent staff and long-term visitors. It encourages outside investigators to collaborate in the use of archival data. It informally advertises the availability of data and research problems. Members of its staff are active in advisory capacities, formulating policy for future missions, suggesting important research areas, suggesting changes in technical procedures, and recommending redirection of resources.

HEAO-2/Einstein Observatory Scientific Consortium

As described in Section II of this chapter, the scientific operation of the Einstein Observatory is a collaborative effort involving a consortium of investigators from the Harvard-Smithsonian Center for Astrophysics, MIT, GSFC, and Columbia University. The consortium is responsible for planning and executing the observing program, for preprocessing scentific data, for carrying out its own scientific investigations, and for operating a Guest Observer scientific program. In these areas the consortium is the interface between the x-ray astronomy scientific community and NASA. To implement the Guest Observer program, NASA has employed the formal procedures of an AO and a proposal review committee (consisting of NASA representatives, consortium representatives, and representatives of the astronomical community as a whole). In the first 2 years of operation more than 400 Guest Observer proposals were approved and carried out. In addition, an official Users Committee was established to review the overall technical and scientific operation of the Observatory.

The scientific consortium (in particular the SAO group for the imaging data) has the responsibility for developing processing software, for processing data, for providing computational facilities for guest observers, for assisting guest observers in their scientific analyses, and for maintaining data bases (primary raw data tapes and secondary processed image files). An identified funding of $500,000 to $600,000 per year has been required to carry out the

activities in support of the Guest Observers program, and this may not be sufficient to provide all the services required. As a result, documentation in the form of a User's Guide for the software and data system structure was not completed to the level desired at the time of launch, although 6 months after launch a functional guide did exist. Also, the distribution of data to Guest Observers was not so rapid as originally expected. This was in part due to the incompleteness of the overall data analysis system and in part due to the desire to complete most, if not all, of a Guest Observers program before having the observer visit SAO to receive and analyze the data. Guest Observers typically spend 1 week at SAO analyzing their data beyond the standard reduction provided and at the same time becoming familiar with the contents of the processed output. Additional analysis is then usually carried out by the individual observers at their home institutions. A system for archiving the data within the consortium is reasonably well established, as are the formal requirements for delivery of processed outputs to the NSSDC for further archiving. It is clear that at least for the duration of the program and the post-program data-analysis efforts, the participation of the scientific community will be through the Consortium rather than the NSSDC.

The Space Telescope Science Institute (ST ScI)

The Space Telescope Science Institute (ST ScI), although several years from operation, provides another example for which we expect a data-management operation as successful as the UCLA Space Sciences Group and the Einstein Observatory.

The ST ScI will be responsible for the scientific program of the Space Telescope. It will solicit and select observing proposals, do the science planning, assist the observers in conducting on-line operations, provide calibration processing for all ST data, archive both raw and processed data, deliver data products, and provide analysis capability to observers and the user community. In addition, the ST ScI will coordinate supporting observations with other instruments (both ground-based and space-based) and provide some facilities for correlating ST and non-ST data.

The ST ScI will employ about 200 people, of whom approximately 40 will be active astronomers. Of these, roughly 15 will be associated with the data-management functions of the ST ScI. These astronomers will be obligated to perform service functions for about 50 percent of their time, with the other 50 percent available to carry out their own research programs. These scientists will provide scientific direction to those members of the technical support staff responsible for calibration processing and data archiving. In addition, they will develop and maintain the analysis software, in some cases writing the software themselves and in other cases providing scientific direction to

professional programmers. These scientists will have an obligation to support the user community in obtaining data from the archives, in understanding the limitations of the data, and in using the ST ScI analysis software.

The ST ScI will have user and advisory committees, although the detailed structure of these committees will not be known until the ST ScI is firmly established. Because the ST will be a prolific producer of high-quality astronomical data (roughly 10^{12} bits per year), it is almost inevitable that the ST ScI will play a leadership role in defining software and data format standards.

There are two attributes that the ST ScI will not have: (1) the ST ScI will not maintain the principal data archives for astronomy or even space astronomy—only ST and related data will be archived; and (2) the ST ScI may not undertake extensive modeling to represent the data in the data bases—this activity may be carried out by individual scientists or groups of scientists. Finally, it should be noted that the ST ScI, not yet in existence, has not provided advice to NASA concerning the scientific requirements for data collection, processing, archiving, distribution, or analysis. Once established, however, it is fully expected that the ST ScI will provide such advice.

The Lunar Consortium

The command module used during the Apollo 15 and 16 missions contained instruments to measure the characteristics of the lunar surface and interior. In particular, the modules were equipped with a gamma-ray spectrometer, x-ray fluorescence spectrometer, magnetometers (on-board subsatellites), and a laser altimeter. In addition, the lunar gravity field was mapped for regions beneath the subsatellites. When combined with photoelectric, multispectral telescopic imaging of the lunar frontside, with geological mapping, and with estimates of the surface age from crater degradation states, these data sets provided a unique, although formidable, opportunity to conduct multivariate analyses designed to elucidate our understanding of the lunar surface and crust.

To facilitate such analyses, a consortium of about 40 scientists was founded at La Jolla, California, in 1974. A research effort was begun at the USGS Flagstaff Image Processing facility to develop data-processing schemes that would accept diverse data, reformat the data, and display results. The Flagstaff facility became the focus for the consortium effort, developing array processing programs to reduce, display, and correlate the large number of data sets. Results are being produced on a continuing basis, with most of the computation being conducted at Flagstaff and interpretations being conducted at individual investigator's home institutions, as well as during visits to the facility. The Lunar Consortium also serves as a pilot program to develop a data-handling system for handling the multivariate data sets that are

anticipated from future planetary orbiter missions, such as the Galileo mission to Jupiter and the Galilean satellites in the mid-1980's.

The Lunar Consortium, with its Flagstaff base, illustrates a number of positive concepts. Specifically, the purpose of the program was to coordinate analyses of diverse data sets, to provide the science community with reduced and comparable data sets, and to develop a software capability generally applicable to comparisons of different kinds of data. Recently, a Mars Consortium was initiated, with Flagstaff serving as the coordinator center, to facilitate comparisons of the diverse Viking and Earth-based data for Mars.

Regional Planetary Image Facilities

The large number of images already acquired during lunar and planetary missions, together with the large number expected from future missions, such as Galileo and the Venus Orbiting Imaging Radar (VOIR) mission, make it prohibitively expensive for each PI to have an extensive data set. Some 500,000 separate images exist at present, along with extensive sets of shaded relief, topographic, and geologic maps (Table 3.1). The NSSDC is the prime depository for those products. It is also the center that researchers contact for acquisition of products. It is difficult, however, to know what to order, even with the extensive catalogs available from NSSDC. The Planetary Division, NASA Headquarters, has recently helped to alleviate this problem by setting up several Regional Planetary Image Facilities at the Jet Propulsion Laboratory, Pasadena, California; University of Arizona, Tucson, Arizona; Astrogeology Branch, USGS, Flagstaff, Arizona; Lunar and Planetary Institute, Houston, Texas; Washington University, St. Louis, Missouri; Cornell University, Ithaca, New York; and Brown University, Providence, Rhode Island.

Each facility has a full set of images, together with all associated maps and other products, such as catalogs. In addition, each facility has a quick-look capability, consisting of an interactive link to the institution's computer, together with the software necessary to conduct interactive interrogation of image engineering data such as latitude and longitude of picture center and slant range. Plans are currently under way to acquire videodisk players to use for quick-look displays of image data. Each videodisk can store up to 52,000 frames on one side. The players would be linked to the interactive terminals so that searches could be conducted on the engineering data base, and those pictures that fulfill search constraints could then be displayed interactively. Such a capability reduces the time that a researcher needs to spend looking at hard copies. The facilities, which were set up at sites where there is a long-term science interest in planetary sciences, thus serve as a regional data base for use by the scientific community. A researcher can visit the facility, utilize the quick-look capability to decide which of the data he or she is interested

TABLE 3.1 Summary of Image Data for the Moons and Planets at Washington University's Regional Planetary Image Facility

Object	Mission	Launch Dates	Remarks	Data Products
Earth's moon	Ranger 7, 8, 9	1964–1965	Probes impacted on the moon after acquiring 17,259 frames	Mission reports
Mars	Mariner 4	1964	Flyby mission, acquiring first 22 frames of Mars	8-in. x 10-in. prints
Earth's moon	Lunar Orbiter 1, 2, 3, 4, 5	1966–1967	Acquired 1474 frames of lunar surface, some with meter-scale resolution	20-in. x 24-in. prints
Earth's moon	Surveyor 1, 3, 5, 6, 7	1966–1968	Soft landers on moon, transmitting back 86,897 pictures, together with data on soil chemistry	Mission reports
Mars	Mariner 6, 7	1969	Mars flybys, returning 235 frames	8-in. x 10-in. prints
Earth's moon	Apollo 15, 16, 17, (pan photography)	1971–1972	4685 panoramic frames acquired from command module	5-in. x 46-in. prints; microfiche
Earth's moon	Apollo 15, 16, 17 (metric photography)	1971–1972	6781 metric mapping frames acquired from command module	5-in. x 5-in. prints; microfiche
Earth's moon	Apollo 8, 10–17 (Hasselblad photography)	1968–1972	Numerous frames taken by astronauts from orbit and from surface	8-in. x 10-in. prints; mission reports; microfiche
Mars	Mariner 9	1971	Photographed 95% of surface; acquired 7329 frames	11-in. x 14-in. prints; microfiche

TABLE 3.1 Continued

Object	Mission	Launch Dates	Remarks	Data Products
Mercury Venus	Mariner 10	1973	Transmitted over 8000 frames of Venus, Mercury, during flybys	8-in. x 10-in. prints; 70-mm negatives; positives; microfiche
Mars	Viking Orbiter	1975	Transmitted 50,000 frames of Mars; high-quality views of Phobos and Deimos	5-in. x 5-in. prints; 20-in. x 24-in. mosaics; microfiche; selected magnetic tapes
Mars	Viking Lander 1, 2	1975	Over 6000 frames of Martian surface acquired	5-in. x 5-in. prints; negatives, positives; 20-in. x 24-in. mosaics; microfiche; magnetic tapes
Mars	Viking Lander 1, 2	1975	Over 6000 frames of Martian surface acquired	5-in. x 5-in. prints; negatives, positives; 20-in. x 24-in. mosaics; microfiche; magnetic tapes
Galilean Satellites (Io, Europa, Ganymede, Callisto)	Voyager 1, 2	1977	Flybys of Jupiter in 1979	5-in. x 5-in. prints; negatives; microfiche

in, examine hard copies, and then order the subset of the data needed from the NSSDC.

The Regional Planetary Image Facility concept has many advantages. First, the Directors of each of the facilities must be PI's in the Planetary Program, thus assuring a continuing interest in the data sets. Second, each facility has a full-time photolibrarian, supported by institutional funds, to assume continual archiving and upkeep of data. There is a Committee, composed of Facility Directors, whose purpose is to oversee general operations of the facilities and to make recommendations to NASA with regard to such things as proper

archival conditions for housing the data or which data sets need to be recovered. For instance, the Planetary Division has no archival facility for housing digital image tapes from lunar and planetary images. Tapes are kept in a given mission's data library at JPL until the library overflows. Then, the tapes are put in storage at JPL or stored at a federal warehouse in Laguna Beach, California. The USGS at Flagstaff and some of the other facilities house a large fraction of the data set but not in archival fashion. One of the future recommendations from the Directors' Committee will be to begin transferring the data to high-density tapes, although it is not clear that the Planetary Division can afford to do so.

Most of the facilities also have computational capabilities for image data analysis. A researcher could arrange time to conduct data-processing tasks on one of the systems, although, at this time, such arrangements are informal in nature. Members of the Directors' Committee have also made recommendations to NASA with regard to image-processing software development, transportability, and distribution.

A Spectral Data Base for Earth Observational Research

The following is a description of a spectral data-base system that was set up over the past 5 years to serve the research community in Earth observational programs.

One of the most dominant characteristics of the field of Earth observations is the complexity of the natural scene. In order to devise successful satellite-based information systems for Earth resources data, a significant degree of understanding of the spectral reflective and emissive characteristics of various types of vegetation and other land cover is required. The processes involved in solar illumination falling on the surface of the Earth and then being re-emitted or reflected are at present not well understood. As a result, only rather simplistic models can be constructed to predict and simulate each mechanism. These models are not complete enough to enable the study of desirable sensor system characteristics nor research into information-processing matters.

Thus, a data base is needed both to permit the required advance in scene understanding and to serve as an empirical model of the scene for other types of system research. It is generally recognized that the problem will require a number of years of continuous research activity.

Several years ago NASA began funding the construction of a suitable data base in which the spectral characteristics for various types of scenes and observation characteristics are represented. This effort has involved a number of different institutions. The data base itself is resident at Purdue University's Laboratory for Applications of Remote Sensing (LARS). This laboratory has the responsibility for scientific leadership as a result of the combination of

Earth-surface sciences, instrumentation sciences, and data-processing and -distribution capabilities that are present there. The specific circumstances of each year's data collection are proposed by appropriate scientists and reviewed by a representative body.

Figure 3.2 shows in overview form the types of data collected for this data base during the Large Area Crop Inventory Experiment (LACIE) program. Central among these are carefully calibrated high-spectral-resolution data gathered with a helicopterborne field spectrometer. Airborne scanner and Landsat data provide an ancillary scene characterization showing the spatial (geographical) variability of such spectra. Large quantities of other ancillary data, many of them in literal (i.e., descriptive) form, are also collected and stored; these document the circumstances of observation (sun angle, view angle, for example) and details of the scene (local weather variables, surface cover species and variety, cultivation treatment, for example).

Sites involved were limited by the cost and the small number of sensor systems and instrumentation teams with the specific skills necessary. This limitation has been alleviated by devising a lower-cost, easier to operate instrument, which can in future years be placed in the hands of a larger number of scientists for observation at a larger number of secondary sites.

In concert with the funding for data-collection operations, NASA also funded a data-archiving and a data-distribution system. Some 150,000 spectra with associated ancillary data now reside in this data base in digital form, and the data are frequently used by scientists at several institutions. Catalogs of data are published as needed, as are the procedures used in collecting and preparing the data for storage. Scientists gain access to the data by applying to NASA, which in turn authorizes its availability from the data center.

Availability may be via shipment of a computer-compatible tape directly to the scientist. However, remote electronic access is also available. In 1970, a remote terminal system was established at Purdue University for use in multispectral image-processing research. Terminals via leased lines have been available at several sites in the East, Midwest, and South. The system serves not only as a research tool for scientists at various sites but also as a means for interchanging ideas and capabilities via software placed on the system by one scientist that can then be used at other sites. The system, from its beginning, has used a resource-sharing type of software system, as compared with a batch mode, such that it is quite possible to provide every user simultaneously immediate system access without interference with another user. Many of the terminals are simple alphanumeric CRT devices; some also include line printers and, in one case, tape units. In a few cases, a minicomputer system with its associated I/O and storage devices serves as the local terminal. Two sites have dial-up ports so that scientists at additional locations may dial into the port nearest them.

Given the existence of this system, it was a natural matter to use it also for

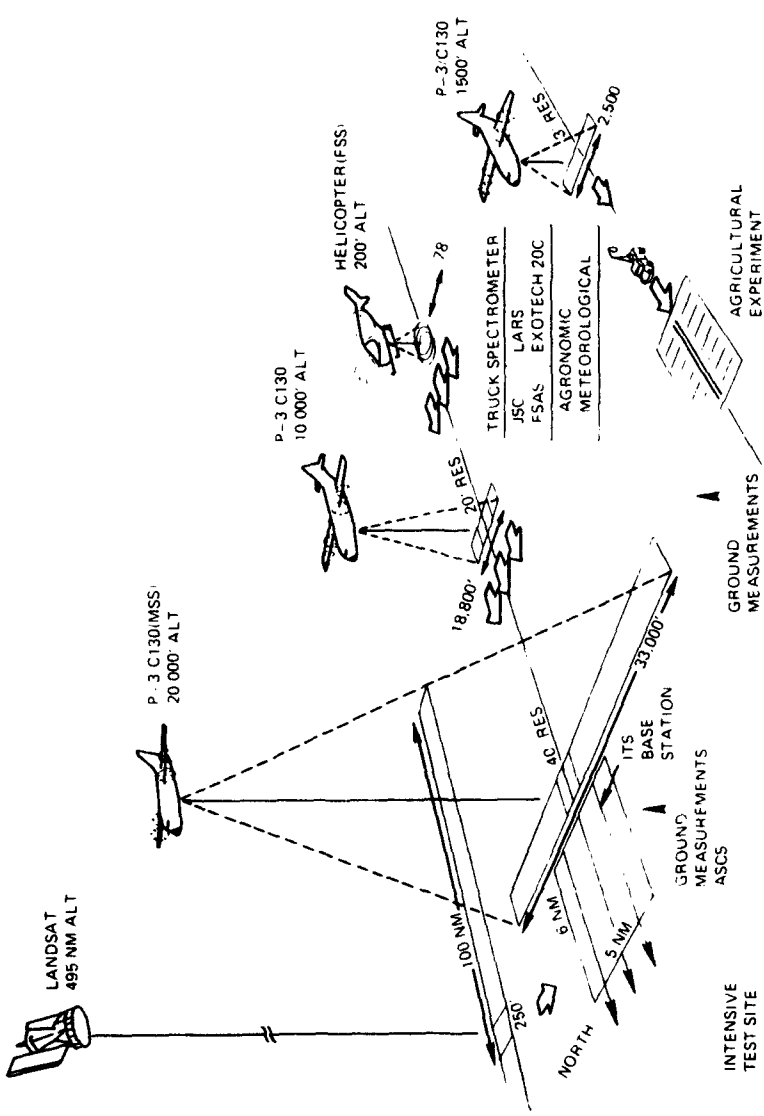

FIGURE 3.2 Schematic illustration of LACIE field measurements data acquisition.

the field spectral data base and to provide various capabilities to manipulate and analyze it. Over the course of the last few years a software system known as LARSPEC has been devised to carry out many data-management utility operations. Examples are spectra selection, retrieval, combination and averaging, and standard graphics operations. A detailed LARSPEC *User's Manual* has been published to aid users. In addition, various types of commercially available analysis packages, such as the Statistical Analysis System, have been added to the central site software. This system provides, for example, a wide range of statistical procedures including general linear models, multivariant analysis of variance algorithms, a variety of linear and nonlinear regression methods, estimation of spectral and cross-spectral densities, and maximum likelihood and nearest-neighbor discriminate analysis algorithms. It also provides a variety of easy-to-use line-printer plot and chart routines, including box graph, pie graph, and contour plots. A number of significant data-management tools are also available.

As time passes and more spectra are gathered, the value of this data base grows. It began with spectra exclusively from the U.S. wheat belt. More recently, spectra of a wide range of U.S. and foreign soils have been added, and currently programs are under way to gather spectra typical of earth cover found in the U.S. corn belt. As a result of this data base, significant advances in the understanding of natural-scene spectral characteristics is made possible.

Coordinated Data Analysis Workshops

Recently, a new approach to the analysis of space data has been developed as part of the International Magnetospheric Study (IMS). This approach is called the Coordinated Data Analysis Workshop (CDAW). A CDAW is a computer-based interactive, graphics-assisted workshop, organized to address a selected physical problem in space research utilizing data from satellites, rockets, balloons, aircraft, and ground-based measurements. A CDAW consists of five phases during which a problem is selected, data are entered into a data base, investigators study results at a first meeting, and a follow-up workshop clears up remaining questions. Finally, publications are generated from the CDAW results by various participants. At present, the CDAW concept is evolving rapidly as a result of experience of the organizers and participants and from the acquisition of new equipment and software.

DATA SYSTEM PLANNING

The CDAW concept is a result of joint planning by the IMS steering committee and members of the NSSDC. The primary factor motivating the development of CDAW was the recognition by the IMS Steering Committee that

since the data-acquisition phase of the study was coordinated for common interests, so too should be the data-analysis phase.

The primary purpose of a CDAW at present is to assemble a large collection of correlated data in a single data base and to provide means for displaying these data graphically. To accomplish this, the CDAW data system includes the following components: computer with disk memory, data-base management software, graphics software, graphics terminals, videodisk memory for graphics displays, intercom system to link various facilities, chalkboards, bulletin boards, slide projectors, audio system, copy machines, microfilm and microfiche readers, auditorium, and meeting rooms.

The first phase of a CDAW is planning a workshop. At present, this is done informally by an interested group that simply requests NSSDC support. If resources are available and no conflicting requests have been made, support is made available. The organizers then advertise the planned meeting and its theme. Simultaneously they request suggested time intervals for study and solicit participation. The organizers make the final selection of the time interval to be studied. The NSSDC circulates a request to submit relevant data to the data base.

The enthusiastic response of the space-science community to the CDAW concept has produced a demand exceeding the resources available to the NSSDC. As a consequence, more formal methods for selection of topics are evolving. Currently a Working Group on the Data Analysis Phase of the IMS is acting as an advisory group attempting to coordinate the choice of topics to realize the goals of the IMS.

PREPROCESSING

Extensive preprocessing of data is required before it is displayed in a CDAW. This preprocessing begins at the various institutions that have decided to participate. Given the selected time intervals, the investigators run programs to select, calibrate, and correct data from their instruments. They create a digital tape containing a time series of measurements and mail the tape to the NSSDC accompanied by a description of the tape format.

On receipt of the tape at the NSSDC, members of the programming staff write a program to read this tape and place it on the disk memory. They then write further software to interface this disk file with existing graphics display programs. Once completed, these programs are used to create listings and plots that are sent to the original investigator for validation.

DISTRIBUTION OF DATA

The primary mode of data distribution in a CDAW is through the workshop. A 3- to 5-day meeting is scheduled at NSSDC. Investigators come to the meeting, often bringing additional data in hardcopy or film format. The first

day of the meeting is usually devoted to presentations by the organizers attempting to formulate questions to be studied. Following the presentation, investigators begin to correlate different data sets. At present, this is done by plotting different parameters with the same horizontal time scale on one graph. Vertical scaling can be accomplished through the use of algorithms that make simple transformations of one or more parameters into a single parameter. Currently, algorithms do not allow measurements from different times to be combined.

To obtain a particular graph, a participant specifies the desired parameters, algorithms, and scales on special request forms. The requests are relayed to a programmer via intercom by an NSSDC coordinator, and the data are plotted on a graphics terminal and transmitted by close-circuit television to a video recorder at the requestor's location.

On receipt of their graphs the investigators discuss the results with their colleagues, including other experimenters, empirical modelers, numerical simulators, and theorists. At the end of each day, subgroup leaders prepare a brief report using transparencies and viewgraphs produced during the day. These reports attempt to summarize initial conclusions, problems, and directions for further work. At the end of the workshop, final reports are prepared.

At the conclusion of the workshop, the disk data base is copied onto tape. All graphs prepared during the meeting are microfilmed and mailed to the participants. The data base remains on disk, and requests for plots are satisfied on a time scale of several weeks. Recently, the capability of copying a limited segment of the data base onto tape has been added to the system so that an investigator may carry out further digital processing on his own computer. An even newer capability is remote access to the on-line data base using appropriate remote graphics terminals.

Subsequent to the workshop, investigators study the data obtained and collaborate in the preparation of papers. If necessary, additional data are obtained via one of the above mechanisms or through a follow-on workshop at NSSDC.

DATA STANDARDIZATION AND FIDELITY

At present, no attempt has been made to standardize the format of the data files submitted to NSSDC. Further, the internal data files used by the CDAW computer have been developed on an *ad hoc* basis by the NSSDC programming staff. Validity of the data submitted to the CDAW is the responsibility of the originating experimenters.

SOFTWARE DEVELOPMENT

Software to support the CDAW has been developed in the same manner as internal data formats. Until recently no one outside NSSDC has been required

to generate transportable software so that a CDAW could be supported by a computer outside NSSDC.

DISTRIBUTION OF COMPUTATIONAL CAPABILITY

The present CDAW concept does not allow digital data to leave the data base except as tapes in response to special requests. As a consequence, all computations must be performed within the CDAW computer. In addition, because this computer has limited power, these computations are limited to simple algorithms that transform several parameters into a new parameter. Currently, the algorithms cannot reference points earlier or later than the time point being generated. In addition, the system does not allow new parameters generated in this fashion to become a part of the data base for future reference.

MASS DATA STORAGE AND RETRIEVAL

The CDAW system utilizes disk storage. All data to be accessed by the system must reside on disk. There is currently no facility for moving data sets to and from tape during a workshop.

INTERACTIVE PROCESSING

The CDAW system is not truly interactive. A user must submit his plot request to a coordinator. The coordinator in turn relays the request to a programmer. The programmer then enters the plot parameters through cumbersome fixed-format, coded control statements. No processing other than the generation of time series plots is provided.

VI. CONCLUSIONS

Based on the reviews of selected space missions, data archives, and data-processing units given above we draw the following conclusions:

1. The active participation of scientists in all aspects of the data stream is a necessary condition for successful data management and computation. Scientists should participate in the earliest stages of mission planning when important aspects of data acquisition and preprocessing are being established. This participation should continue through the further processing and analysis of the data and finally in the archiving process. It is the scientist who has the motivation to assure that maximum scientific results will be obtained from the data. Many of the mission examples presented in this chapter illus-

trate how scientists actively interact with the data-processing activities in order to expedite and improve the process. In programs in which scientific involvement was less substantial, there were many problems that have limited the scientific utilization of the data.

2. Adequate funds for data management and computation must be provided, and these funds must be protected. Realistic budgets for required computer hardware, software, and operations should be established early in mission planning. Many of the examples of missions in this chapter illustrate how insufficient funds for data processing led to unacceptable delays in making the data available to scientists. In some cases, substantial quantities of data will never be processed. In other cases, only a fraction of the potential scientific results can be extracted from data because of inadequate processing. Seasat SAR is a classic example of inadequate funding for data processing. No funds were available to process the SAR data, and only a small fraction of the acquired data has been processed to date.

3. It is necessary to document software, and in many cases it is desirable to standardize software used for data processing. In a significant number of cases the documented software required to utilize data is not available to the interested scientist. In other cases an unacceptable level of effort and expense has been required to adapt nonstandard software. The use of incompatible image formats on the Viking mission has been a particular problem.

4. In many cases the use of dedicated minicomputers in data processing and computation is more cost effective than the use of centralized computers. Other advantages of the dedicated minicomputer are that priorities can be determined locally and peripherals and terminals can be specifically designed for a project. However, some image processing and scientific applications require larger computers and even special-purpose processors.

5. In a number of examples, the availability of quick-look, (sometimes) low-resolution data to serve as a guide to interested scientists has significantly increased scientific results. This type of data has also allowed active monitoring of satellite operation and scientific data acquisition.

6. Data must be archived in an appropriate form and state so that they are available to interested scientists. The data must be sufficiently documented. This problem is particularly evident with activities involving a single PI, where there is little incentive to carry out the necessary documentation of the data once the results have been published. No archival facilities are provided for the digital version of the planetary imagery data nor are the digital image data available from NSSDC.

7. Based on our reviews of the relative roles of the large national data archives and the smaller specialized data-handling units, we conclude that the smaller units also have a valuable role to play. Within these units scientists are actively handling the data, and guest investigators are able to utilize the data.

New calibrations are applied to the data, and the relevant related data are assembled. For sensor-derived data these small data-handling centers incorporating the active interaction of interested scientists can play an essential role. For sophisticated image data, small centers can also be expected to make valuable contributions to the scientific utilization of the data.

4
Technology Directions for Science Data Management

What we anticipate seldom occurs;
What we least expect generally happens.

BENJAMIN DISRAELI, *Henrietta Temple* (1837)

Science data management in the 1980's and beyond has the potential to be dramatically different from what has been experienced in the formative years of the space program. The differences will result through advances in technology at the microelectronic component levels, in storage technologies, in fiber optics, and in many other related areas. The effects of this technology will be evident from both cost and performance viewpoints and will influence every aspect of data management from acquisition in space to final processing, distribution, and presentation of data for interpretation.

CODMAC is of the opinion that there are no technological barriers to achieving a substantial improvement in science data management during the 1980's, even including the higher volumes of data that will be available. Improvements in technology will result largely from advances in the commercial sector; however, NASA funding will be required for certain items unique to space applications. We are also of the opinion that many of the current problems can be solved by employing current technology in new data system architectures and by employing more discipline and the new methodologies for the development of software.

The following sections of this chapter identify the technology drivers for scientific applications; provide a brief summary of pertinent technology developments in the area of hardware, software, and systems; describe the

data-management capabilities that scientists can achieve through the utilization of this technology; formulate criteria for technology implementation to improve the support of science projects; and list areas for technology and systems that should be emphasized by the science community and NASA in order to satisfy the future needs.

Technology is discussed in terms of the state of the art and future developments. Future developments are generally covered through 1985. Developments beyond 1985 are discussed where data to project such developments are available.

I. TECHNOLOGY DRIVERS

Data acquisition, handling, communication, processing, storage/archiving, and distribution requirements that drive technology for future science applications are discussed throughout this report. Beyond the problems that have existed in the past, new scientific requirements and developments in sensor technology are resulting in greatly increased data rates for future satellites. Representative examples of satellite programs that will require the implementation of new technology in order to manage the data from them are discussed in the Appendix.

Specific factors that are driving technology are as follows:

- Higher resolution instruments that generate an order of magnitude more bits/year.
- Requirements for more sophisticated on-board data management.
- High-volume/high-data-rate transmission.
- Processing rates that grow faster than the rate of increase in raw data.
- Growth by an order of magnitude of archival data volume and information that serve multiple users.
- Software development costs that are becoming an increasingly larger fraction of the total cost of a computer system.
- Trends toward dedicated computer systems (a result of technology) that place the computing function either at the source of the data or in the hands of the ultimate user of the data.
- Requirements for interacting with the data bases, the software, and the hardware of other science data users at remote facilities.
- Requirements for timely (sometimes real-time) delivery of data to the user community.
- Requirements for long-term reliability and fidelity of archival storage media.
- Trends toward more interactive processing with intercomparison of diverse (often large) data sets.

- Requirement for adaptive, remote decision making, especially for deep-space missions.

II. TECHNOLOGY PERSPECTIVE

Space data systems developed and used during the first 25 years of the space program have generally reflected the technology available to them. As needs have arisen, the technology to support these needs has either been available or developed at high cost. In situations where the funds were unavailable, the requirements have been adjusted to conform to the technology. Functions have been combined and new techniques have been implemented, but, in general, radical changes in the basic structure of data systems have not taken place. The technology to support new architectures was not previously available. We are beginning to see some significant changes for Space Transportation Systems (Shuttle) payloads that will fly during the 1980's.

Current and future NASA data systems have access to new technologies that can be used to enhance every aspect of space data management from collection in space through the final processing and display. The users and the designers of space data systems have considerable flexibility in where the data are processed, where and in what form they are stored, when and how the data are transmitted to the ground, and how they are handled on the ground. Capabilities that will be available from a technology viewpoint (although some development work may be required for space use and for specific NASA unique applications) for future space data systems include the following:

- Data-management systems with adaptive features and interactive capabilities (including voice interaction with computers), on board, on the ground, and via space-to-ground communications links.
- Space processing options that range from sophisticated, dedicated processing to shared resource processing, with on-line, fast access storage for data and complex software algorithms. Processing capabilities may be distributed at any point from the sensor to the final display.
- High-capacity data-base storage (10^{11} to 10^{12} bits) on board the spacecraft with update capabilities either on board the spacecraft or via the ground-to-space communications link.
- Programmable and adaptive space data-handling systems with asynchronous, variable-bandwidth, priority-controlled message-handling capabilities.
- Communications power and bandwidth to satisfy the most complex orbital requirements and ever-increasing deep-space requirements.
- Use of artificial intelligence and robotics to carry out adaptive remote decision-making actions, especially for deep-space missions. Artificial intelligence and robotics could provide the means whereby data-management

complexity could be reduced by allocating certain central control functions to instruments that can interpret the data themselves.

- Continuous orbital coverage with real-time access to data by users via small satellite-receiving terminals that are economically feasible for scientists.
- More powerful ground data-processing and data-base management capabilities through faster processors and memories, more sophisticated system architectures, and more sophisticated software. Processing power that ranges from a few million operations per second on small-computer terminals to billions of operations per second on superscale computers and/or special-purpose computers.
- Access to and control over spatially distributed processors and data bases using sophisticated distributed computing system networks, architectures, and protocols.
- The ability to store up to 10^{14} to 10^{15} bits of data on-line via centralized storage systems and to access these data via distributed networks.
- Capabilities for both broadcast and interactive distribution of wideband data via low-cost satellite receiving terminals; and increased terrestrial capabilities through the use of future hardware and software technologies, including advanced packet-switching techniques, fiber-optic links, advanced communications software, and sophisticated terminals using single and multiple-chip processors with the capabilities of today's larger minicomputers.
- Minimal need to share computer and memory resources on board future spacecraft. Each subsystem and experiment will be capable of having its own processors and storage devices except for instances where the same data and/or command/control capabilities are shared by many users.

The ability to implement the preceding capabilities within the budgets of NASA and participating scientists will require advances in a number of technologies, including semiconductors, magnetics, fiber optics, and software. The principal element affecting data systems, however, is the availability of low-cost computers, high-capacity memory devices on a chip, and very-low-cost mass storage. Subsequent sections discuss the effects of this technology in terms of the data systems elements that comprise the science data-management chain.

III. TECHNOLOGY SUMMARY

A technology overview for the principal elements that comprise the science data-management chain is presented in this section. Volume 2 of this report, *Data Management and Computation: Technology Trends* (National Academy Press, Washington, D.C., to be published), presents a more detailed discussion

on several of the elements discussed herein. Sensor technology, however, is not addressed in depth in Volume 2.

Sensors

Space data sensor technology will advance in the areas of increased resolution, improved sensitivity and spectral range, and other acquisition-related capabilities; however, the greatest potential for improvements will result from advances in processing and data storage. The integration of high-speed signal processors and/or superminicomputers into high-data-rate sensors will provide the capability to acquire, process, and store data selectively as they are collected. The result will be "smart" or adaptive sensors with the capability to manage data during the acquisition process.

Features include the following:

- The ability to search out specific types, classes, and/or levels of data and to send data back to ground only when those data have been located.
- The ability to modify the sensor's program sequence when unable to perform a preprogrammed function because of anomalies such as cloud cover or the absence of events of provided description.
- The ability to store data/scene from previous observations, to perform complex on-board data processing, and to send to earth only pertinent and/or new data.
- The ability to control resolution and data rate and to send back high-resolution data over selected areas.
- Preprocessing, filtering, and other complex operations under the preprogrammed and/or interactive control of the scientist.
- The ability to adapt or reconfigure automatically the data system in response to changing requirements and/or conditions.

Space Data Processing

Space data processing has traditionally been a highly centralized function with limited and closely controlled resources. The hardware has been highly complex and expensive in order to accomplish the required processing function within the weight, power, reliability, and space limitations imposed on space hardware. The requirements to conserve and share computing resources have created the need for complex software to manage the resources that exist. Also, the limited resources have necessitated the tuning of software in order to optimize its use and have not generally permitted the use of higher-level languages. The future availability of more powerful processors and storage capabilities on board should lessen these burdens.

Space data systems that will be used with Space Transportation Systems (STS) payloads are evolving toward a distributed processing approach, employing what is currently a centralized data-base concept. Although the system is not a distributed system in the strictest sense, and certain elements of the hardware are not state of the art, the system permits the use of hardened versions of state-of-the-art commercial processors that are dedicated to, and under the control of, the scientist. Use of these commercial processors provides a number of advantages, including relatively low cost, upward compatibility of hardware interfaces and software as technology permits more sophisticated systems, availability of a wide variety of low-cost interfaces, existing software (applications and system software), low-cost software development, use of higher-order languages, and familiarity with the computer by the scientist.

It is recognized that time and further investment of resources may be required before these commercial systems are available for certain applications. However, the potential rewards appear to be fully worth the investment in view of the fact that 16-bit systems with 32-bit internal data paths are available today, with processing capabilities on the order of 0.85 million operations per second (MOPS) and memory capacities to 2 million words. 32-bit systems are in preparation with instruction rate capabilities in excess of 1 MOPS and addressing capabilities to several million words.

Technology advances will encourage the proliferation of distributed space data processing to the point that every major subsystem and experiment will have its own processor, and larger subsystems and experiments will have multiple processors. Similarly, each subsystem and experiment will be capable of having its own data base, along with the capability to share elements of that data base with other subsystems and experiments. By enabling the scientist to have control of his or her own data bases and software, the performance of space processing systems should improve, and the overall cost will be less.

Space Data Storage

Data storage technology advances will provide significant advances in both storage capacity and reliability of spaceborne systems. In addition to the multimegabit, high-speed semiconductor memory devices that will be available for high-speed buffering and for supporting processing operations, tape-recorder technology will continue to improve and will be complemented by practical bubble memory systems for on-board playback and data-base storage. Bubble memory devices could, subject to NASA funding, dominate the storage capabilities of small spacecraft in the 1980's because of their inherent flexibility, access time (compared with that of magnetic tape), and projected reliability. However, larger spacecraft will need to utilize both types of

data storage, with bubble memories for on-line storage and tape systems for off-line storage.

State-of-the-art (1980) capabilities for aerospace recorders are on the order of 10^9 to 2×10^{10} bits. Prototype magnetic bubble memory systems have been built for space applications with capacities to 10^8 bits using 100-kbit memory chips. Using 1000-kbit chips that are available during 1980, a magnetic bubble memory system could be built today with a capacity of 10^9 bits. Using mid-to-late-1980's technology, scientists can expect to have access to on-board data-storage devices with capacities on the order of 10^{12} bits (e.g., ten thousand 10^8-bit chips).

Space Data Handling

Space data-handling systems have traditionally been fixed-format systems that multiplex data synchronously and transmit all acquired data to the ground. Although simple adaptive multiplexing schemes have been implemented by changing the sensor sample rates via the controller software during various mission phases, truly adaptive data systems have generally not been feasible previously because of the complexity and cost of hardware. Also, adaptive features such as data compression have not been used effectively, partially because of hardware and software limitations and partially because of the desire of the science community for raw data.

Today's technology offers an opportunity to implement adaptive data-handling systems for spaceborne instruments and experiments, and current trends are pointing toward such systems. An ongoing effort in the NASA NEEDS program provides for asynchronously multiplexed packets of data, buffered in variable-capacity data buffers. A number of activities are also ongoing to develop algorithms within NASA and elsewhere for compressing data. Given today's hardware, the implementation of these algorithms is becoming feasible, and they will have application in many future science programs.

The asynchronous, packetized approach to on-board data management offers a number of advantages on the ground and on board the spacecraft.

Advantages on board the spacecraft are as follows:

- Packets of data can be downlinked either on a downlink "when ready" basis or on a priority basis.
- High-priority experiments and/or instruments have access to the full downlink capacity during periods of high interest.
- Double buffering at the downlink source may be used to respond to requests from the ground for retransmission for purposes of error detection and correction.
- Data that are packetized are formatted in accordance with the wishes

of the experimenter. Thus, the data require minimal preprocessing, and the scientist does not have to reformat the data when received.

- Ancillary data are assigned packet numbers and are transmitted in packets that are available on request. Certain ancillary data (e.g., time) may be included within the individual packets as appropriate.

The principal advantage to packetized data is in the area of ground processing. The data that arrive on the ground are already addressed and formatted. Preprocessing is minimal, and the scientist should be able to receive his or her packet, along with any ancillary data, almost immediately.

Another major improvement in the projected capabilities of general-purpose data systems is the potential for much higher intraspacecraft data rates as a result of fiber-optics technology. Currently, serial time division multiplexed data buses operate at data rates of 1 to 2 Mbps. Fiber-optic data buses are under development for both ground and spacecraft applications. Data rates on the order of a few hundred Mbps appear to be feasible during the late 1980's.

While some scientists have traditionally rejected any attempts at data compression on board the spacecraft on the grounds that it is impossible to predict *a priori* the data characteristics, other scientists (particularly those involved in planetary programs and particles and fields programs) have run successful programs using compaction schemes. As a result of limitations imposed by the former group, the scientific community has sacrificed the potential for obtaining meaningful data and simultaneously processed vast amounts of redundant data. The advances in hardware technology and in algorithm development make it necessary for a re-evaluation of this practice of resisting all attempts to compress data. Current practices for image data provide a lossless compression ratio of 2.5:1. If some minimal loss in data is acceptable, compression ratios on the order of 10:1 to 20:1 are possible. Considering the potential for large data buffers and powerful processors on board future spacecraft, it would appear that the performance for certain types of missions (e.g., image data of planets, where bandwidth is at the greatest premium) could be enhanced even more. The amount of future enhancement could not be established as a part of this report, but it deserves future study by NASA and the science community.

Space-to-Ground Communication

Space-to-ground communication should be transparent to the scientist and as such should not be of direct interest to him. However, the space-to-ground link is not transparent, and, in fact, it frequently impacts on the scientist's approach to data collection. Major considerations related to the space-to-ground communications link are the following:

- Link cost, if directly chargeable to either the scientist or his or her program manager.
- Data timeliness.
- Available bandwidth to the scientist.
- Effects on experiment design (e.g., need to buffer/store data).
- Data quality.
- Potential for data loss, where one element of the link supports multiple users on a priority basis.

While all the preceding considerations have an effect on the scientist, few of the considerations are under his control. NASA has to set and enforce standards and policies for this link. Especially important is the fact that this element of the data-management chain is the one element that is not so readily affected by technology changes because the link for earth-orbiting missions involves communications satellites that have a design/implementation schedule of approximately 5 years and a lifetime of 7 to 10 years. In view of this lengthy lifetime wherein the majority of the link hardware is fixed, the principal choices for improving performance are through operational techniques and the improvements in interfacing hardware and/or software.

The paragraphs that immediately follow address consideration for technology improvements related to Earth-orbiting missions. Deep-space missions are addressed at the end of this section.

Specific limitations imposed on Earth-orbiting missions based on current (1980) technology and link considerations are the following:

- Data timeliness: data can be as much as 400 min old before they reach the users because of visibility limitations and store-and-forward delays at the direct ground site (DGS).
- Bandwidth is not being used efficiently.
- TDRSS single-access service is expensive ($83/min). This pricing policy forces users to store their data on board and read them out in short bursts in lieu of having real-time channels.

Specific improvements to the link capacity of Earth-orbiting missions that are possible through the implementation of 1980's technology are the following:

- Improved link capacity through the use of more efficient modulation (2 bits/Hz).
- Improved timeliness and data quality through the use of "intermediate frequency bent pipe" mode at each DGS in the link.
- Expanded Tracking and Data Relay Satellite System (TDRSS) single-access K-band (KSA) service on the next generation of communication satellites that accommodate more users and reduce link costs.

- With government funding, the development of 20-W, solid-state, reliable, space-qualified transmitters for use on TDRSS KSA.
- Spot-cast transmission of high rate data on an as-needed basis.

Deep-space missions have for many years been optimizing their data handling and communication links. Because deep-space missions are one-of-a-type missions, they have more control over their communication channels than do Earth-orbiting missions. Deep-space users will be able to expand their downlink capabilities through improvements in transmitter and receiving devices, antennas, and encoding techniques. Equally important, the deep-space user can improve communication channel efficiency through the utilization of more sophisticated processing techniques that enable intelligent/adaptive data collection, handling, and transmission.

Computers (Ground Based)

The steady increase in the volume, acquisition rate, and variety of space-science data collection, and in subsequent processing requirements, is increasing the need for faster and more capable computer systems in science data-processing facilities. Fortunately, commercial markets and military and space requirements are at present driving advances in computer technology at a rapid rate. Thus, in principle, it is possible now (in 1981) either to acquire or to develop computer systems to handle most of the projected space-science data processing and short-term storage needs. The situation is not without problems, however; the price of commercial high-speed computers is still high with respect to the budget of a scientist, and superscale computers are not likely to be developed without extensive government funding, thus making them generally unavailable to scientists.

The most rapid advances are taking place in the microcomputer and semiconductor memory devices area. Already on the market are 16-bit microcomputers with sizable random-access memory capacity and sophisticated architectural features (e.g., 32-bit internal architectures) as discussed in more detail in Volume 2. Also on the market are 64-kbit random-access memory chips. 256-kbit and 1-Mbit random-access memory chips are in prototype stages in laboratories abroad and in the United States. 32-bit microcomputers on a single chip are expected in a few years (i.e., by 1983–1985). A consequence of these technology advances for space-science data management is that:

- Microcomputers and memory chips can be integrated with sensors, instrumentation, and control units in spacecraft to increase the versatility and adaptability of these units and to perform data preprocessing on board.

- Large amounts of low-cost computing power and memory can be incorporated into intelligent terminals to enhance interactive computing, display generation and presentation (including sophisticated graphic displays), and provide word-processing support.
- Microcomputers can be assembled into large arrays to increase significantly processing power for scientific modeling, space data collection and processing, and data-base accessing.
- Special-purpose computers can be constructed from microprocessor, memory, and special-function computation chips at relatively low cost for use as data machines, network communications processors, and adjuncts of general-purpose computers (e.g., floating-point processors).

In general, the cost of large-scale integrated (LSI) or very-large-scale-integrated (VLSI) circuit chips is only a few tens of dollars when produced in large quantities. However, at present, custom-made chips are expensive (thousands of dollars). Advances in computer-aided design (CAD) and computer-aided manufacturing (CAM) will reduce this cost considerably, but it will still remain relatively high. Thus, standard microcomputer and memory chips should be used to the fullest extent possible.

Minicomputers and mainframe computer technologies are similarly advancing, so that the cost of a constant level of computing power is decreasing, or, alternatively, more computing power can be acquired for the same level of cost. These systems are evolving toward effective use in networks of computers and distributed processing systems, as well as in the areas of access, retrieval, and maintenance of very large data bases. Finally, high-speed and superscale computers, which provide performance at rates up to a hundred million operations per second (MOPS) today, can be expected to evolve into computers with performance in excess of 1000 MOPS as early as 1985. These machines will dramatically reduce the present problems of processing large volumes of space data and will support computations for complex models of space phenomena. Especially important will be the development of special-purpose architectures for image processing and for management of very large data bases. For the latter function there are data-base computer architectures now in design and development and in prototype implementation.

By 1985, it is expected that computers will exhibit the following performance and cost characteristics:

- Minicomputers and superminicomputers will be capable of processing rates up to 20 MOPS. Prices will be in the $2000 to $40,000 range for "typical" systems with up to 2-3 MOPS performance capability.
- Mainframe computers will have similar performance capabilities, but they will be capable of handling large numbers of peripheral memories and

input/output devices, and they will provide much more extensive software support. These aspects will tend to keep their prices in the $60,000 to $800,000 range, as a function of the capabilities provided.

- High-speed computers will be capable of processing rates up to 100 MOPS by 1985, with prices in the $700,000 to $3 million range.
- Superscale computers can be expected to pass the 1000-MOPS performance capability by 1985 through the use of special vector-processing and array architectures that permit concurrent processing of multiple sets of data. However, high processing rates can be achieved only on processing tasks that are suitable for either vector or array processing. Prices of superscale computers in 1985 can be expected to be in the $5 million to $16 million range.

Dramatic advances in the processing speed and versatility of computer hardware underscore the problems that still complicate their use: man-computer interface and software production. Despite the development of advanced higher-order programming languages, utility programs in operating system software, and application program libraries, space scientists are spending a large part of their efforts in programming their computer systems and learning about the details of hardware architectures in order to optimize their performance and increase the efficiency of hardware use. Advances in hardware technology will be used to reduce the software generation problem.

In general, the software portion of systems cost is approaching the 90 percent mark, yet the selection of computer hardware and attempts to minimize its cost are given much more attention in system acquisition than is software. Project managers tend to view software as something that "will get done when needed." Uniformly, the result has been costly overruns, time delays in project completion, and even scrapping of effort and beginning anew.

While hardware advances can be used to alleviate some of the software generation and man-computer interface problems, there is also a need for management awareness of the software problem and greater emphasis on software from the inception of a project. Ideally, hardware should not be acquired before software requirements have been firmly established and the software is sized. This would permit acquiring more advanced hardware than is now the practice, as well as sizing the hardware to meet the software needs, not vice versa.

Ground Data Storage

Ground data storage requirements for space data fall into two categories: short-term data storage for processing and long-term/archival needs. Short-term needs will be satisfied during the 1980's by a combination of magnetic and optical storage devices, with magnetic devices dominating the field because of their read/write capability and flexibility. Mass data storage of 10^{14}

to 10^{15} bits of data is still evolving, and a clear leader for mass storage media has not surfaced. Current contenders include lasers writing on metallic media and lasers writing on film systems. Other approaches may evolve during the 1980's. The demand for these mass storage systems is limited, and it is not likely that new systems and/or techniques will evolve without government funding.

Short-term ground-based data storage requirements will primarily be met by magnetic disk devices, both rigid and flexible. In addition, there will be a growing use of cartridges for load/off-load and archiving functions.

Product advances will result from continuing increases in areal density, with improvements in access time resulting from weight reductions in actuator mechanisms. It is expected that the areal density of rigid disks will increase by a factor of 4 by the mid-1980's and by almost a factor of 10 during the decade.

In flexible media disk (e.g., floppy disks) and tape even greater gains may be realized since their initial uses did not focus on cost/bit. This type of storage is especially attractive for applications requiring many copies of the same data but not where only a few copies are required or for applications requiring data update. The advances in magnetic recording technology will lead to high-capacity low-cost units in small packages.

It is expected that during the 1980's videodisk, read-only storage will find a role and growing use in data-storage applications.

Data-Base Management

Data-base management is one of the most worked areas in data processing today. The problem is being approached from the standpoints of software, implementations of special processors (back-end processors) that are dedicated to data-base management, and the implementation of special computers with internal architectures that are designed to accommodate the management of data bases. Of more general interest is the trend that appears to be occurring in the area of mainframes. Many experts in the field are of the opinion that the next generation of mainframes will be oriented more toward data-base management than to numerical computing, as has been the tradition among computer manufacturers. Although these mainframes do not provide the capability of the data-base machine, they provide a mix of data-base management and basic computational capabilities. These new machines will provide a great advancement for the majority of users, but they may not be the answer to the scientists' needs for two reasons: (1) the software support available with these machines will probably be oriented more to commercial applications than to spatially oriented, time-series data, and (2) the reduced emphasis on numerical computation may have an adverse effect on the cost/performance ratio for scientific applications.

The majority of the data-base management system (DBMS) developers contend that their systems are general purpose and that their DBMS will satisfy the requirements of NASA and scientists. The NEEDS program has examined the data-base management task within NASA and is developing a DBMS concept that is unique to NASA. The functional requirements for the proposed system appear to be applicable to the need of the scientific community. At the time of this writing, NASA was implementing the system using the capabilities of two commercial systems. The combined systems will serve a much broader class of problems than standard DBMS's by the fact that it handles other NASA data problems (e.g., near-real-time data distribution). This program is important to the scientific community, and members of the community should be involved in its definition and development.

Another consideration that should be taken into account is that several NASA groups are working the data-base problem independently. NEEDS is developing one such system. Several other organizations within NASA and the science community are apparently doing the same thing. All of these efforts are important. However, a coordinated effort within NASA would not only leave more funds for science projects, it would produce results that are more beneficial to NASA and science users overall.

Communications Networks and Distributed Processing

The technology of data communications networks and distributed processing evolved at a rapid pace during the 1970's. This evolution was driven by technological advances in the areas of transmission networks, microelectronics, software, communications protocols, and packet switching. Early communications networks were characterized by high initial cost, single application orientation, limited reliability, limited flexibility in adapting to new applications, and centralized control. Today's networks, however, are characterized by relatively low initial cost, multiple application orientation, a high degree of flexibility to accommodate new applications, and distributed control.

In the area of transmission systems, technology advances have resulted in improvements in bit error rates of 2 orders of magnitude as a result of advanced modulation and coding schemes, a reduction in cost for data transmission, and increases in transmission capacity by a factor of 4. Also, during the 1970's satellite technology improved to the point that satellite links are now cost competitive relative to terrestrial links.

The microelectronics revolution of the 1970's spurred the development of powerful, low-cost communications processors. Both voice and data-switching networks have evolved toward the use of such processors as the key element in their systems. In addition, microelectronics advances have served to reduce

the cost of terminating equipment such as modems, line drivers, and signal conditioning equipment.

Software technology also matured significantly during the last decade. This is important and will continue to be important, because more data communications tasks are being implemented in software. The concepts of top-down design and modularity have played a major role in the development of reliable software for large communication switching systems.

Protocols for data communications received considerable attention during the 1970's. International standardization efforts have resulted in the development of a protocol for interfacing to packet-switched data-communications networks. Standards have also been developed for interfacing asynchronous and synchronous terminal devices to packet networks. This work is still ongoing and will evolve along with hardware and software advances to the point that distributed computing will truly become a reality during the 1980's.

The 1970's were a proving period for packet switching as a viable alternative for data communications in the commercial sector. Building on the success of the ARPANET in the late 1960's and early 1970's, companies such as Telenet and Tymnet were created to provide value-added data-communications services to the public using packet switching. Packet switching has become the most cost-effective choice for data communications that is bursty (i.e., the ratio of peak transmission rate to average transmission rate is high).

In the 1980-1990 time frame we can expect to see the following further advances in the technology of data communications and distributed processing:

- Increased use of fiber optics for terrestrial links. This will greatly increase the capacity and reduce the cost of transmitting large volumes of data.
- Greatly increased use of packet-switching techniques for local and long-distance networks.
- Higher-level services such as "electronic mail," which will be layered on top of existing packet networks.
- Extensive use of distributed processing systems utilizing intelligent terminal work stations interconnected with a local network. This can communicate with remote data-base systems using public packet networks as the communications facility.
- The integration of digital voice and data into a single network that handles both types of traffic.
- Increased use of low-cost (approximately $3000-10,000), high-capacity (several megabits) receiving terminals by scientists.

Distributed processing capabilities will continue to improve during the 1980's and will reach a high level of maturity for relatively low-data-rate

applications (9600 baud and less) as a result of advances in hardware, software, and networks. However, the capability to distribute large volumes of data in other than a broadcast mode will still be prohibitively expensive for the scientist during the 1980's. Multimegabit interactive applications will likely become feasible during the 1990's when fiber optics begin to replace existing wire links and some satellite links on a widespread basis into homes and offices.

Interactive Processing

Interactive processing is not a technology within itself; however, the trend in processing is strongly in the direction of interactive processing. From the time software and data are entered into a system, throughout the lifetime of a given data base, an interactive approach permits the scientist to fine tune his or her models as the data are being processed. Through the use of either alphanumeric terminals or graphics terminals, the scientist can observe the effects of various inputs in real time, and subsequent processing steps can be tailored to the results of previous computations.

The current trend in interactive processing is to use video display terminals with built-in processors with relatively large main memories. These terminals have access to data via either disks, tapes, high-speed electronic memories, or communication links. As the single-chip computers previously discussed become more powerful, and with ever-increasing main memories on the same chip, interactive processing will continue to expand. Future interactive terminals will have access to ever-increasing capabilities for processing, input/output operations, and communications. These terminals will play increasingly larger roles in distributed computing networks.

Display device technology plays an important role in interactive processing. Although there are a number of display technologies available for this purpose, the cathode-ray tube (CRT) will continue to dominate interactive display devices through the 1980's for low-cost users. Display devices have taken advantage of the microelectronic revolution, and today all major display devices have built-in memories that provide internal refresh, which offload the main processor and eliminate the need for storage tubes.

Today's CRT displays provide color, high resolution, and various manipulation capabilities such as zoom, rotate, and translate. Certain displays have capability for three-dimensional presentations. Although these displays are beneficial, they all have one limitation that eventually affects their performance—they are restricted to approximately 25 inches diagonal measurement, and this dimensional limitation will not change significantly in the future. As a result of this limitation, data are frequently crowded, annotations run together, and the human eye frequently cannot discriminate between data ele-

ments. Currently available projection systems tend to provide some improvement in terms of crowding and feature size; however, current low-cost systems are limited in resolution. As the commerical market for these systems expands, one would expect that the technology of these units could be enhanced and combined with integrated circuit technology to provide a high-resolution, internally refreshed graphics system that eliminates many of today's problems.

Software

Data-management programs are almost universally beset with software problems. The problems are manifested in a number of ways, the most prominent of which are:

1. Software is expensive.
2. It is not generally transportable.
3. Software documentation is generally imprecise and inadequate to satisfy requirements for operation, maintenance, sustaining engineering, and transportability. Inadequate documentation leads to the need to duplicate software, which increases costs.
4. Software development schedules frequently extend beyond the projected completion date, and needed software is not available with the remainder of the system.

The preceding problems are the result of both management policies and technological factors and/or issues. Management policies within NASA that have affected scientific software are as follows:

1. Scientific analysis software has never been considered as a deliverable item, and as such it remains the property of the developer. Each new project develops software that has been developed many times previously.
2. Software language and programming methods for scientific processing have not been specified by NASA, and no attempt to support PI software developments has been made.
3. An organized effort to standardize common procedures for the tasks and/or data was never initiated by NASA.
4. Since software was not a deliverable item, development schedules were not encouraged and/or enforced, and portability was never a consideration during the software design process.

The technological issues that contribute to the software problems are as follows:

1. There is a lack of software estimation and forecasting models for typical NASA software development projects. This causes unrealistic and optimistic forecasts and oftentimes results in late software delivery.

2. Software practices have not been standardized within the software community. There is a common belief that software development is a craft. Therefore, no disciplined practice is emphasized.

3. Documentation standards are not well defined and followed by strict enforcement.

4. The software industry did not extend the systems-library concept. Thus, many labor hours are wasted to duplicate the same software that has been developed previously.

5. There is a lack of precise and unambiguous requirements/specification languages to facilitate communication among different personnel working on the same software project.

6. Software design has not been developed into a readily applicable engineering discipline.

Data-processing professionals are beginning to realize the ever-increasing cost of software development, operation, and maintenance. When compared with the decreasing cost of hardware resulting from the hardware technology advances, the effects are even more alarming. Boehm collected historical data and plotted his famous cost-trends curve that predicted that software costs would be three times hardware costs in 1978 and will approach 90 percent by 1985. It was this fear of rising costs of software that led a group of software scientists to advocate using engineering discipline in the software development process; thus, the term "software engineering" was coined in 1968. The group felt that by establishing an engineering-like discipline for software development, operation, and maintenance, it would be possible to reduce overall life-cycle costs of software. This hope inspired many software professionals who have since dedicated their energy to this new field. The results have been many new approaches to developing software, many new tools to aid software development, and many new management philosophies to organizing software teams.

The majority of the new discoveries in software engineering have not been widely accepted by the general practitioners in the software industry. Therefore, the escalating cost of developing software has not been deferred. In part, this is because software professionals do not try new approaches, methods, and techniques. The more important blame, however, should be laid on the people who do research in software engineering. They did not provide software metrics or any other alternatives to gauge or demonstrate the usefulness and effectiveness of the new discoveries in software engineering.

The need for establishing measurements in software (software metrics) is obvious; without a system for measurements there is no way to demonstrate

the merits of new software methodologies, new tools, new principles, and new approaches. An alternative to software metrics is the establishment of a foundation for software engineering experiments. Once the experimental framework is set, many comparative evaluations can be performed and meaningful data can be analyzed to draw conclusions.

We are of the opinion that NASA should instigate a special effort to support the research for developing software metrics and experimental software engineering. These two areas actually complement each other by the fact that newly proposed software metrics can be used in experimental software engineering to be evaluated, and experimental software engineering can use the established metrics to evaluate other new developments. If the two research areas are properly funded, we are of the opinion that there can be an order of magnitude in improvements in the next 5 to 10 years in software development. The time is ripe for the evaluation of software engineering findings that have occurred over the past 10 years.

One recent study by the General Accounting Office found that private software companies that either sell software or produce it on a contract basis are the only organizations that consistently exploit the new software technology currently. Most government and private facilities have made only moderate use of new software tools and techniques. The report concluded that in general the private-sector facilities are more interested in, and devote more organizational efforts to the use of, software techniques than do government agencies. The Office of Management and Budget should motivate government agencies by requiring agency heads to establish software quality-assurance functions and to define more clearly management responsibilities for the acquisition, management, and use of software tools and techniques.

This section can be summarized by pointing out that there is a software technology problem to support future space-science data management. We cannot expect the symptoms illustrated in the beginning of the section to disappear without laboring on finding the real solutions.

The basic recommendations for NASA are the following:

1. Establish disciplined software practices for both individual programmers and group projects.
2. Embark on research to establish software metrics to facilitate software estimation and forecasting.
3. Initiate efforts to set up experimental software engineering frameworks and foundations to validate software methodologies through experimentation.
4. Establish documentation standards and guidelines and strictly enforce their adherence.
5. Establish a unified software library center to make available software products that have already been developed within the software industry.
6. Specify that all software should be transportable.

IV. CRITERIA FOR TECHNOLOGICAL IMPLEMENTATION

One of the principal complaints against NASA is that the Agency does not utilize current technology. This section discusses some of the areas where NASA has fallen behind in technology and provides recommendations and criteria for keeping more current.

One principal problem area is in on-board data systems. A mission and/or program is defined, the data system is designed, and the mission/program does not fly for 7 or more years. The technology that applied to the original data system was probably 2 or more years old when the data system was defined. It is on the order of 9 years old when the system actually flies, and it becomes much older if the system has a lifetime of a few years. There are a number of things that NASA can do to minimize this problem. The approaches involve maximum use of modular designs. First, NASA can specify the data system to take advantage of hardware that is on the cutting edge of technology. To the extent possible, the design should be modular to permit replacement by a later upgrade to that technology with a compatible interface characteristic. This approach requires NASA to develop flexible interface standards for major components such that growth and/or replacement of a given component with a newer technology will not affect existing interfaces. There are a number of interface standards in industry that manufacturers of computers and peripheral devices are conforming with. It is our recommendation that NASA use the conventional standards to the extent feasible but not be controlled by them.

Another area where NASA is noticeably behind in current technology is in ground processors. NASA centers and scientists are frequently processing current data with computers that are 10 years old. This approach to data processing is not prudent. Everything that is saved in hardware cost is probably lost two to five times in software cost and in the intangible costs associated with frustration. How and when does an agency such as NASA decide to replace existing equipment for new technology? The answer is not an easy one, but a criterion for arriving at an answer must be developed and implemented. Further, the implementation process is also not easy to implement since there are no defined methods for measuring some of the considerations identified herein.

Suggested criteria for replacing computers and associated equipment are as follows:

1. When its performance is overtaxed by either new uses or additional users such that turnaround time is excessively long and interferes with the work of the users. A cost analysis is likely to show that, collectively, the time and cost of lost productivity resulting from inadequate computing support will at some point be greater than the cost of obtaining a new, more capable computer system.

2. When the equipment is so obsolete that its reliability is deteriorating, and maintenance takes increasingly longer, or when maintenance support is no longer available at reasonable costs. Also, when spare parts are no longer manufactured and when down time becomes a source of the system being overtaxed as in Item 1 above.

3. When more capability can be obtained at the same or lower cost by acquiring more modern computer equipment.

However, there are factors that also work against changing the current computer:

1. Software conversion can be exceedingly expensive, and errors are likely to be introduced. This can occur when an older computer model is being replaced, when nonstandard features are in the present computer (e.g., word length), when especially constructed adjuncts are part of the present systems that have software implementation implications, and when programmers of the present system have left and documentation is not adequate.

2. Candidate replacement systems may themselves be undergoing change because technology may be just advancing to a new plateau. Sometimes, the candidate replacement equipment is so new that it may not be completely debugged.

3. Uses of the computer system may be undergoing a change that has not been sufficiently defined, such that there is risk that a new system also will not be adequate.

Another important consideration is whether to purchase or to lease a computer. The government generally purchases its computers and keeps them for the next decade. In the age of rapidly advancing computer technology, purchase so as to save expenses over some long time period may not be really cost effective.

Selection of the new architecture is another important consideration when planning to replace a computer. Much has been written about this choice between distributed and centralized architectures. Distributed processing systems tend to be modular and permit easy growth by adding new modules (provided that the interconnect architecture is properly chosen, too). Sometimes there are organizational problems, however, when it is proposed that computing power be distributed throughout the organization rather than kept under the control of some office in the organization's headquarters. This aspect of computing is going to assume increasingly greater importance during the 1980's.

5
Relevant Technology Programs

Science is built up with facts, as a house is with stones. But, a collection of facts is no more a science than a heap of stones is a house.

JULES HENRI POINCARÉ, *La Science et l'Hypothese* (1908)

I. OVERVIEW

There are a number of systems or programs that have been initiated within NASA that will influence the utilization of space-acquired data for scientific investigations during the 1980's and beyond. The Committee on Data Management and Computation received presentations on several of these programs during the course of its deliberations. The sections that follow briefly describe these and discuss their relevance for scientific utilization of space-acquired data. The systems are the following:

- Landsat Assessment System (LAS)
- Satellite Communications
- Applications Data Service (ADS)
- NASA End-to-End Data System (NEEDS)
- Space Science Data Service (SSDS)
- Jet Propulsion Laboratory Planetary End-to-End Information System

CODMAC recognizes that these programs are to a large extent still in a planning or definition phase. Thus the concepts presented to CODMAC in the 1978-1980 period by various NASA representatives are in many cases prelim-

inary and subject to change. Nevertheless, the essential directions intended for these programs seem to be well understood, and for this reason it is appropriate to assess their potential impact on data management and the extraction of information from space data.

II. LANDSAT ASSESSMENT SYSTEM (LAS)

Introduction

NASA initiated the Landsat-D research and development program in September 1977. The program includes launching Landsat-D in 1982 with a thematic mapper (TM) and a multispectral scanner (MSS). The spacecraft is designed to be compatible with the Space Transportation System's satellite retrieval and replacement capabilities. The Landsat-D spacecraft and sensor will be operated through a Tracking and Data Relay Satellite System (TDRSS) to produce an average of 50 TM and 200 MSS scenes per day over selected regions of the Earth.

A minicomputer-based Landsat Assessment System (LAS) is being developed at the Goddard Space Flight Center as part of the Landsat-D ground segment of the data-processing chain. The intent of the LAS is to assess the characteristics and utility of the Earth observations provided by the Landsat-D system, particularly the thematic mapper. To accomplish this, the LAS will be organized to support the assessment activities of investigators in at least five major disciplines: agriculture, hydrology, land use, forestry/rangeland, and geology.

Landsat-D System Overview

The overall Landsat-D system is shown in Figure 5.1. The Operations Control Center (OCC), the Data-Management System (DMS), the Domestic Communications Satellite (Domsat) station, and the Transportable Ground Station as shown in the figure are all located at the Goddard Space Flight Center.

The Landsat-D data will be received at Goddard directly through the Transportation Ground Station or through a TDRS/Domsat link. The latter data path uses a TDRS ground station at White Sands, New Mexico, for relaying the data through Domsat to Goddard and serves as the main data-acquisition link for the Landsat-D ground data-processing system.

The EROS Data Center will receive processed Landsat-D data in digital and film product form from Goddard, will further process the data, and will maintain data archives. The Center will also assume responsibility for disseminating Landsat-D data products to the public. Other domestic and foreign users

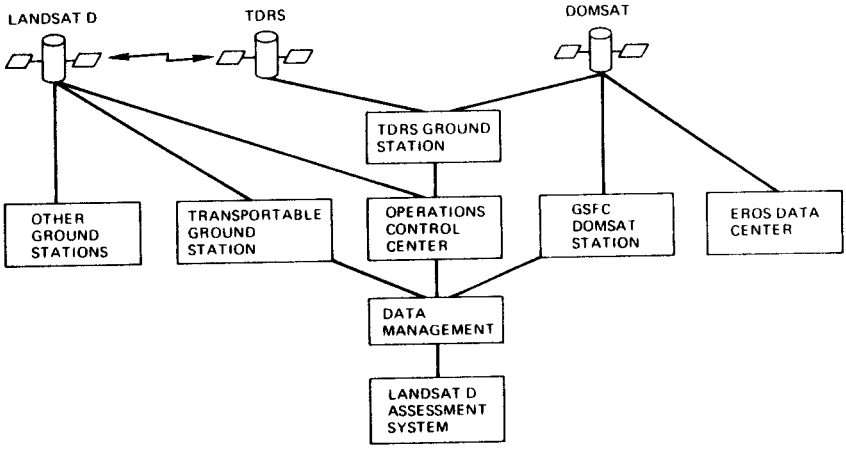

FIGURE 5.1 The Landsat-D system.

may enter into agreements with NASA to utilize their own ground stations to acquire the data directly from the Landsat-D satellite.

Landsat-D Ground Segment

The major elements of the ground segment include the Transportable Ground Station, the OCC, the DMS, and the Landsat-D Assessment System.

The Transportable Ground Station will be used to acquire the Landsat-D data directly for engineering evaluation of the spacecraft and sensors and will serve as a backup data-acquisition system in case of failure in the satellite data-relay communications link.

The OCC will coordinate scheduling of the ground-system resources and will provide the satellite command and control functions required to support the mission.

The DMS of which LAS is a part, is the heart of the Landsat-D ground data-processing system. As shown in Figure 5.2, the system will perform the digital image-processing functions required to produce corrected TM and MSS data products. A key performance requirement for the system is to produce final output products within 48 h of data acquisition at Goddard.

Landsat-D Assessment System Tasks

LAS will be used by investigator teams in cooperation with the Project Scientist to analyze and verify the Landsat-D output products and to develop improved means of processing the data for a variety of applications. The

types of LAS tasks envisaged to accomplish the required support include the following:

1. Calibration and noise removal of TM data.
2. Position determination and image registration analysis.
3. Optimize data-processing procedures and algorithms.
4. Assess radiometric sensitivity, frequency of coverage, and data compression.
5. Assess the benefits from improved TM resolution.
6. Support system-level applications tests and demonstrations.

The above tasks will be performed to refine the data-processing and analysis procedures used with the TM data, to evaluate the incremental improvement offered by the TM over the MSS, and to improve the ability of users to apply the data in representative Earth-observations applications.

LAS System Design

A block diagram of the Landsat-D Assessment System is shown in Figure 5.3. The system includes a Digital Equipment Corporation VAX-11/780 minicomputer with 1 Mbyte of main memory, seven high-speed input/output ports, a low-speed input/output bus for connecting standard computer peripherals, a Floating Point Systems AP-120B pipelined processor, and a variety of other peripheral devices.

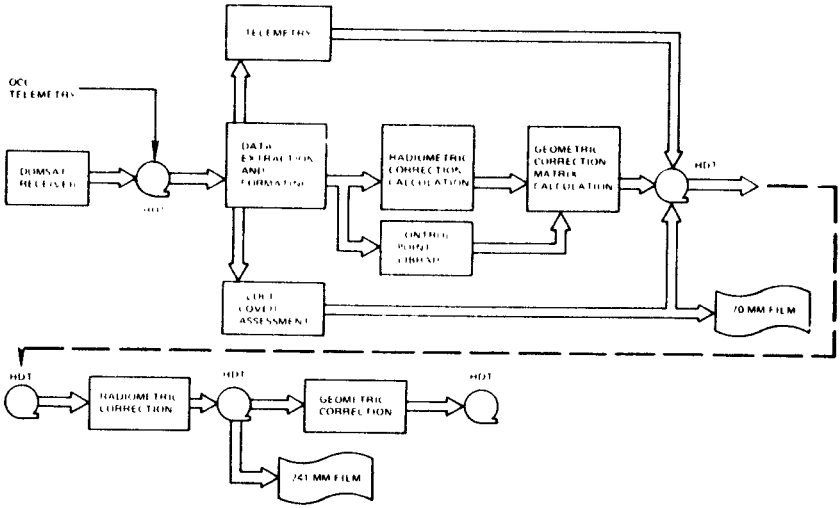

FIGURE 5.2 Landsat-D data-management system—product generation.

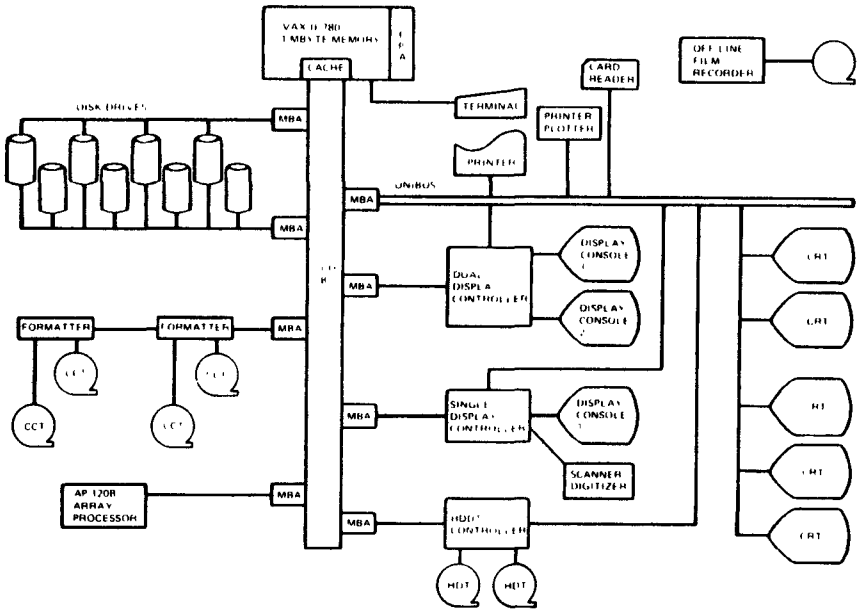

FIGURE 5.3 Landsat-D assessment system configuration.

The inputs to the system are TM and MSS data from computer-compatible tapes and high-density tapes as well as ancillary data from tape, film, maps, charts, and punched cards. The primary interface between the DMS and the Landsat-D Assessment System will consist of a high-density tape recorder for reading the 28-track high-density digital tapes containing radiometrically corrected data in band interleaved by line format and radiometrically and geometrically corrected data in band sequential format.

The LAS system is being designed to be able to process 350 Mbytes/day of image data on a regular basis and 700 Mbytes/day during peak loading periods. In addition, the system is being designed to be able to extract the required image segments from up to 12 (24 for the peak load) separate input scenes. In addition, 30 percent of the selected input data sets are required within 48 h of data acquisition, and the remaining data sets within 2 weeks of acquisition. A typical LAS user session would involve extracting selected image segments, displaying the segments, and performing a variety of analysis operations such as multispectral classification and applications modeling.

LAS Software

Since the intent of the LAS is to support teams of investigators in a variety of Earth-resources applications, the LAS software is being designed to provide a

user-oriented, interactive, and batch-processing system structured for flexibility and rapid response to user requests. There are three main components of the LAS software: the DEC supplied VAX/VMS operating system, LAS Executive, and the specific applications support software packages required by the investigator teams.

The VMS software is a multiple-user, virtual-memory operating system with an automatic program paging feature. VMS schedules and allocates system resources and includes a Fortran IV Plus compiler, a MACRO assembler, and extensive file management capabilities. The LAS Executive and the applications-support software packages are being designed and developed by Goddard personnel. These software packages will operate in conjunction with the VMS operating system and will make use of VMS capabilities wherever possible.

The LAS executive software will provide the interface between the user and the system and will manage all batch and interactive processing operations. The basic design of the LAS Executive includes adapting and extending the Video Communication and Retrieval (VICAR) system, developed at JPL, for use in the LAS. The VICAR extensions currently planned for the LAS will be implemented in Fortran IV Plus and include integrating a menu-driven executive, developing a procedure capability for easy execution of batch or interactive jobs with multiple processing steps, and adding data-management capabilities that make it easier to catalog, store, retrieve, combine, display, and manipulate data bases containing both image and nonimage data. The system will provide these capabilities while maintaining the VICAR command language structure and applications program linkage conventions intact.

Implications for Science

LAS is a step forward in that early planning may allow a working system to be in existence when Landsat-D data begin to be acquired. Such was not the case for Landsat-1, for instance. LAS could serve the purpose of assuring that Landsat-D data quality remains high, through quick looks at the data sets. LAS could also provide a thorough analysis capability on-site. Perhaps one of the drawbacks is that scientific input has not been widely encouraged.

III. SATELLITE COMMUNICATIONS

Introduction

There is a perception that we will be soon facing a shortage of communications space in the electromagnetic spectrum and a shortage of physical space for satellites (slots) along the geostationary arc in view of the United States. The spectrum shortage is the result of heavy usage of the 4-6-GHz

commercial band. The expectation is that the next higher frequency band, the 12-14-GHz band, will be similarly exhausted by the early 1990's. The perceived shortage of slots is due to the combined effect of inefficient use of the available spectral range, thereby limiting the communication capacity available in any one slot, and the need to separate satellites widely enough to limit electromagnetic interference. There is a plan for NASA to move back into satellite communications research and development, through a program called the 30/20 GHz Program.

General Description

Electromagnetic interference currently limits the spacing between geostationary satellites to 5° for 4-6 GHz and 4° for 12-14 GHz in the fixed services. If a mean spacing of 4.5° is assumed, there are available nine positions along the geostationary arc covering the continental United States (48 states) and 18 positions along the arc covering the entire United States (50 states)

TABLE 5.1 Near-Term Capacity of the Geostationary Arc in Gigahertz

Band and Reuse Factor	Bandwidth Allocation (GHz)	Capacity of Service Arc (GHz)	
		9 Slots	18 Slots
4-6 GHz (2x reuse by polarization)	0.5	9	18
12-14 GHz (2x reuse by polarization; 4x reuse by spot beams)	0.5	36	72
		—	—
TOTALS		45	90

TABLE 5.2 Near-Term Capacity of the Geostationary Arc in Transponders

Band	Capacity in Terms of Number of Transponders[a]	
	9 Slots	18 Slots
4-6 GHz	216	432
12-12 GHz	864	1728
TOTALS	1080	2160

[a] Equivalent 36-MHz bandwidth transponders each capable of carrying 1000 voice signals or 50 Mbps of digital data on 1 television signal.

for each of the two bands. The telecommunications and cable television markets have led to circumstances in which total U.S. coverage is desired from single-orbit positions for the 4-6-GHz band. This economic or market factor largely eliminates any possibility of multiplying the utility of this band through reusing the allocated bandwidth of 500 MHz in geographically separate areas. Thus, each position along the arc can provide 500 MHz of bandwidth at 4-6 GHz. Further, the use of orthogonal polarization overlaid on the basic polarization multiplies the effective bandwidth by 2. Thus, the present and future capacity of the arc at 4-6 GHz is 1 GHz per location or a total of either 9 or 18 GHz depending on the required service area.

The market at 12-14 GHz is less entrenched, and there is, therefore, more opportunity to use each arc location efficiently. The combined use of polarization and the use of modest spot beams will permit a multiplication of the allocated bandwidth. Present technology permits a reuse factor of 4 (including the factor of 2 associated with polarization reuse). Modest improvements over the next decade should increase the net reuse factor to 8. Thus, the

TABLE 5.3 Future Capacity of the Geostationary Arc in Gigahertz

Band and Reuse Factor	Bandwidth Allocation (GHz)	Capacity in GHz of Service Arc	
		9 Slots	18 Slots
4-6 GHz (2× reuse by polarization)	0.5	9	18
12-14 GHz (2× reuse by polarization; 4× by spot beams)	0.5	36	72
20-30 GHz	2.5	225	450
TOTALS		270	540

TABLE 5.4 Future Capacity of the Geostationary Arc in Transponders

Band (GHz)	Capacity in Terms of Number of Transponders	
	9 Slots	18 Slots
4-6	216	432
12-14	864	1,728
20-30	5400	10,800
TOTALS	6480	12,960

capacity of each position along the arc at 12-14 GHz is at present 4 times the 500 MHz allocated, or 2 GHz. With the modest extension alluded to above, the capacity will be 4 GHz per position. Therefore, the arc will carry (under the modest improvement assumption) 36-72 GHz, again depending on the service area. The numbers given above for the 4-6-GHz and 12-14-GHz bands are summarized in Table 5.1. The statement of capacity in terms of gigahertz can be restated by observing that each 0.5 GHz represents 12 transponders. A single transponder can carry 1000 voice signals or approximately 50 Mbps of digital data or 1 full-motion television signal. Thus, the formulation in Table 5.2 can be given.

The NASA satellite communications technology program will open a new frequency band (20-30 GHz) that has an allocated bandwidth of 2.5 GHz. While polarization diversity will not be available because of atmospheric propagation properties in this band, the use of advanced multibeam antennas will yield reuse factors up to 10. This gives a capacity of each arc position of 25 GHz in the 20-30-GHz band and, assuming 9 slots, a total arc capacity of

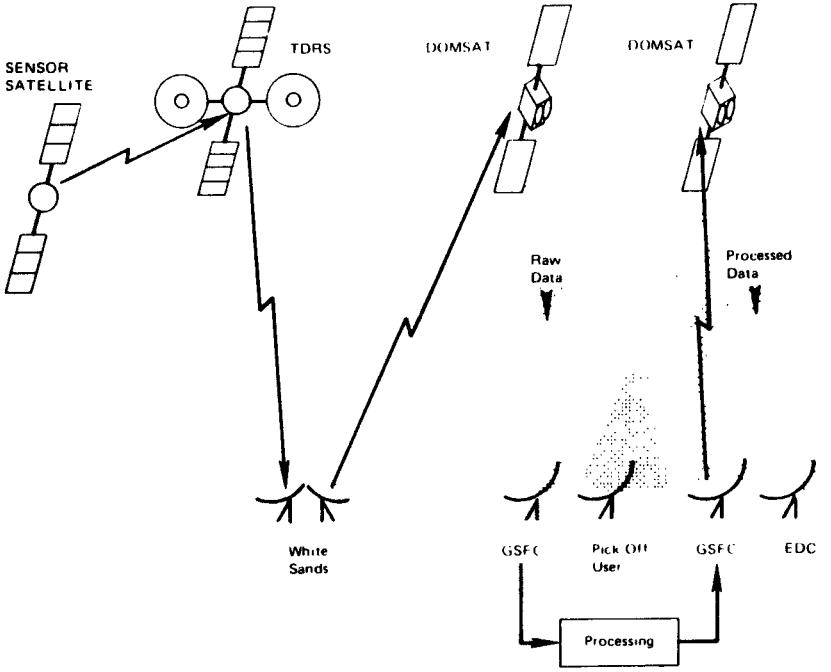

FIGURE 5.4 Mid-1980's data flow.

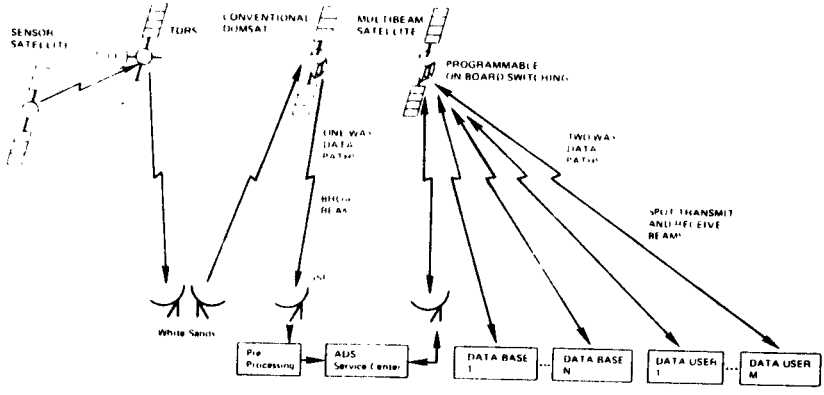

FIGURE 5.5 1990's data flow.

of 225 GHz and for 18 slots 450 GHz. Restating Tables 5.1 and 5.2 to include this new band gives Tables 5.3 and 5.4.

Market studies completed by NASA in 1979 show that even under highly optimistic assumptions the capacities shown in Tables 5.3 and 5.4 vastly exceed all projected requirements through the year 2000.

Implications for Science

Even considering signal attenuation due to severe weather conditions, one reaches the conclusion that it is feasible to consider the 1990's a data-communication network, based on data spotcasting that has unique properties in comparison with any currently in existence or in development. Consider Figure 5.4, which depicts the planned flow of Landsat sensor data in the mid-1980's. Data from Landsat are transmitted to a Tracking and Data Relay Satellite (TDRS) and from there to White Sands. The data are then transmitted via commercial Domestic Communication Satellite (Domsat) to GSFC for processing. The data can be intercepted easily in raw form at that point by a potential user. After processing, the data are relayed via commercial Domsat to the EROS Data Center (EDC). Again the data can be intercepted easily, and in this instance the data are in a processed form. The last intercept offers the sophisticated user the opportunity to bypass the EDC segment of the chain.

A configuration using spotcasting to intercept data is shown in Figure 5.5. After the same initial data flow to GSFC, data sources and data sinks are interconnected by a multibeam satellite with on-board switching under the

control of the Applications Data Service Center (see Section IV, below). For the sake of illustration, assume that a Time-Division Multiplex (TDM) System is combined with a Time-Division Multiplex Access (TDMA) System. A data user requests a set or sets of data from one or more data bases. The request could be in the form of a serial bit stream containing segments devoted to such things as request priority, data source addresses, required precision or resolution, and required scale of format. A Service Center would process, route, and, using the spotcasting concept, transmit the data request as required. In the simpler, receive-only mode, individual scientists or small groups cannot yet afford the $10,000 cost associated with direct data reception. Technological developments might well reduce this cost significantly in the near future, thereby increasing its use. With the present $10,000 cost of receiving terminals, larger organizations such as science institutes could afford to acquire the capability to operate and to receive data directly in the TDRS era. Acquisition of such a capability would address the problems of long delays in data delivery to the user and the high cost of data distribution.

It is more difficult to forecast the cost of more complicated two-way systems, but at present comparable 4-6-GHz systems would cost several hundred thousand dollars, well beyond the budget for most scientific investigations. It is possible that future developments might result in a substantial lowering of costs, making the technology more attractive to the average user.

IV. APPLICATIONS DATA SERVICE (ADS)

Introduction

Many NASA scientific and applications activities require integration of inputs from a variety of sources—sensor data, ground truth, historical, ancillary, and *in situ* data. Since the sources of space data are not available in a single consolidated catalog or directory and since data formats differ, standards for data quality differ, and costs of processing data (particularly image data) are larger, NASA is currently developing an Applications Data Service (ADS). It is the objective of ADS to provide timely, affordable access to readily usable, multisource data products and services. ADS is intended to serve a broad cross section of the user community. Our discussion will address its adequacy with respect to scientific investigations rather than to applications users.

General Description

The ADS would be a communications network interconnecting all NASA and NASA-related applications data producers and data users. ADS is intended to

meet the data-transfer needs of all clusters of data producers, where a cluster is used to indicate a group of scientists operating in a discipline requiring one or more sources of data. Each data user would have access to all data not just to data produced within his particular data cluster. A service center would provide user support in finding and using data. The ADS would be designed to make all data bases at various different locations appear as one transparent data base for the user. The ADS would be modularly expandable to meet future needs. It is intended to avoid duplication of effort and costs by individual data clusters.

The specific functions to be performed by the ADS are as follows:

1. Develop and maintain a network to access and disseminate data.
2. Develop and maintain data directories and catalogs.
3. Develop and maintain data and data systems standards.
4. Provide data preparation and integration services.
5. Provide a user charge-back mechanism.
6. Allow systematic, evolutionary growth of Office of Space and Terrestrial Applications (OSTA) data systems.
7. Advance research and development in data management and usability.
8. Foster technical compatibility with a national data service.

The following functions would *not* be performed by the ADS:

1. Produce or maintain data bases.
2. Analyze data or extract information.
3. Provide or pay for user terminals, user communications charges, or producer data facilities (beyond pilot demonstrations).
4. Require substantial programmatic changes in existing producer data systems or procedures.

Implications for Science

With its data directories and catalogs, ADS will provide the capability of determining what data exist, where they are located, what their general characteristics are, and possibly what their quality is. One important feature is that the directory assistance may identify previously unknown data sources for many scientific investigations. One limitation may be the lack of browse capability; there is no mechanism for second-stage screening of data sets identified from the directories and catalogs. There are also no guarantees that existing data bases will have adequate directory information for use in searching for specific data sets. The ability to use ADS to access data will be only as good as the completeness of the constituent data bases.

The availability of the ADS capability should lead to greatly improved distribution of data. ADS can minimize delays in acquiring relevant data. Also, through the use of documentation standards and possibly data formatting standards, ADS can minimize the hardships involved in accessing different data sets and thereby improve and broaden the scientific uses of space-acquired data. It should provide a mechanism for assembling scientific data sets extracted from a variety of different data bases, spanning a variety of different formats and architectures. This should eliminate the need for the scientist to make costly, time-consuming conversions of data from one computing architecture to another.

This improved distribution of data could be diminished by excessive delays in the creation of user data sets; the ADS approach must be able to provide a reasonable response time to user data requests. It is not clear that there will be sufficient throughput capability in ADS to handle a significant volume of potential users from all scientific disciplines and commercial and other applications users. If the scientists must wait as long or longer to receive a desired data set through ADS than is now required by individual initiatives, then it will not adequately serve the scientific user community.

An additional concern is that the terminals and data links required for the user of ADS, as well as the establishment of the ADS capability itself, may involve high user costs with respect to typical computing budgets on scientific investigations. Service must be provided at reasonable costs to scientific users for the ADS program to be of value to the scientific community.

One additional feature of the ADS approach is the distributed nature of the information extraction process inherent in the approach. If there is sufficient distribution of computational capability for scientists to do their computing locally, then the ADS will facilitate the exchange and distribution of data for the scientists to use in their analysis. This should encourage more widespread use of the data, as should the standardized interfaces provided through the ADS. On the other hand, one can also foresee some situations in which the ADS network could be used by scientists without a local computing capability to make use of a central computer for data processing. For small users this approach might prove cost effective, although image processing would probably be too costly, given present data transmission costs.

V. NASA End-to-End Data System (NEEDS)

Introduction

Space-oriented information management systems for the 1980-2000 era face two critical problems: (a) near-real-time handling of space-derived data and (b) cost-effective analysis and distribution of data and information to the user.

The coming decade may be characterized as an era of correlative research in which space-derived data are merged with data from other sources and integrated to provide the desired information. The system architecture will facilitate exchanges between data sources, will allow users easy access to data stores, and will permit near-real-time control of observing parameters.

In the late 1980's and subsequent years, data system architectures must provide operational capabilities and serve a broad community of users. Typically, such systems will include dedicated mission facilities to provide direct data interpretation services; a distribution network to service the users groups; and appropriate archiving, processing, and control facilities.

General Description

A comprehensive NASA End-to-End Data System (NEEDS) concept is being evolved as part of NASA's activities to provide the technology for an operational information management system. The currently funded NEEDS program emphasizes those aspects of an overall system that involve acquisition through archiving of data and focuses on real-time data management. A later phase, planned for initiation in subsequent years, will pursue that part of the system that involves the distribution and analysis of data, as well as sensor control technology. The stated goal of the NEEDS program is to establish systems and operations concepts and technologies that will increase the effectiveness and efficiency of the End-to-End Data System by two orders of magnitude over current practices.

Major elements of the currently funded NEEDS program include the following:

1. Development of an analysis program for simulating system concepts so that key technologies can be evaluated to assess gains in system performance.
2. Development of information-adaptive systems utilizing advanced spacecraft processors for on-board preprocessing, feature identification, and selection of data sets for distribution. These systems are intended to provide real-time processing, minimizing the amount of data sent from spacecraft to ground.
3. Development of modular data transport systems affecting on-board data system architecture and control software. This involves the development of standardized packet telemetry formats containing sensor identification; sensor data; and auxiliary data such as time, orbit, and attitude. It also involves the development of spacecraft microprocessor-based distributed data-systems architecture to achieve improvements of a factor of 100 in data access time and reduction of a factor of 10 in cost of data retrieved.
4. Development of programs designed to increase efficiency of ground-based data-handling facilities. These technologies include data-base management systems, including modular microprocessor-based data control systems.

5. Development of archival mass memory systems capable of 10^{14} bits of on-line storage and up to 10^{15} bits of archival storage at a data transfer rate of 50 Mbps.

6. Development of advanced, high-speed, special-purpose processors for complex data sets. Examples are the Massively Parallel Processor to process two-dimensional image data in near real time and a digital radar image processor.

Future plans call for NEEDS to examine ways of increasing the efficiency by which data are converted to useful information and placed in the hands of the direct user. Elements of this phase include the following:

1. Development of technology for on-board information calibration and for on-line remote control of spaceborne instruments. This is intended to allow the user direct access to the data source and greater control of the data acquired to satisfy his needs.

2. Development of cost-effective processing techniques such as information compaction and compression algorithms, the definition of standard interfaces between the user and the data-distribution system and the data bases, and the design of universal information structures. These are intended to improve the efficiency of interchange between the user and the processing system.

3. Development of low-cost remote terminals to allow direct user access to the data system, of high-rate data links for *local* information transfer, and of networking technology for multiuser access to spaceborne data sources.

Implications for Science

The specific activities of the NEEDS program can be effective in dealing with several of the data problems identified in Chapter 1 of this report. Specific examples of identified problems that NEEDS is directed toward handling include the following:

1. Insufficient current capabilities for on-board preprocessing to reduce significantly the amount of data that must be returned to the ground.

2. High cost of current computer technology used for preprocessing.

3. Long delays between the receipt of data on the ground and delivery to the user.

4. High cost of data distribution sometimes preventing its acquisition by small users.

5. Lack of standardization of data formats requiring considerable efforts to allow data to be used.

6. Insufficient auxiliary data such as time, attitude, and bit provided by data archives.

7. Inadequacy of current mass storage technology to archive data at sufficiently low cost.

8. Wide variety of man-machine interfaces with little standardization of hardware and software.

9. Current technology often not exploited or implemented in present data systems.

The primary activity of the NEEDS program is directed toward the front end of the data chain—the acquisition, distribution, and archiving of data. These are intended to meet major deficiencies that exist at present and that are likely to become more severe as data rates increase in the next decades. *As such, these programs are most deserving of our support.*

At the same time, there are uncertainties associated with these new programs. If substantial financial resources are required to carry out these activities, the funding available for the actual extraction of scientific information from the data may be adversely affected. For example, the success of some aspects of the NEEDS program will depend on the availability of low-cost, remote terminals for individual scientific users and on the availability of low-cost data lines or links for data distribution. These represent somewhat uncertain factors at present.

It should also be recognized that while standardization and simplification represent directions that are generally desirable, there are scientific activities for which these directions may not be effective. For example, the transmission of raw attitude information to the ground from an astronomy satellite may allow the data to be processed several times (*post facto*) with increasing precision in subsequent runs due to a better understanding of uncalibrated systematic effects. Experience indicates that unexpected effects can compromise the accuracy of scientific information if only preprocessed results are transmitted to the ground. Also, there is some risk of real data loss if only selected data segments are preprocessed and transmitted to the ground; it may simply not be possible to determine what information is required in advance.

The advertised cost and technical efficiencies of standardized and simplified data systems must be balanced against special requirements of the user community on a case-by-case basis. An environment should be created that is capable of supporting both approaches. This same consideration applies to the development of scientific software and to the use of specialized data-processing computers. Where the centralization fills a void and the scientists actively involved in a discipline are consulted in a real way, substantial gains can be achieved.

VI. SPACE SCIENCE DATA SERVICE (SSDS)

Introduction

The Space Science Data Service (SSDS) is seen as a computer system, perhaps directly associated with the Space Applications Computing Center (SACC) at GSFC and capable of providing coordinated data-processing and data-storage facilities for the scientific community served by NASA's Office of Space Science (OSS). In terms of satisfying needs of investigators, two configurations have been deemed feasible. One configuration involves full centralization of computer facilities with terminals linked to experiment teams. The second configuration is called a distributed environment, with some centralized computers and a mix of terminals and minicomputers to handle the needs of investigators. A decentralized system comprising individual minicomputers and no central mainframe was rated unsatisfactory, primarily because of the total absence of a large mainframe for longer-running processing jobs, for archival mass storage, and for timely exchange of data bases.

General Description

The centralized option would comprise large-scale mainframes, limited to two at any single location because of administrative issues. Processors would be selected that are compatible with existing SACC computers to minimize the impact of changeover. Processors would have major upgrade potential to provide for future SSDS expansion. Mass storage requirements would be at least 10^{11} bits on-line and 5×10^{11} bits for archival storage.* High-speed data connections (up to 50 Mbps) could be installed between the SACC and the various GSFC facilities used for operations control and other data processing. Outside users could be connected to the mainframe through conventional terminals with minimal graphics and output capability or through intelligent terminals with interactive graphics capability and hard-copy printers. In addition, there would be a category of users who require a greater computational capability for interactive graphics and/or software development. These users would probably be restricted to local or on-site processing, because of limitations in data communication lines.

The distributed option would provide a central facility with less computing power but with an archival mass storage system sufficient for a central data base. Processor time would be made available for outside users as needed and as available. The basic system would be sized to meet the SACC needs directly. Eventually, data links would be configured to connect the central computer to the individual minicomputers so that data could be transferred

*These volumes are relatively low because they involve only nonimaging OSS programs and no high-data-volume applications programs.

without mailing magnetic tapes. Major innovations in data-communications services will be required before this option becomes cost effective.

The present plan is to proceed with the hardware upgrade of the SACC facility in the near future. A detailed conversion plan will be required to cover the hardware changes and the updating of the operating system. The upgraded SACC could then be interfaced with other SSDS facilities as they materialize. Individual SSDS facilities may be developed before the SACC upgrade is complete. An example is the upgrade of the Atmospheric Explorer (AE) computer required for the Dynamics Explorer Project. Among the considerations of this current upgrade activity is the feasibility of expanding the upgraded system to meet future SSDS requirements, such as future major processor upgrades and archival mass storage of up to 10^{14} bits.

Decisions will also be made about high-speed data links between SSDS facilities. Procedures will be established for storing and distributing large volumes of scientific data. Interfaces between SSDS facilities and independent investigator computers will be studied. A final decision will be made between a centralized and a distributed SSDS system, with the requirements of scientific investigators being taken into account.

Implications for Science

Implementation of the SSDS would have a substantial impact on several of the problem areas described in Chapter 1. Those problem areas that could be alleviated by SSDS include the following:

1. Insufficient long-range data-system planning, including adequate funding and protection of funding.
2. Delays in delivery of data to scientific users. High costs of data distribution. Failure of PI's to provide data to data centers.
3. Multiplicity of formats limiting use of different data sets and archived data.
4. Inadequate documentation of software. Lack of transportability of software. High cost of software development. Incomplete software at time of launch.
5. Requirement for scientists to travel to computer centers.
6. Inadequacy of current mass store technology to archive all NASA data at acceptably low cost.
7. Multiplicity of man-machine interfaces required for interactive processing on a wide range of systems.
8. Failure to exploit current technology in present data systems.

With respect to long-range data-system planning, the SSDS should be helpful in either of its configurations. To the extent that the SSDS studies lead to

a better understanding of data requirements for the next decades, they should help to create an environment better able to provide the technical and fiscal resources required for adequate data-system planning and development.

The distribution of data to scientific investigators and the multiplicity of formats encountered in different data sets are areas that could also be helped by the SSDS. The SSDS could be used to create a centralized data base. This would require the development of a satisfactory data-base management system with a minimum number of data formats. The centralized data bases would facilitate transfer of data to the National Space Science Data Center for distribution to secondary users after the PI-proprietary period has elapsed. The availability of the data and the standardization of formats should greatly enhance scientific utilization of the data, including synergistic studies using multiexperiment data. Coordinated Data Analysis Workshops (CDAW) would also be enhanced under this approach.

Costs of duplicating, handling, and mailing experimenter data tapes would be minimized in the centralized option under which data sets would be accessed and analyzed at the central computer. In the distributed system, advances in data-communications capabilities would be required before the data tapes approach could be replaced. In particular, present costs and data transfer rates prohibit practical transmission of imaging data sets. This situation might be somewhat ameliorated by the division of the central computing facility into several nodes and/or by the establishment of regional centers to share the costs of data lines among individual investigators.

The implications of the SSDS for software development are somewhat more uncertain. Proponents of the SSDS have described an approach under which the SSDS would be responsible for the operating system and for general software, with individual investigators responsible for specific applications software. If users request the development of specialized software by the SSDS, funds would have to be provided by the appropriate project office. Advantages claimed for this approach include cost savings through the sharing and reuse of software, greater compatibility and transportability of software because of the establishment of standards and the use of common hardware, and protection against the disruption caused by conversion of a non-NASA institution to a new computer configuration and/or a new cost-accounting system. These benefits would obviously be more easily realized in the centralized rather than the distributed approach, although the setting of standards and the sharing of software would be possible in either case.

The potential drawback of this SSDS approach to software development would involve the role of the individual scientists or teams in this picture. The approach could significantly limit the active involvement of those people most knowledgeable about the software requirements. Particularly in the centralized option, situations could develop that do not permit the effective in-

volvement of these key scientific personnel. The "system" could lose sight of its original objectives and establish unreasonable ground rules and operating procedures that limit the effective development of specialized software. NASA software subcontractors might not understand the technical requirements as well as a staff associated with the PI's, and the resultant software might not be adequate to do the job and might not be developed in a timely fashion. Decisions to adapt existing software or develop new software might be influenced by nontechnical management issues and might not always be made in the best interest of the individual projects. In the distributed mode, individual scientists (or Scientific Data Management Units, described in Chapter 8) would retain more control over software development.

There are also a range of implications of the SSDS for the distribution of computing capability and of mass store archival capability. In the centralized option, the individual investigator would use the central computer in one of the three modes described above. The central facility would provide more processor capability and more data-storage capability than a typical off-site user's system. The central computer would have the potential for substantial growth and would involve a minimum duplication of resources. This would be of significant value to scientific investigators, provided a viable set of operating rules are established to permit equitable allocation of processor, storage, and output devices. In other words, individual PI's would need assurances that their needs would not be neglected in a centralized system and that the system would also have some flexibility for handling special requirements. This has not been the case in the past for individual outside users of SACC computing capabilities. A centralized facility would also have limitations for the remote processing of imaging data, thereby requiring the on-site presence of investigators wishing to carry out such studies for large data sets. There might also be some limitations on bulk printer outputs, requiring some data printouts to be mailed to individual users. On the other hand, the centralized facility would reduce the variety of man-machine interfaces that individual investigators are often forced to deal with at present.

The distributed system would have the advantage of providing more control to the individual scientists or teams. An individual could use an individual minicomputer for special applications; could set his or her own internal priorities and schedule usage accordingly; and would have immediate access to the system when he or she required it. The scientist could work at his or her home institution most of the time, thereby minimizing requirements for travel to a central computer facility. A limitation on the distributed approach would be the requirement to continue the mailing of data tapes in most instances, since data lines are not adequate for data transmission for high-data volumes. Other limitations include the possibility that local minicomputers would not have adequate computing power, adequate data-storage capability,

microfilm capability for output, and reproducible graphics capability. Links to the central computer facility might be required under these circumstances, and a variety of distributed machines would mean extra costs and activity required to establish compatible interfaces to the central computer. A potential problem would be the remote, infrequent user of the central system who could not receive sufficient priority to complete his or her work on schedule. Cost analyses have suggested that the distributed approach would be 7-12 percent more expensive than the centralized approach, but the analyses presented to CODMAC are not completely convincing on this point.

The ability of the SSDS approach to affect favorably the exploitation of current technology in present data systems is subject to a range of arguments. The approach currently planned for any SACC upgrade would provide the capability for significant upward expansion in the future. The presence of more than one processor will provide for improved hardware redundancy as well, resulting in greater reliability and system availability. At the same time, the problems that have been encountered during the planning of the present SACC upgrade and the lack of flexibility in utilizing systems different from those currently in use are indicative of the limitations in using new technology that occur when a relatively expensive, centralized computing facility is established. More specifically, rapid technological evolutions are often more readily implemented with relatively smaller, less costly computer systems than with larger, more expensive systems. This would appear to be a major argument in favor of a more distributed, less centralized SSDS approach.

VII. JET PROPULSION LABORATORY PLANETARY END-TO-END INFORMATION SYSTEM

Introduction

The Jet Propulsion Laboratory (JPL) is in the process of significiantly upgrading its computing capabilities. The new computing facility is intended to support space sciences, including the functions of mission operations, data collection, and data reduction. The JPL End-to-End Information System (EEIS) is being considered for streamlining mission operations and data collection, for reducing data, and for archiving data sets. The EEIS would be concerned primarily with planetary data sets. This overall computing facility will directly influence the manner of conducting mission operations and of on-site data handling in the coming decade. The possibility of links for outside scientific users may also greatly influence scientific data handling, reduction, and analysis.

General Description

The JPL computing facility will consist of three Univac 1100/80 computers, with possibly a high-speed link to the JPL Image Processing Laboratory IBM 360/67 computer. Mass data-storage capacity on the system will be approximately 6×10^{12} bits. An option under consideration at present is the use of data links to connect remote users to the system.

The EEIS is a long-range plan for the operation and control of remote-sensing spacecraft, for the real-time acquisition and delivery of data, and for the utilization of these data for the extraction of scientific information. The architecture currently envisaged for this system is represented schematically in Figure 5.6. The various facets of project-dedicated satellite operations are indicated at the top of the diagram, while the elements of a multimission control and operations network are shown just below. The lower left portion of the figure shows the data flow from the spacecraft sensors through the real-time data-acquisition network. The lower right portion indicates the use of the computer facility to establish a multimission planetary (solar-system exploration) data base and also suggests the possible interactive links to the end-user (scientific) community. This final stage represents the point at which the primary scientific data analysis would occur.

Implications for Science

The overall JPL computing facility has been designed primarily to meet the load projected for mission operations and primary data reduction (preprocessing before data distribution to scientific users) over the next several years. As such, this system will be providing an essential service; however, its effectiveness for science depends on the details of the interface to the scientific users.

In the extreme case, the mainframes could be used as centralized data banks. Individual investigators would remotely interrogate the data base, implement their computation using the remote mainframes, and have their output transmitted to them by data links. One advantage of such a system is that the relevant data would be archived at one data center location. Also, standardization and documentation of data formats and software systems would probably be facilitated. The availability of substantial mass data storage capacity would result in efficient data retrieval for users.

One possible drawback to this approach is the lack of adequate interactive processing for off-site users, particularly those users who require intermediate data sets (e.g., image-processing users). The problem is that the cost of high-speed data links between the centralized facility and the remote user becomes prohibitively expensive. For instance, a 52-kbps link between Pasadena, California, and St. Louis, Missouri, would cost $6500 per month. With such a link

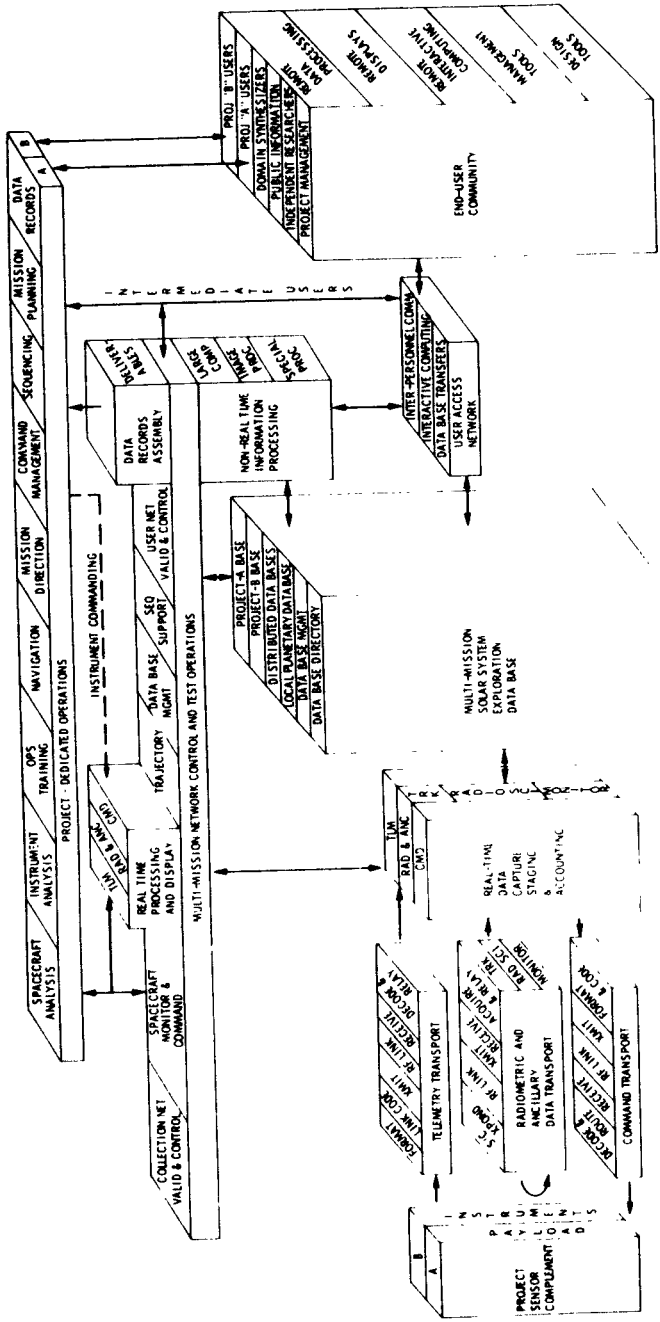

FIGURE 5.6 An architecture for the JPL planetary information system.

it would take about 2½ min to transmit a single Viking Orbiter image (1000 x 1000 pixels with 8-bit dynamic range), assuming no system delays. For intensive processing, the time delays and the high cost makes such processing impractical at present.

Another potential disadvantage to the centralized approach is that users will be dependent on the availability of the centralized computers. In a system that is concerned with mission operations control and real-time acquisition of data, the scientific user will receive the lowest priority. If the system becomes fully utilized (or oversubscribed) the negative impact on the scientific users could be serious.

There is also some potential problem in the software development chain. It is our contention that the active involvement of discipline scientists is essential for the timely and cost-effective development of scientifically usable software. The centralized computational approach may find means of effectively integrating scientists not currently at JPL into this activity. On the other hand, the problems noted earlier in this report with the Seasat program suggest that merely installing scientists as advisors on a program will not be sufficient to guarantee successful software development. At the same time, Chapter 3 of this report provides examples of programs in which individual PI's or scientific teams responsible for software development have been more successful, although a few counterexamples can probably also be identified.

VIII. CONCLUSIONS

It is apparent that there are real problems and limitations with the present methods used to obtain, distribute, reduce, analyze, and archive space-acquired data. Thus, NASA initiatives to deal with these problems are essential. The descriptions of these new systems, however, raise several major concerns:

1. The first concern is the potential cost of the programs. If the costs become too high, the opportunities for carrying out meaningful scientific studies may disappear because of lack of remaining funds. For these new programs to be effective in maximizing scientific returns within limited financial resources, the overall cost of the new programs must not come from funds needed for scientific data analysis.

2. Another concern involves the centralization of scientific activity within various NASA centers as a part of these new initiatives. Advantages and disadvantages of current methods were deduced by studying what worked and what did not. One conclusion is the importance of scientific involvement in mission planning, in software development, and in data utilization. The new NASA systems described have the potential for greatly restricting the leader-

ship or even the involvement of non-NASA scientists in these activities. In those disciplines where the scientific community is not currently organized, however, the centralization of these data activities within a NASA center may well be a major improvement.

A major recommendation of CODMAC is for the active involvement of scientists in all elements of the data chain—planning, collection, distribution, processing, analysis, and archiving. The centralization of many of these data functions within NASA centers, if several of the new initiatives are implemented as currently described, could seriously and unnecessarily limit the involvement of non-NASA scientists in many of these areas. For example, the Seasat experience described in Chapter 3 details the data problems, the lack of adequate funding, the absence of appropriate algorithms and software system for data reduction and analysis, and an inadequate system for the delivery of data from GSFC to JPL and on to the scientific users. This has effectively thwarted the analysis of SAR imagery.

To the extent that speculative cost analyses have been used to generate arguments for centralizing various data activities, CODMAC suggests that additional factors, such as the benefits of active involvement and leadership by experienced members of the scientific community, must also be weighed with great care before decisions with far-reaching consequences are made. This point is further addressed by our specific recommendations concerning the role of the scientists in the overall data chain.

3. Another concern involves the apparent lack of overall planning as to how these new NASA systems will interact with each other. We see study activity in several NASA offices and centers (e.g., OAST-NEEDS, OSS-SSDS, OSTA-ADS, JPL, GSFC) that promises new data systems, yet we see no central direction by NASA to organize each of these separate elements into a larger design. It appears that each of the elements is being pursued independently, and the possibility of competition among them, or even of conflict among them, is quite real.

4. The final concern involves the interaction between NASA and industry. There are technologies that are commercially available at present that appear to address some of the areas that NASA is studying. For example, commercial networks (such as Telenet) seem entirely capable of providing some of the communications networking that we see NASA pursuing in concepts like ADS. We hope that NASA is fully aware of the commercial products and services available and plans to use them rather than attempting to re-invent them.

6
Trends in Computing Facilities

I do not say: Science is useful because it allows us to construct machines; I do say: Machines are useful for by working for us they permit us more time to study science.

JULES HENRI POINCARÉ (1854–1912)

I. INTRODUCTION

Because of the rapidly decreasing costs and increasing capabilities of minicomputers and peripherals, a number of research groups now conduct much of their data processing with relatively small, but dedicated, stand-alone systems.

There are a variety of ways in which computing and data processing are conducted:

- Centralized systems—where processing is done on a shared computer located in a single location.
- Decentralized systems—where facilities consist of a number of essentially independent computers, geographically separated.
- Distributed systems—where several computers are interconnected so that users can access and share their multiple resources in a routine manner.

There are also a number of variations between completely centralized and completely decentralized systems that, in a general sense, can be considered

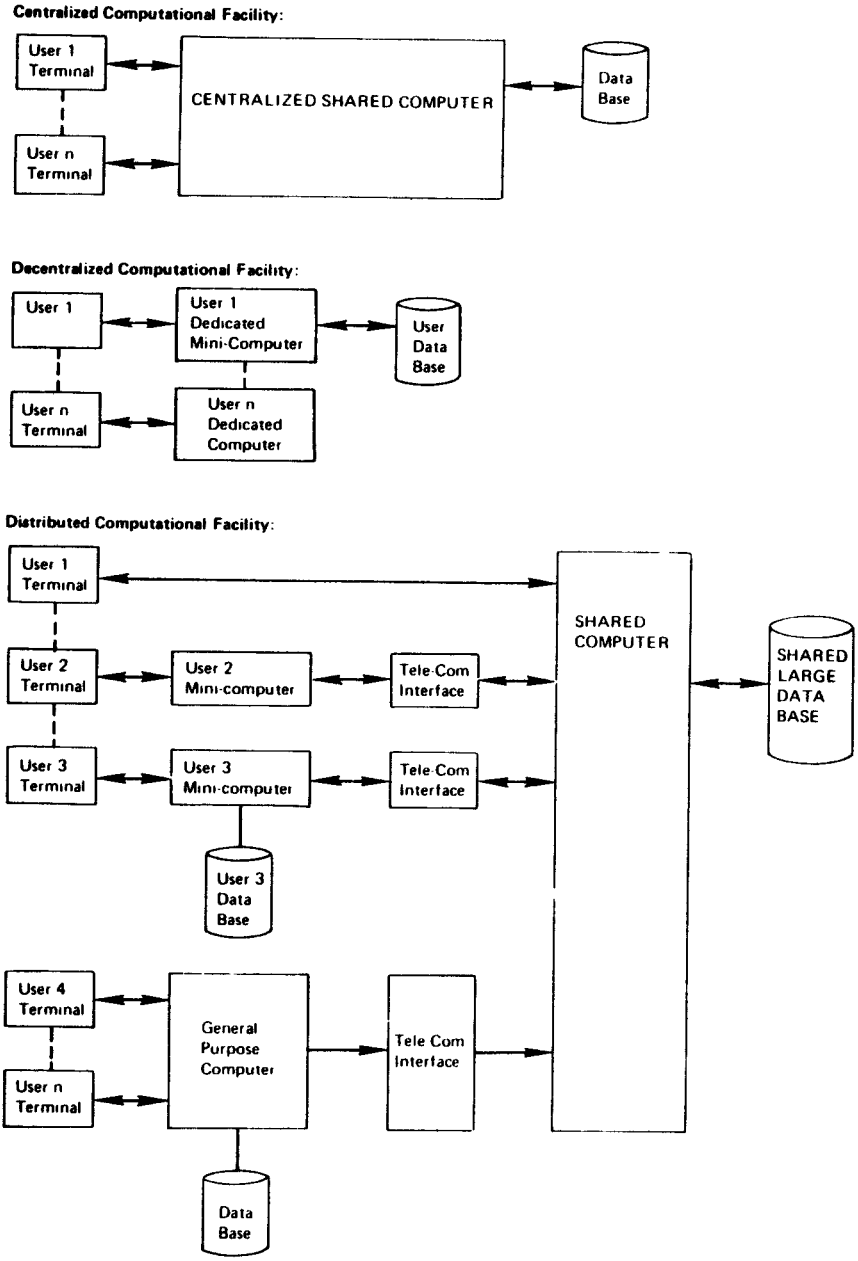

FIGURE 6.1 Combinations of simple data distribution and computational configurations.

to be distributed systems. For example, a centralized facility can have remote users with terminals connected by telephone or other communications media, or local computing systems could access data from a large centralized storage system. Figure 6.1 illustrates the variety of combinations of computational and data distribution facilities in use today.

In this chapter, we discuss the use of all three types of facilities for space-science data processing, including examples of existing systems and trade-offs between (1) costs, (2) differences in CPU capacity, (3) ease of maintenance and development of archives and software, and (4) efficiency of use for centralized, decentralized, and distributed computing facilities.

II. DECENTRALIZED FACILITIES

There is a large and increasing number of research groups that own or share a dedicated computing facility, where the CPU capabilities are significantly less than those of most modern mainframes. Yet, these facilities are able to process data in a timely fashion because they provide dedicated interactive processing and thus provide efficiency of system use. In addition, most of the systems are committed to a given research group so that they are able to process data on a full-time basis.

We discuss three examples of computing facilities that are, in effect, part of a decentralized system for processing space-sciences data:

1. The Laboratory for Atmospheric and Space Physics (LASP) at the University of Colorado has a minicomputer-based system for processing ultraviolet spectral data from the Pioneer-Venus, Voyager, International Ultraviolet Explorer, and other spacecraft. The system operates in a multiuser, interactive operating environment. It interactively supports six terminals and an image/graphics display. Although currently a stand-alone facility, this system temporarily was linked to the NASA Ames Research Center via a high-speed phone line for the first 245 days of the Pioneer-Venus mission (in this mode it was in effect part of a distributed computing system). The link was needed to acquire primary sensor data rapidly in order to process it in a timely fashion. Remaining data were distributed on tapes, which usually arrived 3 weeks after receipt of the data from the satellite. The LASP system is a good example of the use of a relatively inexpensive, dedicated local computing facility, with interactive processing capabilities to maximize throughput. It is also an excellent example of mixing decentralized and centralized facility concepts as a distributed system. Access into the NASA Central Facility can be implemented when needed. The elements of the LASP system are given in Table 6.1.

TABLE 6.1 Laboratory for Atmospheric and Space Physics, University of Colorado, Six-User Interactive Computer System

Quantity	Item	Model
1	PDP-11/34 Computer with: 25.2-Mbyte Disk 128-kbyte memory console Terminal RSX software license	SM-32LLB-CK
1	Floating-point processor	FP11-A
1	Cache memory	KK11-A
1	128-kbyte additional memory	MS11-LB
1	Eight-terminal interface	DZ11-A
1	Line printer and control	BCS
1	Tape controller	Datum
2	9-Track, 75-in. per sec tape drives	Kennedy 9100
1	80-Mbyte disk system	Diva DD-54
6	Graphic terminals	Tektronix 4006
2	Hardcopy units	Tektronix 4631
1	*X-Y* plotter	Tektronix 4662
1	Color image display system	Grinnell GMR27
1	IDL software license	

TABLE 6.2 Regional Planetary Image Facility, Washington University, User Interactive Image/Graphics System

Quantity	Item	Model
1	PDP-11/34 Computer with: 256-kbyte memory Floating-point processor FIV+, Pascal RSX-11-m operating system	
1	9-track, 800-bpi tape drive	
3	MIME 1 video terminals	AMR-27
1	Grinnell imaging system with monitors	CONRAC SNA 17/C CONRAC 5211C19
1	Able interface	QUADASYNC/B
2	80-Mbit disk drives with minicomputer technology interface	CDC9762 SMC-11
1	Disk player	Magnavoc VH800 Serial Interface WICAT
1	Color camera system	Matrix System

2. The Regional Planetary Image Facility (RPIF) at Washington University has a minicomputer-based system for processing digital data from imaging sensors. The system was purchased largely because of a set of frustrating experiences in trying to process Viking data at JPL, where science processing necessarily was pre-empted by mission operations processing. Arguments were made that a small, dedicated system, even having slower CPU and disk/tape speeds would be more efficient than utilizing JPL's Image Processing Laboratory facilities if used in an interactive environment. In addition, the costs per product were projected to be lower on this type of dedicated minicomputer system. In fact, costs per product have run about an order of magnitude less on the minisystem compared with JPL's existing computer, including capital costs and maintenance of the system.

The heart of the RPIF system is an interactive graphics/image display and interactive terminals. Just as with the LASP system, this configuration maximizes the system efficiency for scientific investigations. In addition, the RPIF system supports a videodisk player for the storage and display of analog images. There are 52,000 525-line × 525-element images on one side of a videodisk. The player was purchased at a cost of $774, with an additional cost of $1800 for installation of a serial interface unit. Thus, for about $2600, the player is able to store and retrieve for "quick look" up to 52,000 planetary images. The RPIF system uses a version of the VICAR image-processing software used at JPL. A description of the system is summarized on Table 6.2.

3. The Einstein Data Processing Facility (EDPF) at the Harvard-Smithsonian Center for Astrophysics consists of two relatively high-powered minicomputers that are used to support various functions associated with the Einstein Observatory. These functions include tasks associated with experiment development, mission planning, software development, data analysis, and data archiving/distribution. The facility serves as a focal point for the mission. Fixed-cost constraints on the mission led to the decision to acquire a stand-alone minicomputer-based system. EDPF has two dedicated minicomputers. One computer is used for mission operations and primary data reduction. As such, access is limited, and it serves as an "automated data pipeline." The other system was designed to be more interactive, including incorporation of an interactive image/graphics capability. Data are routinely transferred between these machines. The interactive system typically supports some 20–30 scientists (residents, visitors, and guest observers). The system configurations are given in Table 6.3.

Pros and Cons of Decentralized Facilities

Many other examples of use of minicomputers in a decentralized fashion can be cited in addition to the three given above, cases in which a relatively small,

TABLE 6.3 Einstein Data-Processing Facility Computer, 20–30 User Interactive System

Quantity	Item	Model
Production (Data Reduction) Computer		
1	Eclipse with 448-kbyte memory	S/230
3	Disks (200 + 200 + 100 Mbytes)	
5	Tape drives (800/1600 BPI)	
1	Printer/plotter	Versatec
7	Terminals (9600 Baud)	
1	NASCOM line (9600 Baud)	
1	Connection to user computer and to display system	MCA
User (Analysis Computer)		
1	Eclipse with 512-kbyte memory	M600
2	Disks (200 Mbytes each)	
2	Tape drives (800/1600 BPI)	
1	Printer/plotter	Varian
1	Lineprinter (600 lines per minute)	
12	Terminals (9600 Baud)	
1	Connector to production computer and to display system	MCA
1	Measuring engine (two-axis digitizer)	
1	Display system	Lexidata

dedicated computing facility has been configured to process space-science data. However, even the three examples cited above serve to show the advantages and disadvantages of distributed versus centralized facilities.

The *advantages* of the decentralized or stand-alone system include the following:

1. It is a dedicated facility, usually under control of an individual or a small group of users. Thus, the user is not competing with other tasks for resources or priority as in a central facility. Nor is the user dependent on the schedule of a centralized facility, which may not be optimized for that user. Most of the dedicated facilities operate around the clock.

2. System acquisition and maintenance costs are modest as compared with those for a centralized facility. For the off-site user, connecting to a centralized facility using high-speed data lines is at present too costly for most tasks. However, the LASP example illustrates a case in which timeliness in data reduction overrode cost considerations for such a link.

3. Most modern systems today (whether large or mini) are interactive. Therefore, this is not an exclusive advantage. However, the batch orientation of many of NASA's older computers has been a factor in users acquiring their own interactive minicomputer systems.

The *disadvantages* of decentralized computing facilities include the following:

1. Software developed tends to be specialized (customized) even for higher-level languages. Because of the inherent differences in hardware and architecture of various minicomputers, operating systems and assembly languages typically are quite different and lack standards. To some extent, these differences drive the development of higher-level language software such as Fortran or Pascal. This results in wasteful duplication of software.
2. System CPU and I/O throughput rate are usually significantly less for minicomputers than for mainframes. As a consequence, those tasks that are essentially CPU bound or involve large amounts of data are necessarily performed on large, centralized facilities.
3. Data-archiving practices are diverse (essentially different from one facility to another) and do not lend themselves to ready exchange or sharing of data with other computational facilities, and they also lack standards.
4. The range of available utility software (e.g., subroutines, plotting, statistical analysis algorithms, FFT) is more limited than on larger, centralized computers. Commonly, utility software must be developed or acquired at considerable cost in manpower and time.

The Future of Decentralized Facilities

The outlook for the next decade is that CPU, memory capacity, and I/O throughput rates for small computers will increase dramatically, while the costs will not. Thus, it seems highly probable that a decentralized system will continue to be used and perhaps be even more efficient. It also seems probable that many of the CPU-bound tasks that at present need mainframe-scale CPU's may be feasible to do on smaller systems in the future. For example, array processor costs are dropping rapidly, making them more affordable as an add-on to minicomputer CPU capabilities. However, data link costs will also be decreasing, perhaps making it feasible to have relatively high-speed data connections to centralized facilities, where large data sets have been reduced and archived, and to other decentralized facilities, when there is a need to obtain a portion of the data set for analysis. Thus, more of the decentralized facilities will become distributed computing systems once data networks

are established, while retaining all the inherent advantages to the scientific user that present decentralized systems afford.

III. CENTRALIZED COMPUTING FACILITIES

Examples

The trends with regard to centralized computing facilities are clearly away from batch-mode processing and much more toward interactive processing with a large variety of shared resources. Central computing systems are especially well suited for handling and accessing large data sets such as those for an archive of digital images, for problems requiring a large amount of CPU capability, and for providing specialized input/output, such as graphics devices.

We discuss several examples of centralized systems for space data processing:

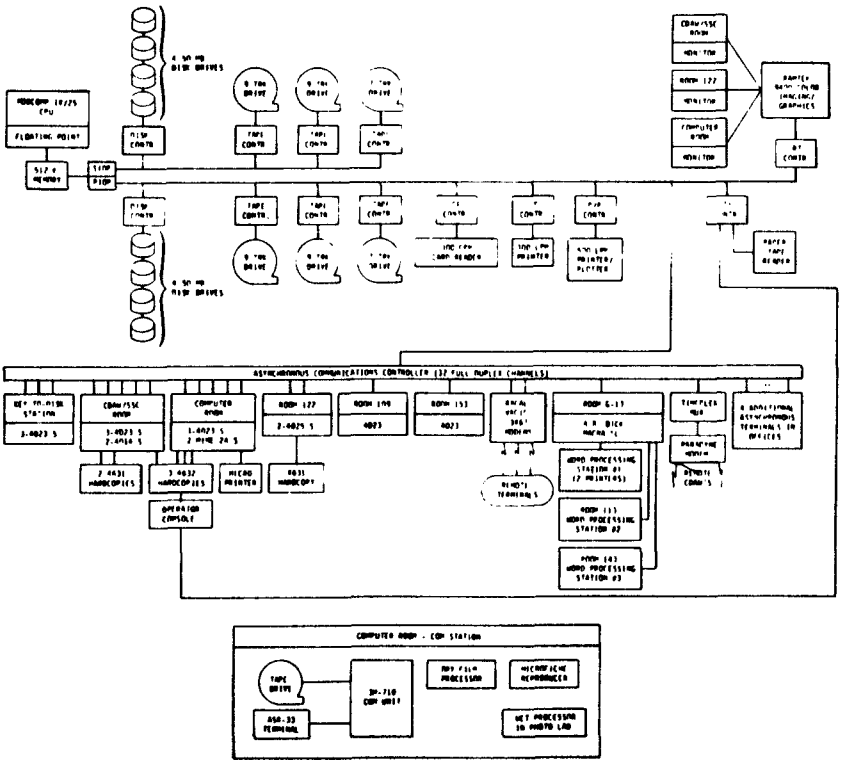

FIGURE 6.2 NSSDC computing center system configuration.

TABLE 6.4 NSSDC Computer System

Quantity	Item	Model
1	Computer with floating point hardware, I/O processor plane, and 512 kbytes of core memory	MODCOMP IV/25
1	300-cards per minute card reader	
1	300-lines per minute printer	
1	500-lines per minute matrix printer/plotter	
2	7-track 800/556 bpi tape drives	
4	9-track 1600/800 bpi tape drives	
8	50-Mbyte removable media disk drives	
1	Display, 1000 lines	Ramtek Model 9400
1	Asynchronous Communication with:	
2	Vector graphics terminals	Tektronix 4014
2	Vector graphics terminals	Tektronix 4025
11	CRT terminals	Tektronix 4023
2	CRT terminals	MIME 2A
1	Modem	Racal-Vadic 3467
1	Word-processing system	A.B. Dick Magna SL
Scheduled to be integrated		
3	CRT terminals	MIME 2A
1	Vector graphics option	Tektronix 4025
2	Modems	Racel-Vadic 3467
1	Statistical multiplexer	Timeplex
2	Printers	
3	Microcomputer work stations	

THE NSSDC SYSTEM

SYSTEM DESCRIPTION The NSSDC Computing Center's central processing unit is Modular Computer's MODCOMP IV/25 16-bit minicomputer, with a floating-point hardware plane, a secondary (I/O processor plane, and 512 kbytes of core memory (see Figure 6.2 and Table 6.4)

The NSSDC Computing Center also includes a stand-alone 3M-710 computer-originated microfilm (COM) unit, equipment for processing dry film, and a microfiche duplicating machine. Support for the COM unit is provided by the NSSDC photographic services section, and microfilm reader/printers are located throughout the building. Software exists for converting magnetic tapes into microfilm for a variety of graphics devices, and several different print tape formats can be processed. A unique factor of the NSSDC COM unit is the software for processing Tektronix 4010/4014 graphics data.

A tape staging area with a capacity of 1000 magnetic tapes is maintained in the computer room for current runs and frequently accessed user and system tapes. A tape library of 35,000 magnetic tapes is maintained in a climate-controlled environment in the basement.

An assortment of audio and video equipment (cameras, monitors, cassette recorders, video tape recorders) is maintained primarily for use in coordinated data-analysis workshops (CDAW's). Appropriate cabling and input/output jacks are installed in certain rooms around the building, and a sophisticated control panel exists to allow a large variety of hardware configurations (and rapid reconfiguration) in support of these computer-assisted workshops.

In addition to the hardware associated with the in-house NSSDC Computing Center, a number of terminals and serial printers are used for interfacing with the IBM 360/75 and 360/91 computers of the Science and Applications Computing Center (SACC) located at NASA/GSFC. A courier service is used for transporting cards, tapes, and printouts between the SACC and the NSSDC. Although the NSSDC Computing Center is the primary computing resource for the NSSDC, certain work, because of equipment limitations, must be done on the SACC. Examples of this are work for which the use of the 6250-bpi tape drives is required; it is required that the input (or output) tapes remain in the SACC tape library; it is not cost effective to duplicate processing software that already exists on the SACC computers; and processing requires more memory and/or CPU resources than are available on the MOD-COMP IV computer.

NSSDC COMPUTING CENTER–OPERATING SYSTEM AND SUPPORT SOFTWARE

The vendor-supplied operating system was designed for a process-control environment. This system, used for NSSDC, was found to result in about 70 percent CPU idle time. To stop this waste, NSSDC has significantly modified the operating system to produce a custom NSSDC operating system that maximizes throughput and resource utilization, while minimizing response time for interactive users. The NSSDC supports a high volume of magnetic-tape duplication, verification, dumping, and reformatting activity; a high-usage information storage, retrieval, and report-generation system; CPU-intensive magnetic-field line tracing and other programs in support of the Satellite Situation Center; scientific data-reduction and synthesis programs; coordinated data-analysis workshops; documentation and other word-processing applications; interactive graphics; key-to-disk data entry terminals; and extensive ongoing software development to implement and support these activities. In short, the NSSDC requires a comprehensive general-purpose computing system that, at the same time, is highly responsive to interactive users, rather than the somewhat limited-purpose applications that are more typical of minicomputer installations in the MODCOMP IV/25 class.

The NSSDC operating system is essentially an I/O driven time-sharing system. Currently the system is set up to handle up to 13 general time-sharing users, along with the key-to-disk entry system (with an arbitrary number of entry terminals) and a background CPU-intensive task. Many of the operator console functions have been made available to the user, so each user essentially has available to him a 128-kbyte (conventional overlaying is required for larger programs) virtual machine, and operator intervention is normally only required for tape-drive allocation and the handling of magnetic tapes. Upon I/O completion, a task is scheduled for an immediate small (adjustable) time-slice of CPU; so, for truly interactive tasks (i.e., text editing), the task will proceed as if it were alone in the system, regardless of any intensive CPU requirements of other tasks in the system, with negligible degradation of any CPU-intensive tasks.

Much system support software to facilitate data set duplication and verification, data management, data set backups, graphics support, and other utilities have been added to the MODCOMP system. Other special software, such as that to support communications between the MODCOMP IV and the A.B. Dick Magna SL word processor and to prepare tapes for the 3M 710 COM unit have also been added.

IMAGE PROCESSING LABORATORY

The Image Processing Laboratory (IPL) at JPL consists of an IBM 360/67 and peripherals that are dedicated to image-processing tasks. Tasks are run, processed results can be examined on either TV's or on Polaroid film, and then tapes can be generated with labeled image data. Those tapes can then be brought to one of a number of film recorders at JPL for production of hard copies. Much of the second-order processing of image data acquired during planetary missions is carried out at the IPL. IPL supports VICAR, a set of software that has been highly modified during more than a decade of use at the IPL. VICAR supports such tasks as contrast enhancements, filtering, and geometric rectifications. Although some interactive capabilities exist, including CRT's and TV monitors, most of the processing is done in batch mode. Users are commonly assigned a block of time and, if needed, an analyst to help process their tasks. Charges were approximately $250/hour during 1976-1977, when processing Viking data at IPL.

Almost all of the special planetary image products that have appeared in the scientific and professional literature have been processed at IPL. As such, the Laboratory must be labeled as a success. However, a number of difficulties have been encountered by scientists who have tried to use the facility, especially those who are not directly associated with JPL and who are not on-site long enough to assure that their request has been processed. The main

problem has been relying on IPL analysts to interpret a user's written instructions properly and then to process the job. Significant delays to the scientific users are inherent in this approach.

IBM SCIENTIFIC CENTER IMAGE PROCESSING FACILITY

The IBM Palo Alto Scientific Center supports 60 interdisciplinary scientists, whose interests include meteorology, seismic research, artificial intelligence, robotics, remote sensing, atmospheric physics, image-processing science, image-graphics computer science, and computer language development. An IBM 370/158 computer with 2 Mbytes of memory is used to support the data processing (see Table 6.5). Interactive graphic terminals and color displays are used to provide two-dimensional graphic and image results. Image and graphics hard-copy products can be produced on-line. The operating system used is the Conversational Monitor System (CMS). The software most commonly used by the scientific staff includes Fortran, APL, PL-1, Pascal, and assembly language.

All programming, compiling, debugging, and execution is done interactively; each of the 60 users has a terminal of his or her own or one nearby and defines a "machine" with the size memory (500 kbytes to several million bytes of "virtual" memory) and peripherals needed for his or her particular application. The resources are then allocated to the user on a shared basis. Productivity is high because of the interactive nature of the processing envi-

TABLE 6.5 IBM Palo Alto Scientific Center Computer Facility

Quantity	Item	Model
1	IBM 370/158 with 2-Mbyte memory	IBM U 34
12	Storage disks	IBM 3330
1	Printer	IBM 3211
1	Laser printer	IBM 6670
2	Card reader punches	IBM 3505 and 3525
7	Displays (graphics)	Tektronix
60	Interactive terminals	IBM 3277 and 3279
1	Information distributor	IBM 6670
1	Color display (525 x 640)	Ramtek 9300
1	Color display (1000 x 1000)	IBM
1	Black and white display	CONRAC
1	Drum scanner/plotter	IBM
1	Film scanner	Optronics
5	Graphics attachments	Tektronix
1	Color processor	Kreolite
3	9-track 6250-bpi tape drives	
3	Minicomputers	IBM Series 1

ronment and the availability of many different languages, routines, programs, and peripherals. The scientific staff perform all of their own application programming. A computer support staff is available to help in the event of system problems and to develop new support software as needed.

The throughput performance of the system is variable. In the normal environment, each user effectively has the full resources of a midsize computer. This includes a comprehensive set of software programs and subroutines. Response time is generally excellent. During maximum-use periods, brief delays in program computation or execution can be noticed but generally do not create problems. In the case of image processing, when a large amount of data is to be processed, the activity is sometimes given higher priority or scheduled for nonpeak use times in order to maximize productivity. The system is running 24 h per day and 7 days per week, with a first-shift operator in attendance only 5 days per week. Extensive word processing is done on the system. Secretaries and scientists prepare manuscipts, technical papers, and reports interactively and use a laser printer to produce high-quality hard-copy products.

Although the system is primarily a centralized processing system, a user can access a geographically remote computer system or data base. It then becomes a distributed system for that user or application. The system is connected to other computers at various IBM Research and Scientific Centers, and messages, programs, and even image data are routinely sent computer-to-computer via satellite or ground communications links. There has been no problem with data degradation from the communications channels, primarily because of error checking routines and the use of reliable communications links.

The use of a shared midsize computer in an interactive, virtual machine environment has provided extensive scientific computational facilities to many scientists in an economical form, without the need for the user having to maintain or operate the system or its peripherals.

Pros and Cons of Centralized Facilities

The following summarizes the major advantages and disadvantages of centralized processing facilities.

The *advantages* of a centralized facility include the following:

1. Each user has the advantage of having a large, high-performance computer with a variety of peripheral devices. Thus, the user is generally not limited by computational resources or I/O devices.
2. Generally, a comprehensive library of programs and utility routines are developed, maintained, and documented in standard formats and protocols.

Thus, each user can benefit from software development that has previously been implemented, reducing the incidence of wasteful duplication of effort. It is common for users to help each other with programming or system problems and to share software.

3. Centralized processing facilities generally provide data-management, storage, and retrieval capabilities, reducing the need for the user to physically locate and handle the data and maintaining data integrity.

4. System maintenance and operation and the development of utility programs are not the responsibility of the user, thus allowing the scientist to do more science.

5. System personnel are generally available to help with hardware, software, or system problems.

6. For some applications, a midsize or large centralized facility used by many scientists in an interactive mode is a more economical approach than a small dedicated facility. The total system and operating costs, divided by the number of users, is frequently less than the cost of a minicomputer and its peripherals.

The *disadvantages* of a centralized facility include the following:

1. The user does not have complete control of the facility. He cannot easily or quickly change the configuration (beyond the usual virtual machine configuration) for his application but must go through a bureaucratic procedure to change computational and peripheral hardware or system software.

2. When one user or group of users requires and uses extensive computational or peripheral resources, this may temporarily hinder the individual user, reducing his productivity.

3. The system is usually designed to meet the requirements of all the users and therefore not optimized for a particular application.

The Future of Centralized Systems

With the rapid decrease in costs and the increase in performance of computers, the situation is rapidly occurring that each scientist or organization can afford and will have his or her own computer, structured or configured to support particular requirements. When there are increases in computational complexity, it is likely that individuals or groups of scientists may combine their resources in order to obtain and share a more powerful computer with more software support and peripherals.

It is difficult to predict what the future will hold, except that both directions will be followed, depending on computational requirements and, in some cases, the personality of the user.

IV. DISTRIBUTED COMPUTING FACILITIES

Examples

The trends toward increased computing and storage capability of minicomputer systems at low cost, coupled with improved data-transmission capability, make it feasible and desirable to interconnect computing facilities of all types.

LABORATORY FOR APPLICATIONS OF REMOTE SENSING (LARS)

The Purdue University Laboratory for Applications of Remote Sensing (LARS) was one of the first organizations to establish a computer system for processing Earth-observation data. LARS also specified and funded the development of the first interactive digital image display for processing science and applications data. This sytem has been and is currently being used for processing remotely sensed aircraft and spacecraft data by investigators throughout the United States. The LARS system is an example of a medium-sized, general-purpose computer facility with distributed processing and networked features.

Research into machine data-analysis methods for earth-observational systems was begun in the middle 1960's. Broadly stated, the research problem was to learn how to use the vantage point of space for the creation of up-to-date, accurate, and less expensive information relating to the Earth's resources and to its environment. At the outset, it was recognized that in such a system there would be large quantities of data that would need to be processed. In 1966 a stand-alone research image and data-processing system was designed and implemented to support the early research. The large central campus computation facility was not used because it had objectives that did not permit it to meet the specialized requirements of such research.

The early research efforts were ultimately successful and led to the requirement for a processing system with the same basic characteristics as the stand-alone system but with much greater capacity to carry the work forward. Thus, in 1970 a new, larger processing capability was designed to serve researchers locally at Purdue as well as at other institutions.

Based on the lessons learned both in attempting to use the central campus facility and the small dedicated in-house capability, some fundamental principles for its design were defined. Specifically, it was intended to create a computing environment tailored to research and with the following characteristics (stated in their approximate order of priority):

1. Very high researcher control and flexibility.

2. Economical accessibility of large blocks of computer time as needed by the researcher.

3. Access to sufficient memory to avoid frequent requirements for software overlays.

4. A computer schedule designed for research jobs.

5. Interactive programming debugging and data-processing capability.

6. Self-directing training materials designed to introduce researchers to new system capabilities at their convenience and with minimum effort on their part.

7. A strong consulting support.

8. A programmer pool.

9. Appropriate software package support.

A system meeting these requirements was implemented and has now been in operation for approximately 10 years. The key to achieving these objectives was more in the software system than in the hardware. As a result, a key choice in the system design and implementation was the selection of a resource-sharing type of operating system that allows the user to load and execute programs conveniently with the system resources needed for his particular job. For example, the current system is capable of serving about 50 users at any given time, each one of whom may be utilizing 500,000 bytes of main memory. It is not uncommon for individual users to execute programs requiring two or more megabytes of main memory within the normal job mix.

Available under the system operating software referred to above are a number of user software packages that have evolved out of the efforts of various researchers over the years. Several of these packages are large (greater than 10,000 lines of code). Their purpose is to provide researchers with a convenient shared software base, giving them many of the routine and utility capabilities that would normally have to be developed independently. Also supplied are a number of more general-purpose software packages. Examples are the Statistical Analysis System, the Statistical Package for the Social Sciences, the VSMA Graphics Compatibility System, and the Interactive Graphics Library.

The system is operated as a self-supporting activity, meaning that if its services are not in demand, its resources decline. This provides its staff with a strong motivation to be helpful and responsive to research users. New capabilities are added as justified by user need, and old capabilities are eliminated when there is no longer a call for them by the user community.

There are currently 66 hard-wired terminal ports into the system at Purdue from around the eastern and southern United States. Telephone dial-in ports are also available at Purdue and at a remote site at the NASA Johnson Space Center in Houston, Texas. Using these terminals, researchers are able

to carry out a broad range of processing tasks over a considerable geographic area from the writing of small Fortran programs to the multispectral analysis of Landsat image data.

The system hardware currently consists of an IBM 3031 with 2 Mbytes of main memory; 2 Gbytes of direct-access storage; 10 tape drives; appropriate line printers, card readers, and punches; and a communications controller for interfacing the computer with the terminals.

The 66 terminals vary in size and complexity, from a simple CRT keyboard terminal to terminals containing line printers, tape units, and the like. The terminal with the most complexity is one in which a PDP 11/34 serves as an intelligent terminal to drive a table digitizer, electrostatic printer/plotter, and color and graphic display systems.

A second approach to accomplishing the same objectives also exists on the Purdue campus. In this case, a distributed computer network has been established with a VAX 11/780, a PDP 11/70, and a PDP 11/45 as the central machines. High-speed links are established to six other PDP 11 machines and an additional VAX 11/780. The system operates under a UNIX software system in which each researcher may process his job on his or her own local machine or, if the user desires, the system will automatically search out the machine with the lowest current-usage level and execute the program on that machine.

PENNSYLVANIA STATE UNIVERSITY METEOROLOGY LABORATORY

Pennsylvania State University's Meteorology Laboratory has in current operation a Digital Equipment Corporation (DEC) PDP 11/34 minicomputer with a 28-Mbyte RK07 disk and three 2.5-Mbyte RK05 disks, a 9-track tape drive, line printer, two Decwriter terminals, and three Tektronix 4010 or 4012 storage-cube CRT display terminals together with a Tektronix 4610 hard copier. The laboratory also has a DEC PDP 11/10 minicomputer with one RK05 2.5-Mbyte disk and a Harris Digidata tape drive that can control a WSR-74 C-band radar and store, analyze, and display digital output from it. The 11/10 handles digital satellite imagery received in analog form via a C-5 conditioned telephone line from Washington, D.C., and re-digitized on site. Video display functions of the Model 10 are carried out through a Grinnell intelligent terminal system.

Data are received from several sources. The FAA's 1200-baud 604 circuit provides conventional meteorological data with excellent coverage in real time. Software has been developed to analyze and display many areas at the surface and aloft, on maps, cross sections, and other plots, based on these data. In addition, national facsimile (NAFAX) service is available, as is a direct line to the National Environmental Satellite Service in Washington, D.C., for real-time satellite data. Programs have been written to display

and enhance, and in a limited way, animate, the satellite images and radar data in digital form. All the systems can operate 24 h per day.

Pros and Cons of Distributed Computing Facilities

It is apparent that more and more systems are directly or indirectly becoming distributed. The advantages and disadvantages are as follows:

The *advantages* of a distributed system are:

1. Data, software, and knowledge resident in other systems or geographic locations can be accessed and used.
2. Computational resources, such as large-scale computers, can be accessed from another computer when needed, thus allowing the user to select and purchase a machine structured for his average rather than his maximum computational workload.
3. Scientific dialogue is improved. Technical material, papers, references, data, and even messages can readily and rapidly be transferred between users.

The *disadvantages* of a distributed system include the following:

1. Communications terminals and large bandwidth are required, which add to system costs.
2. Unauthorized use of facilities or data can occur if proper safeguards are not used.
3. There is some loss of system control when multiple users can access systems that are geographically separate.

Future of Distributed Computing Facilities

With the anticipated decrease in costs of satellite and ground data communications in the 1980's and the concurrent requirement for and implementation of larger data bases, combined with the development of high-performance and special-purpose processors, it is likely that there will be an explosive growth in the interconnection of many existing and planned computers. Communications and computer protocols are being developed that will allow a user at a particular geographical location to access data or programs at another location in a transparent manner, as if this capability were inherent in his system. This approach is the most progressive and cost-effective approach but will not happen efficiently without planning and resource commitment.

7
Principles for Successful Science Data Management

All the world over and at all times there have been practical men, absorbed in irreducible and stubborn facts; all the world over and at all times there have been men of philosophic temperament, who have been absorbed in the weaving of general principles.

ALFRED NORTH WHITEHEAD, *Science and the Modern World* (1925)

Based on the experiences of the past two decades of space activities, a number of principles have been identified that are fundamental to the successful management of scientific data. These are stated here in some detail because they form the basis on which the appropriate Data Management Units for individual space research projects, programs, or disciplines should be constructed. We have seen examples in Chapter 3 of how scientific involvement in the data chain (from mission planning to distribution and archiving) has affected the quality and timeliness of the scientific return from space missions. Table 7.1 summarizes our evaluations of the match between the examples (Chapter 3) and the principles. The principles are as follows:

I. SCIENTIFIC INVOLVEMENT

The object of science and applications research missions is to obtain new scientific knowledge and understanding. It seems reasonable, therefore, that since the data acquired are for scientists, the scientists should have a substantial

TABLE 7.1 Match between Activities and Scientific Principles[a,b]

	P.I. Units					Project Units				Units of Broader Scope				
Principle	Uhuru	ISEE	Landsat	Seasat	GOES	Viking	HEAO-2 Einstein	Atmosphere Explorer	Space Telescope	CDAWS	UCLA Group	LUNAR Consortium	Planetary Image Facility	LARS
Science involvement	G	G	P	P	P	G	G	G	N/A	G	G	G	G	G
Peer review	P	P	P	P	P	M	G	P	N/A	G	M	G	G	G
Data quality availability	M	P	M	P	G	M	M	G	N/A	P	G	G	G	G
Computational capabilities	P	M	M	P	M	M	G	M	N/A	P	G	G	G	G
Software practices	P	P	P	M	P	M	M	M	N/A	P	G	G	G	G
Archival services	M	M	M	M	G	M	G	G	N/A	P	G	G	G	G
Adequate finances	M	M	M	P	P	M	M	M	N/A	G	M	M	M	G

[a]G, good; M, mixed; P, poor; N/A, not applicable.

[b]Since the Space Telescope has not flown and the Science Institute has not been established as of the time of this writing, the program cannot be evaluated.

role in determining what data are required, how they are acquired, and the ultimate disposition of these data after they are acquired.

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

II. SCIENTIFIC OVERSIGHT

Usually the management of scientific data has been the responsibility of project offices, principal investigators, and eventually the archives. Missing in this process are those individuals outside the project who actually use the data for research. CODMAC found that the most successful cases of data management involved user oversight, especially over the data-storage and -archiving activities. In particular, the purging of archives, rescuing of data sets, organizing of related data bases, and providing for long-term storage for data should be processes that are conducted according to the advice of the users of those data.

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

III. DATA AVAILABILITY

Since the cost of acquiring data by spaceborne instruments is so high, we would expect that the process of making them available to the research community would be a prime goal of every mission. Unfortunately, we find that this is not the case. We have frequently found that:

- The extraction of physical units from reduced data is complicated by the lack of ancillary data.
- Comparison of diverse data sets is a problem because of different formats.
- There are long delays in making data available to users.
- Documentation of archived data sets is incomplete or entirely absent.
- Users are unaware of what data are available or are unable to obtain accurate catalogs of the data.

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

1. *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.*

2. *Appropriate ancillary data should be supplied, as needed, with the primary data.*

3. *Data should be processed and distributed to users in a timely fashion as required by the user community.*

4. *Contractual obligations by users to return data to the archive in a modified form should be enforced.*

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

IV. FACILITIES

In order for a scientist to make use of data once they are obtained, he or she must have access to certain facilities to process the data, store them, and analyze them. All too often scientists must use equipment that is slow, overburdened with other users, or with inadequate capacity. On the other hand, sometimes equipment is made available that far exceeds the scientist's requirements and thereby makes the job too costly. The productivity of the scientist should be the prime goal, but attention must also be paid to the costs of facilities so that the research will not be curtailed by overinvestment in hardware.

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

V. SOFTWARE

We have previously recounted many of the problems that scientists face with computer software (see Chapters 1 and 3). Because of the high costs of software, these problems will cause a reduction in productivity since resources are being wasted mainly through unnecessary duplication and costly maintenance. We further expect that the problems will become even worse in the era of severely constrained financial resources that we are now experiencing.

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

VI. SCIENCE DATA STORAGE

Data-archiving facilities are a primary resource of users of science data. Maximum utilization of the available science data requires such services as quick-look capabilities, data catalogs, and timely distribution of the data. Of course, a genuine interest in performing well the role as a service organization will eliminate practically all nontechnical problems associated with data distribution.

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.

VII. DATA-SYSTEM FUNDING

Since the object of space and applications research missions is to obtain new scientific knowledge and understanding, it is imperative that funds be available to use the resulting space-acquired data in order to obtain that knowledge and understanding. However, CODMAC has identified far too many examples of where inadequate planning time, delays, and cost overruns have significantly reduced the funds available for extraction of information from the sensor data. Science is not served if the financial resources for a mission are exhausted before the data can be analyzed.

Adequate financial resources should be set aside early in each space 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.

8
Types of Scientific Data-Management Units

The open society, the unrestricted access to knowledge, the unplanned and uninhibited association of men for its furtherance—these are what may make a vast, complex, ever growing, ever changing, ever more specialized and expert technological world, nevertheless a world of human community.

J. ROBERT OPPENHEIMER, *Science and Common Understanding* (1953)

I. INTRODUCTION

The principles of science data management that are discussed in Chapter 7 were derived, in part, from an analysis of examples of present data-management methods, ranging from Principal-Investigator-oriented, through project-oriented, to systems that involve management and processing of data from a number of missions or sources (units of broader scope). Most likely, all three types of data-management units will continue to exist in the future. In this chapter we summarize the problems encountered in the past for each type, and we then show how the application of the principles would have led to a much greater science return. Our discussion does not include the major archival facilities, although these facilities would also benefit from application of the principles.

We end this chapter with a discussion of the Discipline Data Management Unit. We see this type of unit as a natural evolutionary goal. It would serve as

a focus for collecting, processing, archiving, and distributing data pertinent to a particular scientific community or for a specific scientific program.

II. PRINCIPAL-INVESTIGATOR DATA-MANAGEMENT ACTIVITIES

Many NASA projects operate with the Principal-Investigator (PI) format, whereby a scientist is selected to provide an instrument or to oversee the development of an instrument to be flown on a spacecraft and to process and reduce the data. The data are usually provided directly to the PI, who then processes them and forwards raw or reduced data to an appropriate data center facility. As discussed, the success of this type of data-management unit has been variable. The best way to illustrate the problems is to summarize in a general way the extent to which each of the principles has been followed:

1. *End-to-End Science Involvement* The PI, usually a scientist, has generally been involved from the beginning through the end of a mission. At times, however, problems arise because too little thought or resources have been given to data management in the planning of the mission.

2. *Oversight Mechanisms* Peer review of the data-management tasks has largely been nonexistent. Peer pressure, to some extent, has been applied. Little thought has been given to archival storage of data.

3. *Production of Usable Data* The main task for the PI has been to reduce the data for his own needs. As a consequence, the data are sometimes not useful to, or not interpretable by, the rest of the scientific community. There have sometimes been long delays between the PI's receipt of the data and the deposition of the data into an archive. The merging or comparison of different data sets (from different experiments) has suffered.

4. *Computational Capabilities* This requirement has been met to some extent, although lack of early planning for data-reduction activities, combined with limited resources, has led in some cases to a poor match and the inability of the PI to process all the data needed.

5. *Software Considerations* The extent to which structured, transportable, documented software has been developed is hard to measure, although usually the software has been developed within the PI's own research group, where little incentive or few resources exist to make software transportable.

6. *Archival Services* The degree to which the PI has felt an obligation to help the user community has varied largely with the personality and philosophy of the PI.

7. *Adequate Financial Resources* Lack of early planning to establish adequate resources for data management and computation has often plagued the PI data-management unit.

III. THE PROJECT DATA-MANAGEMENT UNIT

In more complex missions, teams of investigators are often selected and a number of instruments are flown. In most cases, the project has assumed responsibility for processing and managing the returned data within a common data base. Either raw or processed data sets are then distributed to the teams of investigators. In some of the applications missions, even the teams are nonexistent in early phases of the mission, and the processing, storing, and sometimes the distribution of data are the concern of the project. The success of past Project Data Management Units can be measured against how well the principles were followed:

1. *End-to-End Science Involvement* The degree of science involvement has been variable, ranging from incorporation of scientists in the preplanning through the data-reduction phases to having no scientific involvement at all.

2. *Oversight Mechanisms* In some cases, no oversight mechanism has existed. In other cases, a science steering group existed whose charter has included oversight of data-management tasks. Even in the latter case, too little attention has been given to data-management activities.

3. *Production of Usable Data* For science missions, the Project Data Management Unit has usually led to a greater degree of control of data production than in the PI case. In applications missions, the lack of science involvement has led to severe problems in production of data useful to the scientific community. Problems still exist within Project Units in terms of producing interpretable, documented data with usable catalogs. Again, the incentive for the teams has been largely to reduce, interpret, and publish results for their own needs. Even with science steering groups there has sometimes been a lack of appreciation of the utility of comparing diverse data sets.

4. *Computational Capabilities* There frequently has been less than adequate funding for proper scientific computation capabilities. In addition, computational facilities for processing the data at NASA centers have usually been 5 to 10 years behind in computing technology.

5. *Software Considerations* Numerous examples exist of both poor and good practices with regard to software development. In some cases software written for a given mission has not been documented or used after the mission terminated. In other cases, older software has been inherited and implemented in new missions.

6. *Archival Services* Usually, investigators involved with the project, either directly or indirectly, have been fairly well served by the project data facility. Usually, outside investigators are not serviced (or not well served) by Project Data Management Units.

7. *Adequate Financial Resources* Even with early science involvement, major problems exist with planning and protecting financial resources, mainly

because of cost overruns related to hardware development and because of a lack of early thought in regard to needed financial resources.

IV. DATA-MANAGEMENT UNITS OF BROADER SCOPE

There are a number of examples of Data Management Units that either transcend or are broader than PI and Project Data Management Units. In some cases, Project Data Management Units have grown to a broader scale to include data from other sources and to provide a focus for consortia activities. The Atmospheric Explorer and Einstein Observatory missions are good examples of such cases. In general, units of broader scale consider diverse data sets, have continuing scientific involvement, conform more to the principles for science data management, and thereby maximize the scientific return for the investment. An analysis of how the principles are adhered to follows:

1. *End-to-End Science Involvement* Units of broader scope are usually initiated and implemented by scientists. As such, these units commonly have the highest and most vigorous degree of science involvement in all phases of implementation.

2. *Oversight Mechanisms* There are a variety of structures, both formal and informal, existing within these units.

3. *Production of Usable Data* This task has been the main goal of many of these units. Examples of groups that have this goal include the Lunar Consortium and the UCLA magnetospherics group. Data produced or handled by this category of management unit have been documented to varying degrees, depending on the extent to which the data have been produced for consumption by the user community. In some cases, advanced capabilities such as an electronic browse capability to interrogate the data sets exist. In general, there is a greater degree of concern for timely processing and availability of data; funding limitations have been the primary limitation to data management.

4. *Computational Capabilities* The match between computational requirements and capabilities has been variable. Usually, the minimum capability needed to complete the task is provided because of funding limitations. Systems have been developed that closely match the needs of the units, although existing computational resources are frequently used because of their availability.

5. *Software Considerations* The degree to which these units have taken the lead in proper software development techniques is usually a step above either the PI or the project units. This includes both software for scientific processing and for data management. The skills and needs of multiple users

have resulted in the development of more capable systems. There is more concern with software transportability and documentation.

6. *Archival Services* The service orientation is reasonably well handled by this type of management unit, if only because of informal contact between the scientists involved and the user community. Funding limitations frequently limit the degree of service to a user community, since many of these units are supported on yearly renewable grants and contracts and do not have a funded service obligation.

7. *Adequate Financial Resources* Problems often exist in this area because of insufficient early planning, and in some instances, lack of a long-range financial commitment.

V. THE DISCIPLINE SCIENTIFIC DATA-MANAGEMENT UNIT

It is clear that many of the problems associated with maximizing the science return per investment cost for the PI, the project, and the broader scope units could have been alleviated by application of the principles described in Chapter 7. Science involvement in all phases of the data chain may be the most crucial element. This statement is supported by the observation that the data-management units of broader scope, which have largely been initiated and run by scientists, have usually produced data of higher quality and reliability and of more general use to the user community (see Table 7.1). As noted, however, none of the units of broader scope have really met all the principles.

Although we recognize the need of retaining the PI and the Project Data Management Units, we also recognize that, in many cases, the broader scope management units seem to provide greater science return. We can envisage a data-management unit that is largely based on examples of Units of Broader Scope—the Discipline Data Management Unit (DDMU).

These units would evolve as the need for them arises. For example, a unit of broader scope might start collecting additional data sets because of interest in particular scientific problems and eventually find that its collection has become the primary data and information source for an entire scientific community. Such a unit could even provide the major computational resources and software programs for its discipline. Perhaps the Space Telescope Institute, which will have charge of planning, reducing, archiving, and distributing Space Telescope science data, is close to our concept of a Discipline Unit, though only Space Telescope data and not all astronomical data would be included in the data base. We could also envisage the creation of a Geodynamics Data Management Unit as another example of a DDMU. Such a Unit might house geophysical information (data on crustal movements or potential field

data, for example) and geologic information (geologic map data or digital topography, for example) pertinent to tracking and understanding crustal movements and information patterns.

Although each scientific discipline must formulate its own specific requirements for a DDMU, there are several generic requirements that should be applicable:

1. The DDMU employs archive scientists as well as technical and administrative support staff.
2. The DDMU provides the interface between investigators in a scientific discipline and the NASA data-collection process.
3. The DDMU archives and distributes data relevant to its discipline.
4. The DDMU provides scientific oversight for data processing when required.
5. The DDMU exercises scientific oversight of any data purging and is prepared, if necessary, to "rescue" selected data for scientific use.
6. The DDMU provides coordination among disciplines in support of interdisciplinary analysis efforts.
7. The DDMU develops software for general use in a discipline and coordinates software development with other data-management units and with the user community.
8. The DDMU assumes a leadership role in developing software and data standards and formats.
9. The DDMU provides computation facilities for investigators who desire to use the data in the DDMU archives but who do not possess their own dedicated computational facilities.
10. The DDMU advises the user community of the availability of relevant data and assists the user community in obtaining access to these data.
11. In some cases, the DDMU operates the principal archives for its discipline.
12. In some cases, the DDMU creates specialized data bases for general disciplinary use.
13. In some cases, the DDMU creates and maintains models representing the data in its discipline.
14. The DDMU is responsive to advice from a committee of active investigators in its discipline—a users committee.
15. The activities of the DDMU are periodically reviewed by an oversight committee.
16. The DDMU, its users committee, and its advisory committee are jointly responsible for advising NASA on the scientific requirements for data collection, processing, archival, distribution, and analysis.

In addition to the specific technical facets of a DDMU operation, there are several other requirements for a DDMU to operate successfully.

Since a DDMU for handling space data will likely be funded largely by NASA, there should be a formal overview of the unit by the appropriate NASA office. A formal statement of work should be negotiated, which includes regular reporting. In addition, a proposal review committee should be established to determine which scientific proposals are to be supported and to resolve potential duplications or conflicts. The proposal review committee should include NASA representatives (other government agency representatives, if they are involved), DDMU scientific representatives, and science discipline representatives; it might be more broadly constituted to include an overall advisory function concerning general objectives and performance of the DDMU. In the case of larger units, the advisory committee should probably be separate from the proposal review committee.

Still another committee, namely a users committee, is required. This committee should meet frequently to review and advise the DDMU on detailed matters involved in the day-to-day activity of the unit. As the name suggests, the users committee should be composed of individuals actively involved in the use of the data and services provided through the DDMU.

In addition to the above oversight activities, the DDMU will require adequate funding, a long-term commitment, an appropriate technical support staff, and a high-quality scientific staff if it is to be successful. The funding inadequacies for the data phase of many of the programs described above cannot be dissolved by simply creating a DDMU. The DDMU provides a mechanism for optimizing planning and performance to maximize scientific yield, but it clearly requires adequate funding. At the same time, the software development, data distribution, and data-archiving functions of the DDMU all imply a long-term commitment. If the scientists in a given discipline are convinced that a DDMU is required and the appropriate review committees and NASA offices agree, then a long-term commitment is required to establish the unit and to attract a qualified staff. The DDMU should not necessarily be limited to a single program but, in fact, should be involved in all relevant programs from the earliest planning stages. This again implies continuity and longevity.

Although it is necessary that a DDMU be provided with sufficient continuity to perform its functions, we do not envisage that a DDMU is a "permanent" commitment. It is essential that there be established procedures for terminating a DDMU as well as for establishing one. In many cases the function of a DDMU will be for a relatively short time period; in other cases a DDMU may fail to perform satisfactorily or may lose the key personnel necessary for its function. We recommend that a DDMU be established for a prescribed period of between three and five years. At the end of this period a formal re-

view of the unit should be carried out to determine whether it should be continued, discontinued, or revised.

Although the DDMU concept is expected to enhance significantly the scientific utilization of space-derived data, it must also be recognized that there are several limitations.

DDMU's should not be expected to replace the major data centers. The success of the unit depends on having a manageable group of scientists with an active interest in the data being handled. An operation of the size necessary to perform the functions of the data centers could become so large that the scientific objectives would be lost.

On the other hand, it would be inappropriate to establish a DDMU dedicated to a single, short-term mission. Such a unit would not be of sufficient size and longevity to assemble the personnel required for a successful operation.

Finally, it must be remembered that the requirements of each discipline are unique. Although the generic requirements for the DDMU have been outlined above, each discipline must take the responsibility of tailoring the concept to its specific requirements and determining those areas within the discipline for which it is appropriate.

Appendix: Present and Expected Data Volumes

In attempting to estimate the magnitude of the data problem at present and for the future, data have been obtained from NASA providing information on projected annual data rates from existing and planned space missions. The information is organized in Tables A.1-A.3 according to whether the missions are Applications, Geocentric, or Deep-Space. The tables are not exhaustive, particularly concerning planned missions for which the expected data rates will depend strongly on the instrumentation ultimately selected and flown. It should be noted that because of recent budgetary decisions, some programs may be delayed or canceled.

We also present a graph (Figure A.1) showing the trend for the quantity of data expected to be collected in the future.

In order to assess the present volume of data, we have included a brief description of the NSSDC and EDIS data bases (see Tables A.4-A.7).

Some historical Landsat statistics are provided, showing the growth of the distribution of Landsat data between 1973 and 1980. Of particular interest is the dramatic increase in the number of digital-image scenes provided to the user community beginning with about 10 scenes per year in 1973, rising to about 3000 when the digital capability became available in 1978, and currently exceeding 4000 scenes per year.

Finally, at the end of this Appendix, we summarize our conclusions about the amount of space-acquired data that we must plan to deal with.

TABLE A.1 Applications Missions

GOES	1.5×10^{13} bits per year
HCMM	70,000 scenes per year
Landsat-3	8.5×10^{13} bits per year
Magsat	6.0×10^{10} bits per year
Nimbus-7	2.9×10^{12} bits per year
SAGE	1.6×10^{10} bits per year
Tiros-N	2.0×10^{13} bits per year
ERBE	2.5×10^{11} bits per year
ICEX	3.6×10^{13} bits per year
Landsat-D	9.5×10^{13} bits per year
OERS	2.0×10^{12} bits per year
Stereosat	1.5×10^{14} bits per year
UARS	2.0×10^{12} bits per year

TABLE A.2 Annual Data Rates of OSS Geocentric Missions in Kilobits per Second

Mission	Launch Date	Potential	Avg Bit Rate (kbps)	Calendar Year 79	80	81	82	83	84	85	86	87	88
AE-5	11/20/75	9/30/81	2.4	2.4	2.4	1.8							
DE-A	7/31/81	9/30/84	8.0			3.3	8.0	8.0	6.0				
DE-B	7/31/81	7/31/83	4.8			2.0	4.8	3.6					
CCE	10/ 1/83	4/30/85	1.8					0.5	1.8	0.6			
COBE	10/ 1/85	9/30/86	1.0							0.2	0.8		
EUVE	10/ 1/85	9/30/86	3.0							0.7	2.3		
GRO	8/ 1/85	9/31/87	20.0							8.3	20.0	15.0	
HEAO-2	11/13/78	5/31/82	6.4	6.4	6.4	6.4	2.7						
HEAO-3	9/20/79	5/19/81	6.4	1.9	6.4	2.9							
IMP-8	10/26/73	9/30/80	1.0	1.0	0.8								
IRAS	8/27/81	8/26/82	1.3			0.4	0.9						
ISEE-1	10/22/77	9/30/84	6.5	6.5	6.5	6.5	6.5	6.5	4.9				
ISEE-2	10/22/77	9/30/84	3.3	3.3	3.3	3.3	3.3	3.3	2.5				
ISEE-3	8/12/78	9/30/93	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0

IUE	1/26/78	9/30/85	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.3			
OAO-3	8/21/72	11/30/80	0.5	0.5	0.5								
OPEN EML	1/ 1/86	12/31/89	13.9→6.4								13.9	10.2	6.4
OPEN GTL	1/ 1/85	12/31/89	6.4							6.4	6.4	6.4	6.4
OPEN IPL	1/ 1/85	12/31/89	2.7							2.7	2.7	2.7	2.7
OPEN PPL	9/30/85	12/31/89	25.5							16.4	25.5	25.5	25.5
SAS-3			1.0	0.5									
SL-1	4/ /82		3.8				3.8						
SL-2	4/ /82		9.1				9.1						
SL-3	3/ /81		9.1				9.1						
SL's[a]	3/yr		27.3					27.3	27.3	27.3	27.3	27.3	27.3
SM D/L	9/15/81	9/30/82	6.0			1.8	4.5						
SM D/M	6/ 1/83	9/30/86	8.5					5.0	8.5	8.5	6.4		
SME	9/ 1/81	8/31/83	1.0			0.3	1.0	0.7					
SMM	1/31/80	1/31/85	18.0		16.5	18.0	18.0	18.0	18.0	1.5			
ST	12/ 1/83	11/30/98	23.2					1.9	23.2	23.2	23.2	23.2	23.2
UK-5	10/15/74	9/30/80	2.0	2.0	1.5								
TOTAL RATE				26.9	46.7	48.8	74.1	77.2	94.6	98.1	130.5	112.0	93.5
TOTAL in 10^{12} bits				0.84	1.47	1.53	2.33	2.42	3.00	3.08	4.10	3.52	2.84

[a] Assumes three 7-day missions with an average data rate of 472 kbps.

TABLE A.3 Annual Data Rates of OSS/DSN Missions in Kilobits per Second

Mission	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988
Pioneer-10	0.1	0.1	0	0	0	0	0	0	0	0
Pioneer-11	0.1	0	0	0	0	0	0	0	0	0
Vikings-1 and -2	1.0	–	–	–	–	–	–	–	–	–
Voyager-1	31.8	31.8	2.1	2.1	2.1	2.1	2.1			
Voyager-2	31.8	2.1	31.8	2.1	2.1	2.1	2.1	28.6		
Pioneer Venus										
Orbiter	0.7									
Probe	0									
Galileo				5.4	7.3	7.3	28.3	35.0	6.4	
ISPM										
NASA					1.6	1.7	1.7	1.7	1.2	
ESA					0.8	0.8	0.8	0.8	0.6	
VOIR										3758.0
Helios-1 and -2	2.0	2.0	2.0	2.0						
TOTAL	67.5	36.0	35.9	11.6	13.9	14.0	35.0[a]	66.1	8.2	
Nonimaging	6.4	6.4	6.4	9.6	12.7	12.7	31.8[a]	31.8	6.4	

[a]Does not include VOIR.

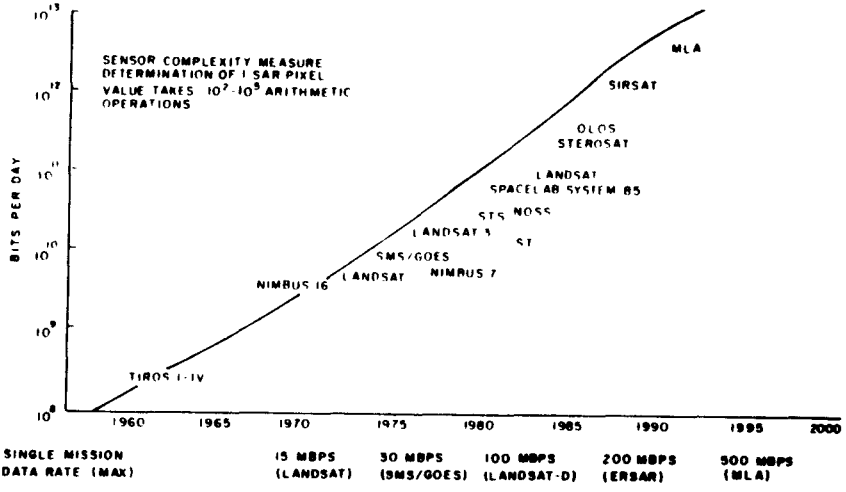

FIGURE A.1 System driver: data quantity/complexity.

TABLE A.4 Existing EDIS Digital Data Base, as of December 1980

	Currently Archived (Bytes)	Growth/Year (Bytes)
Conventional Observations		
Oceanographic	6.8×10^9	1.0×10^9
Geophysical and solar	4.0×10^{10}	5.0×10^9
Atmospheric	2.4×10^{11}	1.2×10^{10}
Satellite Observations		
Meteorological	10.3×10^{12}	6.3×10^{12}
Oceanographic	6.25×10^{10} [a]	(2.3×10^{12}) [a]
Land		(3.1×10^{12}) [b]
		11.718×10^{12}

[a] Beginning 1986.
[b] Beginning 1985.

TABLE A.5 NSSDC Tapes by Program as of December 31, 1979

Category	Satellites	No. of Data Tapes
Physics and astronomy	AE	38
Domestic	Explorer	4,850
	Injun	39
	OAO	424
	OGO	2,389
	OSO	1,561
	Pegasus	2
	RAE	2
	SAS	1
	S-Cubed	2,417
	Skylab	1,169
	Vanguard	1
Cooperative	Alouette	215
	Ariel	569
	ASTP-Apollo	1
	Eole	1
	GRS	47
	ISEE	63
	ISIS	419
	IUE	77
	TOTAL	14,285
Planetary	Mariner	145
	Pioneer	440
	Viking	304
	TOTAL	889
Lunar sciences	Apollo	3,271
	Lunar Orbiter	100
	Surveyor	10
	TOTAL	3,381
Earth observations	GOES	1,433
	Nimbus	6,678
	SMS	3,968
	TIROS	1,047
	TOTAL	13,126

TABLE A.5 Continued

Category	Satellites	No. of Data Tapes
Earth and ocean physics	BE	98
	Echo	5
	GEOS	486
	LAGEOS	54
	PAGEOS	69
	Magsat	2
	TOTAL	714
Communications	ATS	1,335
	Relay	17
	TOTAL	1,352
Department of Defense	DMSP-F2	1
	ERS	47
	OV	157
	Radsat	1
	Seasat	20
	Solrad	20
	TIP	2
	Vela	3
	1963-038C	543
	1964-083C	1
	1976-059A	2
	1977-007A	2
	TOTAL	799
Private industry	Telstar	13
	TOTAL	13
Foreign	Cosmos	1
	D1-C	4
	D1-D	4
	ESA-GEOS	23
	ESRO	3
	HEOS	21
	Intelsat	1
	Starlette	76
	TOTAL	133
	GRAND TOTAL	34,692

TABLE A.6 NSSDC Microfilm Data (100-ft Reels) as of December 31, 1979

Category	Satellites	Ephemeris	Nonephemeris	Total
Physics and astronomy	AD	35	–	35
Domestic	AE-A	3	1	4
	AE-B	4	1	5
	AE-C	28	–	28
	AE-D	3	–	3
	AE-E	19	–	19
	AEROS	10	–	10
	ANS	7	–	7
	DME-A	20	67	87
	EPE	39	48	87
	Explorer	1,675	3,191	4,866
	Hawkeye	14	–	14
	HEAO	5	–	5
	IE-A	12	1,127	1,139
	Injun	57	13	70
	Meteoroid Tech Sat	6	–	6
	OAO	39	110	149
	OGO	197	1,046	1,243
	OSO	110	938	1,048
	Pegasus	50	–	50
	RAE	28	10	38
	RM-1	1	–	1
	S15	7	–	7
	S55	15	–	15
	S-Cubed-A	10	151	161
	SAS	34	1,736	1,770
	Skylab	7	182	189
	TD1A	7	–	7

	TOPO	1	–	1
	Vanguard	24	–	24
Cooperative	Alouette	193	8,043	8,236
	Ariel	55	54	109
	Eole	4	–	4
	FR-1	11	2	13
	GRS	2	3	5
	Intasat	43	–	43
	ISEE	8	16	24
	ISIS	327	4,729	5,056
	IUE	4	–	4
	San Marco	25	–	25
	UK-5	16	–	16
	TOTAL	3,155	21,468	24,623
Planetary	Mariner	–	65	65
	Pioneer	4	205	209
	Viking	–	182	182
Lunar sciences	Apollo	1	1,651	1,652
	Lunar Orbiter	–	991	991
	Surveyor	–	9	9
	TOTAL	5	3,103	3,108
Earth observations	ESSA	115	–	115
	Gemini	1	–	1
	GOES	14	–	14
	HCMM	4	–	4
	ITOS	4	–	4
	Landsat	37	–	37
	Nimbus	86	–	86

TABLE A.6 Continued

Category	Satellites	Ephemeris	Nonephemeris	Total
	NOAA	36	–	36
	Seasat	2	–	2
	SMS	24	–	24
	TIROS	141	2	143
	TOTAL	464	2	466
Earth and ocean physics	BE-B	30	3	33
	BE-C	47	1	48
	Echo-1	38	–	38
	Echo-2	33	–	33
	GEOS-1	7	–	7
	GEOS-2	31	–	31
	GEOS-3	14	–	14
	PAGEOS	24	–	24
	TOTAL	224	4	228
Communications	ATS	86	68	154
	Early Bird	1	–	1
	HERMES	8	–	8
	Relay	40	2	42
	Syncom	13	–	13
	TETR	18	–	18
	TOTAL	166	70	236
Department of Defense	1963-38C	–	3	3
	1965-65E	1	–	1
	ANNA 1B	1	–	1
	Aurora 1	3	–	3

	Cannonball 2	2	–	2
	Courier	6	–	6
	Discoverer 25	–	1	1
	DMSP	–	564	564
	ERS	55	3	58
	Geophys Res Sat	2	–	2
	GGSE	14	–	14
	LOFTI 2A	3	–	3
	Midas	–	1	1
	OV	29	2	31
	P11	–	1	1
	P14	–	3	3
	SECOR	45	1	46
	Solrad	53	2	55
	TRAAC	3	–	3
	Transit	26	1	27
	Vela	20	2	22
	TOTAL	263	584	847
Private industry	Telstar	16	–	16
	TOTAL	16	–	16
Foreign	Aurorae	1	–	1
	Boreas	1	–	1
	Cosmos	2	1	3
	D5-A	11	–	11
	D5-B	1	–	1
	Diademe 2	1	–	1
	ESA-GEOS	2	2	4
	ESRO	6	–	6
	Intelsat	41	–	41

TABLE A.6 Continued

Category	Satellites	Ephemeris	Nonephemeris	Total
	HEOS	3	6	9
	Kyokko	–	6	6
	Meteor	–	1	1
	Peole	5	–	5
	Salute	1	–	1
	Shinsei	1	–	1
	SIGNE 3	1	–	1
	Vostok	2	–	2
	TOTAL	79	16	95
Miscellaneous	Biosatellite 2	14	–	14
	PAC-A	9	–	9
	SERT-2A	9	–	9
	TOTAL	32	–	32
	GRAND TOTAL	4,404	25,247	29,651

TABLE A.7 NSSDC Photographic Data as of December 31, 1979

Satellite	Quantity of Data
Apollo	
Hasselblad photography	16,235 frames (70-mm film)
Maurer photography	26 canisters of 16-mm film
Stellar photography	3 reels of 70-mm and 1 reel of 35-mm film
Metric photography	10,153 frames (5-in. film)
Pan photography	4,732 frames (5-in. film)
Pan photography (rectified)	4,454 frames (9-in. film)
Microfiche catalogs	30,245 frames (874 microfiche cards)
TV kinescope photography	88 canisters of 16-mm film
Nikon photography	3 canisters of 35-mm film
Lunar sample photography	17,209 frames (various formats)
ATS-6	3,860 frames (70-mm film)
Gemini	1,200 frames (70-mm film)
GOES-1	3,467 frames (70-mm film)
IUE	4,654 frames (8-in. x 10-in. film)
Luna-22	
Panoramas	8 frames (4-in. x 5-in. film)
Lunar Orbiter	
Boeing enhancements	93 frames (16-in. x 20-in. film)
LARC	3,102 frames (20-in. x 24-in. film)
AMS	3,300 frames (20-in. x 24-in. film)
LARC Framelets	75,000 framelets (35-mm film)
Microfilm catalog (AMS)	3,300 frames (5 reels of 35-mm microfilm)
Microfiche catalog (AMS)	3,300 frames (75 microfiche)
Microfiche catalog (LRC)	3,102 frames (75 microfiche)
Mariner-6 and -7	
Mariner-6 and -7 mosaics	7 frames (70-mm film)
Mariner -6 and -7 photography	1,415 frames (70-mm film)
Mariner-9	
IPL photography	16,821 frames (70-mm film)
MTVS photography	25,000 frames (70-mm film)
IPL microfiche catalog	3,500 frames (251 microfiche)
Limb microfiche catalog	850 frames (166 microfiche)
JPL mosaics	96 frames (4-in. x 5-in. film)
Cal Tech microfiche	20,015 frames (467 microfiche)
Limb microfiche index	16 microfiche
Press release negatives	55 frames (4-in. x 5-in. film)
MTVS microfiche catalog	23,000 frames (771 microfiche)
Mariner-10	
Mariner-10 Venus photography	7,187 frames (70-mm film)
Mariner-10 Earth/Moon photography	918 frames (70-mm film)
Mariner-10 Earth/Moon microfiche	16 microfiche
Mariner-10 globe mosaics	452 frames (70-mm film)

TABLE A.7 Continued

Satellite	Quantity of Data
Mercury photography	5,511 frames (70-mm film)
Press release negatives	44 sheets (4-in. x 5-in. film)
Venus microfiche	120 microfiche
Mercury microfiche	94 microfiche
SEDR support data on microfiche	2 microfiche
Mercury IPL photography	655 frames (70-mm film)
Mercury IPL stereo	264 frames (70-mm film)
Mercury IPL Cal Tech microfiche	17 microfiche
Nimbus	
IR photofacsimile	109,949 frames (70-mm film)
IR photofacsimile	4,941 frames (4-in. x 5-in. film)
ESMR color photography	43 frames (8-in. x 10-in. film)
ESMR B/W images	10,459 frames (70-mm film)
SCAMS images	4,705 frames (70-mm film)
Pioneer-10	
Polarization data	185 microfiche
Photography	250 frames (5-in. x 7-in.; 8-in. x 10-in. film)
Pioneer image log	1 microfiche
Press release negatives	25 frames (4-in. x 5-in. film)
Pioneer-11	
Polarization data	226 microfiche
Photography	376 frames (5-in. x 7-in.; 8-in. x 10-in. film)
Pioneer image log	1 microfiche
Press release negatives	28 frames (4-in. x 5-in. film)
Ranger	17,260 frames (35-mm film)
SMS-1	3,593 frames (70-mm film)
SMS-2	2,726 frames (70-mm film)
Surveyor	
Original version	100,000 film chips (70-mm film)
Mosaics	2,925 mosaic frames (4-in. x 5-in. film)
Enhanced version	760 enhanced frames (35-mm film chips)
Venera-9 descent craft	
Photography	1 frame (8-in. x 10-in. film)
Venera-10 descent craft	
Photography	1 frame (8-in. x 10-in. film)
Viking-1 Lander	
Press release B/W photography	18 frames (4-in. x 5-in. film)
Press release color photography	4 frames (4-in. x 5-in. film)
TDR Lander imaging photography	1,412 frames (5-in. x 12-in. film)
EDR Lander photography	5 microfiche
TDR color imaging photography	61 frames (5-in. x 12-in. film)
Donut projection imaging photography	4 frames (8-in. x 10-in. film)

TABLE A.7 Continued

Satellite	Quantity of Data
High-resolution mosaic imaging photography	4 frames (8-in. × 10-in. film)
TDR color picture catalog	1 microfiche
Multi-CE label color photography	2 frames (5-in. × 12-in. film)
TDR-IPL prime mission catalog	2 microfiche
Magneti imaging photography	32 frames (5-in. film)
EDR Lander imaging photography	500 frames (5-in. film)
Viking-1 Orbiter	
Press release B/W photography	49 frames (4-in. × 5-in. film)
Rectilinear orbital photography	27,830 frames (5-in. film)
Orthographic orbital photography	12,790 frames (5-in. film)
Press release color photography	7 frames (4-in. × 5-in. film)
Mosaic imaging photography	460 frames (4-in. × 5-in. film)
Stereo pair imaging photography	28 frames (5-in. × 5-in. film)
Index by 10° box and lat/long	6 microfiche
Mosaic summary and index	4 microfiche
Index to Phobos, Deimos, Star, Limb, and Terminator	1 microfiche
Index by roll and file	4 microfiche
IPL false color imaging photography	12 frames (5-in. × 12-in. film)
IPL B/W imaging photography	300 frames (5-in. film)[a]
Prime/extended mission picture catalog	501 microfiche
USGS photomosaics	60 frames (8-in. × 10-in. film)
Viking-2 Lander	
Press release B/W photography	13 frames (4-in. × 5-in. film)
Press release color photography	2 frames (4-in. × 5-in. film)
TDR Lander imaging photography	1,361 frames (5-in. × 12-in. film)
EDR Lander imaging photography	1,378 frames (5-in. film)
EDR-IPL picture catalog	6 microfiche
TDR color imaging photography	47 frames (5-in. × 12-in. film)
High-resolution mosaic imaging photography	24 frames (8-in. × 10-in. film)
Donut projection imaging photography	6 frames (8-in. × 10-in. film)
TDR-IPL prime mission catalog	3 microfiche
Multi-CE label color photography	2 frames (5-in. × 12-in. film)
TDR color picture catalog	1 microfiche
Magnet imaging photography	47 frames (5-in. film)
Viking-2 Orbiter	
Press release photography	10 frames (4-in. × 5-in film)
Mosaic imaging photography	234 frames (4-in. × 5-in. film)
Rectilinear orbital photogrpahy	20,698 frames (5-in. film)

TABLE A.7 Continued

Satellite	Quantity of Data
Orthographic orbital photography	9,793 frames (5-in. film)
Stereo pair imaging photography	24 frames (5-in. film)[a]
Index by 10° box and lat/long	6 microfiche
Mosaic summary and index	4 microfiche
Index to Phobos, Deimos, Star, Limb, and Terminator	1 microfiche
Index to roll and file	4 microfiche
IPL false color imaging photography	26 frames (5-in. x 12-in. film)
Prime/extended-mission picture catalog	304 microfiche
IPL B/W imaging photography	300 frames (5-in. film)[a]
Voyager	
Press release photography	169 frames (4-in. x 5-in. film)
TV imaging	3,925 frames (5-in. film)

[a] Stereo pair imaging and IPL B/W imaging previously reported together as stereo pairs.

SUMMARY

1. In addition to the data currently being acquired from existing and planned space missions, there is a large volume of data already existing from past space missions and from associated ground measurements. These are sometimes resident in integrated data bases, which also often contain additional data that have been processed into a form for a specific investigation. Thus, the total volume of data estimated in this Appendix is actually only a fraction of that available.

2. The data volumes currently being acquired are of the order of 10^{14} bits per year.

3. The data volumes from applications spacecraft (i.e., Landsat, GOES) are about 2 orders of magnitude larger than those from OSS-type missions. This situation is likely to continue for the rest of the decade, at least. When the application satellites become operational (e.g., Landsat in the latter half of the 1980's), even more data will be acquired.

4. Many other countries are planning both space-science and space-applications missions. These countries include France, Japan, Germany, the Soviet Union, and India. This presents both data opportunities and data problems that need to be addressed.

Abbreviations Used In Text

ADS	Applications Data Service
AE	Atmospheric Explorer
bpi	Bits per inch
CAD	Computer-aided design
CAM	Computer-aided manufacture
CCT	Computer-compatible tape
CDAW	Coordinated Data Analysis Workshop
CMS	Conversational Monitor System
CODMAC	Committee on Data Management and Computation
COM	Computer-originated microfilm
CPU	Central processing unit
CRT	Cathode-ray tube
DAL	Data Analysis Laboratory
DBMS	Data-base management system
DDMU	Discipline Data Management Unit
DEC	Digital Equipment Corporation
DGS	Data ground site
DMS	Data-management system
DOD	Department of Defense
DOI	Department of the Interior
Domsat	Domestic Communications Satellite
DRSGS	Direct Readout Satellite Ground Stations
EDC	EROS Data Center
EDIPS	EROS Digital Image Processing System
EDIS	Environmental Data and Information Service (NOAA)

EDPF	Einstein Data Processing Facility
EEIS	End-to-End Information System
ENDEX	Environmental Data Index
EROS	Earth Resource Observation System
ESA	European Space Agency
FAA	Federal Aviation Administration
FOVLIP	First-order Viking Lander image processing
GOES	Geostationary Operational Environmental Satellite
GSFC	Goddard Space Flight Center
GU	Geophysical units
HDT	High-density tape
HEAO	High Energy Astrophysical Observatory
IDT	Investigation Definition Team
IMS	International Magnetospheric Study
INORAC	Inquiry, Order and Accounting System
IPD	Information Processing Division (GSFC)
IPL	Image Processing Laboratory (JPL)
IR	Infrared radiometer
ISEE	International Sun-Earth Explorer
JPL	Jet Propulsion Laboratory
kbps	Kilobits per second
KSA	Single-access K band
LACIE	Large Area Crop Inventory Experiment
LARS	Laboratory for Applications of Remote Sensing (Purdue University)
LAS	Landsat Assessment System
LASP	Laboratory for Atmospheric and Space Physics
LSI	Large-scale integration
Mbps	Million (mega) bits per second
MDOD	Missions and Data Operations Directorate (GSFC)
MIT	Massachusetts Institute of Technology
MOPS	Million operations per second
MSA	Magnetic solar activity
MSS	Multispectral scanner
NAFAX	National facsimile (service)
NASA	National Aeronautics and Space Administration
NEEDS	NASA End-to-End Data System
NESS	National Environmental Satellite Service (NOAA)
NGSDC	National Geophysical and Solar-Terrestrial Data Center (NOAA)
NOAA	National Oceanographic and Atmospheric Administration
NSSDC	National Space Science Data Center (NASA)

OASIS	Oceanic and Atmospheric Information System
OCC	Operations Control Center
OSS	Office of Space Science (NASA)
OSTA	Office of Space and Terrestrial Applications (NASA)
PI	Principal Investigator
RBV	Return-beam vidicon
RPIF	Regional Planetary Image Facility
SACC	Space Applications Computer Center
SAO	Smithsonian Astrophysical Observatory
SAR	Synthetic-aperture radar
SAS	Small Astronomy Satellite
ScI	Science Institute (Space Telescope)
SDMU	Scientific Data Management Unit
SDSD	Satellite Data Services Division (NOAA)
SI	Scientific instruments
SMMR	Passive microwave radiometer
SOGS	Science Operations Ground System (for ST)
SSB	Space Science Board
SSDS	Space Science Data Service
SSG	Science Steering Group
ST	Space Telescope
STOCC	Space Telescope Operations and Control Center
STS	Space Transportation System (Shuttle)
SWG	Science Working Group
TDM	Time-Division Multiplex
TDMA	Time-Division Multiplex Access
TDRSS	Tracking and Data Relay Satellite System
TM	Thematic mapper
UA	Unified Abstract
uhf	Ultra-high frequency
USGS	U.S. Geological Survey
vhf	Very-high frequency
VICAR	Video Communication and Retrieval System
VLSI	Very-large-scale integration
VMS	Virtual Memory System
VOIR	Venus Orbiting Imaging Radar
WDC	World Data Center