

ASSESSMENT OF THE USEFULNESS AND AVAILABILITY OF NASA'S EARTH AND SPACE SCIENCE MISSION DATA

**SPACE STUDIES BOARD
BOARD ON EARTH SCIENCES AND RESOURCES
NATIONAL RESEARCH COUNCIL**

Other Reports of the Space Studies Board

Toward New Partnerships: Government, the Private Sector, and Earth Science Research (2002)

The Quarantine and Certification of Martian Samples (2001)

U.S. Astronomy and Astrophysics: Managing an Integrated Program (2001)

Readiness Issues Related to Research in the Biological and Physical Sciences on the International Space Station (2001)

“The Next Generation Space Telescope” (2001)

Assessment of Mars Science and Mission Priorities (2001)

The Mission of Microgravity and Physical Sciences Research at NASA (2001)

Transforming Remote Sensing Data into Information and Applications (2001)

Signs of Life: A Report Based on the April 2000 Workshop on Life Detection Techniques (2001)

Assessment of Mission Size Trade-offs for Earth and Space Science Missions (2000)

“Assessment of NASA’s Office of Space Science Strategic Plan 2000” (2000)

“Assessment of Scientific Aspects of the Triana Mission” (2000)

“Continuing Assessment of Technology Development in NASA’s Office of Space Science” (2000)

Ensuring the Climate Record from the NPP and NPOESS Meteorological Satellites (2000)

Future Biotechnology Research on the International Space Station (2000)

Issues in the Integration of Research and Operational Satellites for Climate Research: I. Science and Design (2000)

Microgravity Research in Support of Technologies for the Human Exploration and Development of Space and Planetary Bodies (2000)

Preventing the Forward Contamination of Europa (2000)

Review of NASA’s Biomedical Research Program (2000)

Review of NASA’s Earth Science Enterprise Research Strategy for 2000-2010 (2000)

The Role of Small Satellites in NASA and NOAA Earth Observation Programs (2000)

“Assessment of NASA’s Plans for Post-2002 Earth Observing Missions” (1999)

Institutional Arrangements for Space Station Research (1999)

Radiation and the International Space Station: Recommendations to Reduce Risk (1999)

A Science Strategy for the Exploration of Europa (1999)

A Scientific Rationale for Mobility in Planetary Environments (1999)

Size Limits of Very Small Microorganisms: Proceedings of a Workshop (1999)

U.S.-European-Japanese Workshop on Space Cooperation: Summary Report (1999)

Assessment of Technology Development in NASA’s Office of Space Science (1998)

Development and Application of Small Spaceborne Synthetic Aperture Radars (1998)

Evaluating the Biological Potential in Samples Returned from Planetary Satellites and Small Solar System Bodies: Framework for Decision Making (1998)

The Exploration of Near-Earth Objects (1998)

Exploring the Trans-Neptunian Solar System (1998)

Failed Stars and Super Planets: A Report Based on the January 1998 Workshop on Substellar-Mass Objects (1998)

Ground-based Solar Research: An Assessment and Strategy for the Future (1998)

Readiness for the Upcoming Solar Maximum (1998)

A Strategy for Research in Space Biology and Medicine in the New Century (1998)

Supporting Research and Data Analysis in NASA’s Science Programs: Engines for Innovation and Synthesis (1998)

U.S.-European Collaboration in Space Science (1998)

Copies of these reports are available free of charge from:

Space Studies Board
National Research Council
2101 Constitution Avenue, NW
Washington, DC 20418
(202) 334-3477
ssb@nas.edu
www.nationalacademies.org/ssb/ssb.html

Assessment of the Usefulness and Availability of NASA's Earth and Space Science Mission Data

Task Group on the Usefulness and Availability of NASA's Space Mission Data

Space Studies Board
Division on Engineering and Physical Sciences
Board on Earth Sciences and Resources
Division on Earth and Life Studies

National Research Council

NATIONAL ACADEMY PRESS
Washington, D.C.

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 task group responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by Contract NASW 96013 between the National Academy of Sciences and the National Aeronautics and Space Administration. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

Copies of this report are available free of charge from:

Space Studies Board
National Research Council
2101 Constitution Avenue, N.W.
Washington, DC 20418

Copyright 2002 by the National Academy of Sciences. All rights reserved.

Printed in the United States of America

THE NATIONAL ACADEMIES

National Academy of Sciences
National Academy of Engineering
Institute of Medicine
National Research Council

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Bruce M. Alberts is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Wm. A. Wulf is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Kenneth I. Shine is president of the Institute of Medicine.

The **National Research Council** was organized 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 advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Bruce M. Alberts and Dr. Wm A. Wulf are chairman and vice chairman, respectively, of the National Research Council.

TASK GROUP ON THE USEFULNESS AND AVAILABILITY OF NASA'S SPACE MISSION DATA

SIDNEY C. WOLFF, National Optical Astronomy Observatories, Tucson, *Chair*
THOMAS A. HERRING, Massachusetts Institute of Technology, *Vice-Chair*
JOEL BREGMAN, University of Michigan, Ann Arbor
MICHAEL J. FOLK, University of Illinois, Urbana-Champaign
RICHARD G. KRON, University of Chicago
JAMES F.W. PURDOM, Colorado State University
DONNA L. SHIRLEY, University of Oklahoma, Norman
WALTER H.F. SMITH, National Oceanic and Atmospheric Administration
NICK VAN DRIEL, U.S. Geological Survey/EROS Data Center
DONALD J. WILLIAMS, Johns Hopkins University
ROGER V. YELLE, Northern Arizona University, Flagstaff
JAMES R. ZIMBELMAN, Smithsonian Institution

National Research Council Staff

ANNE M. LINN, Study Director
MONICA R. LIPSCOMB, Assistant Program Officer
JOSEPH K. ALEXANDER, Director, Space Studies Board
ANTHONY R. DE SOUZA, Director, Board on Earth Sciences and Resources
CLAUDETTE BAYLOR-FLEMING, Senior Program Assistant
EDMUND M. REEVES, Consultant

SPACE STUDIES BOARD

JOHN H. McELROY, University of Texas at Arlington (retired), *Chair*
ROGER P. ANGEL, University of Arizona
JAMES P. BAGIAN, National Center for Patient Safety, Veterans Health Administration
JAMES L. BURCH, Southwest Research Institute
RADFORD BYERLY, JR., Boulder, Colorado
ROBERT E. CLELAND, University of Washington
HOWARD M. EINSPAHR, Bristol-Myers Squibb Pharmaceutical Research Institute
STEVEN H. FLAJSER, Loral Space and Communications Ltd.
MICHAEL H. FREILICH, Oregon State University
DON P. GIDDENS, Georgia Institute of Technology/Emory University
RALPH H. JACOBSON, The Charles Stark Draper Laboratory (retired)
CONWAY LEOVY, University of Washington
JONATHAN I. LUNINE, University of Arizona
BRUCE D. MARCUS, TRW, Inc. (retired)
RICHARD A. McCRAY, University of Colorado
HARRY Y. McSWEEN, JR., University of Tennessee
GARY J. OLSEN, University of Illinois at Urbana-Champaign
GEORGE A. PAULIKAS, The Aerospace Corporation (retired)
ROBERT J. SERAFIN, National Center for Atmospheric Research
EUGENE B. SKOLNIKOFF, Massachusetts Institute of Technology
MITCHELL SOGIN, Marine Biological Laboratory
C. MEGAN URRY, Yale University
PETER W. VOORHEES, Northwestern University

JOSEPH K. ALEXANDER, Director

BOARD ON EARTH SCIENCES AND RESOURCES

RAYMOND JEANLOZ, University of California, Berkeley, *Chair*

JILL BANFIELD, University of California, Berkeley

STEVEN R. BOHLEN, Joint Oceanographic Institution

VICKI J. COWART, Colorado Geological Survey

DAVID L. DILCHER, University of Florida

ADAM DZIEWONSKI, Harvard University

WILLIAM L. GRAF, University of South Carolina

RHEA GRAHAM, New Mexico Interstate Stream Commission

GEORGE M. HORNBERGER, University of Virginia

DIANNE R. NIELSON, Utah Department of Environmental Quality

MARK SCHAEFER, NatureServe

BILLIE L. TURNER II, Clark University

THOMAS J. WILBANKS, Oak Ridge National Laboratory

ANTHONY R. DE SOUZA, Director

Preface

Space flight missions, which generate vast quantities of data, are the most visible and costly elements of the National Aeronautics and Space Administration's (NASA's) earth and space science programs. However, the acquisition of data by flight missions is only one step in generating knowledge. Advances in scientific understanding also require the ability to collect, share, and save data; the computational power to reduce data and create models; communications to move data from one place to another; structures to manage the data and associated resources; and access to data over extended time periods. Through these activities, data from flight projects are transformed into knowledge about the world and universe in which we live. The analysis of data also provides the foundation and often leads to the enabling technology for planning future NASA missions.

In recognition of the importance of space mission data and data management in space research, the House conference report on Fiscal Year 2000 appropriations for NASA noted:

The conferees are concerned that the large amount of data being collected as part of NASA science missions is not being put to the best possible use. To allay these concerns, the conferees direct NASA to contract with the National Research Council for the study of the availability and usefulness of data collected from all of NASA's science missions. The study should also address what investments are needed in data analysis commensurate with the promotion of new missions.¹

In response to a subsequent letter from NASA's associate administrators for earth science and for space science (see Appendix A), the National Research Council (NRC) charged the Space Studies Board and the Board on Earth Sciences and Resources to undertake a study. The Task Group on the Usefulness and Availability of NASA's Space Mission Data, composed of experts from earth, space, and information sciences (see Appendix D), was established to address three sets of questions, as follows:

1. How available and accessible are data from science missions (after expiration of processing and proprietary analysis periods, if any) from the point of view of both scientists in the larger U.S. research community, as well as U.S. education, public outreach and policy specialists, and private industry? What, if anything, should be changed to improve accessibility?
2. How useful are current data collections and archives from NASA's science missions as resources in support of high priority scientific studies in each Enterprise? How well are areas such as data preservation, documentation, validation, and quality control being addressed? Are there significant

¹House Conference Report 106-379 to Accompany H.R. 2684, Making Appropriations for the Departments of Veterans Affairs and Housing and Urban Development, and for Sundry Independent Agencies, Boards, Commissions, Corporations, and Offices for the Fiscal Year Ending September 30, 2000, and for Other Purposes, 106th Congress, U.S. Government Printing Office, Washington, D.C., October 13, 1999, p. 155.

obstacles to appropriately broad scientific use of the data? Are there impediments to distribution of derived data sets? Are there any changes in data handling and data dissemination that would improve usefulness?

3. Keeping in mind that NASA receives appropriated funds for both mission development as well as analysis of data from earlier or currently operating missions, is the balance between attention to mission planning and implementation versus data utilization appropriate in terms of achieving the objective of the Enterprises? Should the fraction of a mission's life-cycle cost devoted to data analysis, processing, storage and accessibility be changed?

Because NASA's earth and space science programs have generated thousands of data sets that are stored in dozens of facilities and are used by several hundred thousand users in the United States and abroad, it is not possible to analyze every data set or consider every use within the confines of an NRC study. Consequently, the task group focused on the usefulness, availability, and accessibility of data for the scientific community, while remaining cognizant of a second tier of users who are interested in educational, commercial, or policy applications of space mission data. Only the major data-handling facilities were considered; individual data sets held in mission databases, in science project offices, or in the hands of principal investigators were not examined in detail. Finally, issues of documentation, validation, and quality control of individual data sets were only indirectly addressed, as a measure of the usefulness of the data. The task group concluded that this approach was appropriate, given what it understood to be the primary intent of the Congress, the science focus of NASA, and the need to stay within the bounds of the schedule and resources available for the study. The task group also concluded that the charter was directed primarily toward data collected and stored digitally, such as imaging data, rather than toward returned physical materials (e.g., samples). In keeping with the NASA letter of request, the task group focused its attention on the NASA earth and space science programs and did not consider related activities in the NASA Office of Biological and Physical Research.

Most of the issues raised in the charge have been addressed by previous NRC or NASA advisory committees, commonly at a level of detail not possible in this broad study. Rather than duplicate their efforts, the task group used their reports whenever possible. In addition, the task group gathered its own information through briefings at its three meetings; interviews with chairs of NASA advisory committees, working scientists, and archive managers; and a questionnaire to 16 NASA earth and space science archives, data centers, and data services (Appendix C). The task group also invited input from the two parent NRC boards—the Space Studies Board and the Board on Earth Sciences and Resources—and their relevant disciplinary committees. Finally, most of the members of the task group are users of NASA data. In addition to drawing on their own experience, they reviewed relevant Web pages and retrieved data for this study.

The task group wishes to acknowledge the assistance of the many individuals who gave presentations or provided information for the study: Mark Abbott, Waleed Abdalati, Charles Acton, Michael A'Hearn, Raymond Arvidson, Bruce Barkstrom, Reta Beebe, Bruce Berriman, David Black, Kirk Borne, Joseph Bredekamp, Bruce Caron, Robert Chen, Cynthia Cheung, Donald Collins, James Conner, Jacques Descloîtres, Elaine Dobinson, Eric Eliason, Wendy Freedman, Andrea Ghez, David Glover, Sara Graves, Vanessa Griffen, Joseph Gurman, Frank Hill, Lee Holcomb, Thomas Kalvelage, Thomas Karl, Jack Kaye, Steven Kempler, Joseph King, Susan LaVoie, John Leibacher, Francis Lindsay, Jeffrey Linsky, Dawn Lowe, Barry Madore, Martha Maiden, Richard McGinnis, Blanche Meeson, Mike Moore, Richard Mushotzky, Philip Nicholson, Frazer Owen, Dolly Perkins, Judith Pipher, Marc Postman, Guenter Riegler, Jeff Rosendhal, Cassandra Runyon, Ethan Schreier, Mark Showalter, Roger Smith, Paul Steinhart,

Terry Teays, John Townshend, Larry Voorhees, Raymond Walker, Ronald Weaver, Ming-Ying Wei, Steven Wharton, and Nicholas White. We also appreciate the valuable contributions of David DeWitt, who served on the task group through December 2001. Finally, the task group wishes to express special thanks to the NRC study director, Anne Linn. The broad knowledge of NASA programs in earth sciences and in data management that she has acquired during her years of service at the NRC played an essential role in ensuring that the task group acquired quickly and efficiently the information and perspectives needed to arrive at its assessments and to complete this report on schedule.

Acknowledgment of Reviewers

This report has been reviewed by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the National Research Council's (NRC's) Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the authors and the NRC in making the published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The contents of the review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their participation in the review of this report:

Otis Brown, University of Miami,
John Christy, University of Alabama-Huntsville,
Lennard Fisk, University of Michigan,
Steve Holt, Franklin W. Olin College, Babson College,
Melissa A. McGrath, Space Telescope Science Institute,
David G. Sibeck, Johns Hopkins University,
Charles L. Steele, Stanford University,
Leon Stout, Pennsylvania State University Libraries, and
John R.G. Townshend, University of Maryland, College Park.

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations, nor did they see the final draft of the report before its release. The review of this report was overseen by Richard M. Goody, Harvard University (emeritus), and Mark R. Abbott, Oregon State University. Appointed by the National Research Council, they were responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the authoring task group and the institution.

Contents

| | |
|---|----|
| EXECUTIVE SUMMARY | 1 |
| 1 NASA: A KNOWLEDGE AGENCY | 10 |
| Introduction, 10 | |
| Space Science Enterprise, 11 | |
| Earth Science Enterprise, 19 | |
| The Changing Paradigm for NASA, 23 | |
| 2 ACCESSIBILITY OF DATA: THE ARCHITECTURE OF THE ARCHIVES | 24 |
| Space Science Data Systems, 26 | |
| Earth Observing System Data and Information System, 34 | |
| Strategic Evolution of ESE Data Systems, 39 | |
| Long-Term Maintenance of Data, 41 | |
| 3 THE USERS OF NASA DATA | 47 |
| User Profile, 47 | |
| Availability of Earth and Space Science Data, 50 | |
| Scientific Community Users, 51 | |
| Nonscientist Users, 54 | |
| Conclusions, 61 | |
| 4 STRATEGIES FOR MANAGING EARTH AND SPACE SCIENCE DATA | 62 |
| A Comprehensive Approach to Information Management, 65 | |
| Issues of Balance: Acquisition, Analysis, and Archiving, 67 | |
| Federated Databases, 71 | |
| Elements of Effective Data Management, 76 | |
| APPENDIXES | |
| A Letter of Request, 81 | |
| B The Data Life Cycle, 83 | |
| C Questionnaire to the Active Archives, Data Centers, and Data Services, 88 | |
| D Biographies of Task Group Members, 89 | |
| E Meeting Agendas, 93 | |
| F Acronyms, 99 | |

Executive Summary

The National Aeronautics and Space Administration (NASA) has become a knowledge agency. Long after the Mars Surveyor has gone silent, Hubble has met the same fate as Mir, and the Moderate Resolution Imaging Spectroradiometer has produced its final set of images, what will endure are the volumes of valuable data that these instruments and many others have collected over their lifetimes. NASA data sets are revolutionizing the fields of astrophysics, solar system exploration, space plasma physics, and earth science. As this impressive collection of observations has grown, NASA's mission has also expanded—evolving from an emphasis on mission planning and execution to include the collection, preservation, and dissemination of earth and space data.

Spacecraft that will be launched during the next decade will increase the data volume returned by NASA missions a hundredfold. These rich data sets will open new eras in precision cosmology and in understanding of the complex linkages in the forces that shape the Earth's environment. Addressing the increasingly complex questions that can now be asked—and answered—through the use of NASA data will require the capability to compare and combine observations of different types and to discover patterns and relationships through sophisticated querying tools. The user community will need still-to-be-developed tools and methodologies for accessing, analyzing, and mining data; recognizing patterns; and performing cross-correlations that are scalable to a billion or more objects. Developing the necessary tools will present new challenges to space scientists, to the information-technology community, and to NASA. Investments in scientific analysis and in packaging data in formats useful to other potential users, including educators, those in industry, state and local government officials, and policy makers, will be needed in order to exploit the full potential of existing data sets. *The end product of each mission—knowledge—must be the key factor in determining mission design and budget allocations.*

AVAILABILITY AND USEFULNESS OF NASA'S SPACE MISSION DATA

The Task Group on the Usefulness and Availability of NASA's Space Mission Data was charged by NASA's associate administrators for earth science and space science to evaluate the availability, accessibility, and usefulness of data from earth and space science missions, and to assess whether the balance between attention to mission planning and implementation versus data analysis and utilization is appropriate. Based on input from various sources—recent National Research Council (NRC) and other advisory committee reports; interviews with the chairs of relevant NASA advisory committees and discipline committees within the NRC; information gathered from NASA headquarters; and the task group's survey of the archives, data

centers, and data services and use of their Web sites—the task group’s answers to the charge (see Appendix A) are summarized below:

Charge 1. How available and accessible are data from science missions (after expiration of processing and proprietary analysis periods, if any) from the point of view of both scientists in the larger U.S. research community, as well as U.S. education, public outreach and policy specialists, and private industry? What, if anything, should be changed to improve accessibility?

As few as 10 years ago, NASA’s data collections were accessible mainly to researchers involved with specific missions. With the advent of a NASA network of active archives, data centers, and data services, most newer data sets have become widely available, especially to researchers. Enhancements in bandwidth and planned increases in the number of online data sets available through publicly accessible data facilities will improve the accessibility of NASA’s earth and space science data still further over the next decade. However, much of the older data (e.g., in the fields of solar and space physics and planetary science) is still in the hands of principal investigators (PIs) or is not available in formats that users need. Other data or information products (e.g., education and nonscientific applications products) are available on project Web sites but may require extensive searching to find, and their long-term availability is not assured. Further improvements in cataloging and documentation will be required to help users find data.

Charge 2. How useful are current data collections and archives from NASA’s science missions as resources in support of high priority scientific studies in each Enterprise [i.e., NASA’s Earth Science Enterprise and Space Science Enterprise]? How well are areas such as data preservation, documentation, validation, and quality control being addressed? Are there significant obstacles to appropriately broad scientific use of the data? Are there impediments to distribution of derived data sets? Are there any changes in data handling and data dissemination that would improve usefulness?

The use of archival data has contributed to a number of scientific advances in the earth and space sciences (e.g., confirmation of the Antarctic ozone hole and the accelerating expansion of the universe). The large and growing number of users—coupled with the positive results of user surveys, external reviews, and the task group’s own experience with the data facilities—attests to the usefulness of the data in a wide variety of investigations.

Many data sets will grow in value as the time period covered by the measurements lengthens. However, getting the most out of existing data sets will require the development of software tools for handling the data (e.g., for changing formats, subsetting large data sets, and querying and visualizing data sets) and improvements in documentation, user interfaces, and technical and scientific support. These improvements will be even more important for dealing with the projected growth in the volume of data (one to two orders of magnitude over the next 5 years) and the increasing need to integrate disparate data sets for both research and applications purposes. Maintaining accessibility and compatibility with changing standards for storage media, software tools, and so forth in the long term will present substantial challenges in terms of both cost and management. Although issues of validation and quality control of individual data sets were not directly addressed in this study, the task group’s generally positive findings about data usefulness suggest that these issues do not now pose either major or widespread obstacles to data

use. However, they will require heightened attention in the future as demands on the active archives increase.

NASA data have the potential to benefit society in many ways, but in order to exploit this potential it is necessary to provide support for the translation of scientific data into data products that are tailored for specific applications. These data products must be easily accessed and interpreted by people who are experts in the fields to which the data are being applied, but who will very likely have limited or no training in fields for which the data were originally collected. The work of Earth Science Information Partners, Regional Earth Science Application Centers, Infomarts, and similar applications programs is an important step in increasing the usefulness of NASA data. However, meeting the needs of the broader community would require a very substantial additional investment of resources, and such investments should be preceded by an assessment of the market for NASA information and a prioritization of investments according to cost-effectiveness and likely impact.

Charge 3. Keeping in mind that NASA receives appropriated funds for both mission development as well as analysis of data from earlier or currently operating missions, is the balance between attention to mission planning and implementation versus data utilization appropriate in terms of achieving the objective of the Enterprises? Should the fraction of a mission's life-cycle cost devoted to data analysis, processing, storage and accessibility be changed?

Declines in funding for analysis of space science data in the 1990s have been reversed in recent years, although funding remains insufficient for analyzing data during extended missions or after missions have been completed. The major exception to this generalization is for long-lived astrophysics missions, where funding for data analysis, including analysis of archival data, is made available for a decade or more after launch. Despite changes in the way budgets are reported, the fragmented budget structure of both enterprises makes it difficult to quantify the adequacy or inadequacy of funding.

Rigid guidelines for the balance between support for mission planning and implementation on the one hand and data utilization on the other are inappropriate. However, in view of the expected growth and diversification in the data products from future missions, NASA should address more explicitly the issues of balance in its planning and management of missions and programs and it should do so utilizing mechanisms that involve the user communities. Trade-offs within the life-cycle budget should be made in such a way as to optimize the overall scientific return, even if that means reducing mission capabilities for data acquisition.

Specific recommendations related to the task group's charge are presented in the sections that follow.

MANAGEMENT OF DATA WITHIN NASA

Concerns about the management of NASA data sets have been identified in several earlier NRC and General Accounting Office reports. The task group concludes that the management of science data and information has become a function of sufficient scope and importance that its successful execution requires leadership with the expertise to carry out these tasks:

- Provide strategic planning, oversight, and advice concerning the collection, processing, archiving, and dissemination of data and information collected by NASA's space missions;
- Be the advocate for the appropriate balance of investment in data analysis;
- Ensure the preservation and accessibility of valuable space mission data and information;
- Require a data management plan for each mission and monitor its implementation;
- Provide oversight for the design and implementation of software, hardware, and database systems for processing and storing NASA's massive data sets;
- Develop a long-term software plan for NASA's Earth Science and Space Science Enterprises;
- Require interenterprise communication and sharing of successful methods and systems for data management;
- Work out the memorandums of understanding governing access to data from those missions that are carried out cooperatively with other countries; and
- Determine how information generated by the space programs of other countries can be accessed and effectively used by U.S. scientists and institutions.

The person(s) charged with the tasks listed above should also create and draw on the experience of an advisory panel composed of instrument scientists, computer scientists, chief information officers (CIOs) from major corporations and government organizations, and an electronic-records expert from the National Archives and Records Administration. Analogous to the position of CIO in a major corporation, the NASA person(s) in charge of the information-management function should have budgetary responsibility for the collection, analysis, and long-term maintenance of all earth and space science data sets. This responsibility could consist of either holding the budget for designing the data collection, analysis, dissemination, and archiving function for each mission or having the right of refusal for projects and programs that do not handle it adequately, or both. In parallel with the title of CIO in industry, this person might appropriately be called the chief science information officer(s) (CSIO; this title distinguishes the functions addressed here from those of the chief information officer at NASA, who is primarily responsible for NASA business systems and security). The CSIO(s) would have responsibility for the data acquisition and utilization component of every mission and would advocate investment in data management at a level that optimizes the overall scientific return of a mission when trade-offs between hardware and data must be made.

Some of the responsibilities outlined above relate to cross-NASA issues, while others are more specific to individual program offices. Accordingly, they could be carried out either by a single individual or by individuals assigned to each of the enterprises. However, whatever administrative structure is selected, it should be one that supports cross-enterprise communication and cooperation and provides the support and authority needed to ensure that the CSIO is effective in carrying out the functions identified here.

The recommendation to consolidate the information-management function does not imply that NASA should centralize all data aspects of all missions. The task group believes that a combination of distributed and centralized activities is necessary. For example, analysis and production of data products should probably continue to be performed in a distributed manner by scientists, while long-term maintenance of data is probably best handled centrally. The NASA CSIO(s) would be responsible for overseeing the development of the overall architecture of the data and information "production line," while leaving much of the actual design,

implementation, and operation to the scientists and engineers directly responsible for each mission.

Recommendation. NASA should assign the overall responsibility for oversight and coordination of NASA's data assets to a chief science information officer (CSIO) (or alternatively to multiple science officers). The CSIO(s) would provide leadership; long-term strategic planning; and advice on the collection, processing, archiving, and dissemination of data and information collected by NASA's space missions to ensure the preservation and accessibility of these valuable resources. If a single CSIO is named, then this individual should report to the NASA administrator. Alternatively, CSIOs might be appointed for each of the enterprises and report to the heads of the enterprises, but in this case a mechanism should be established to ensure cross-enterprise coordination and communication of best practices.

INVESTMENTS IN SOFTWARE AND DATA ANALYSIS

The scientific productivity of a space mission depends as much on the readiness of software and data flow pipelines as on the readiness of the sensor and spacecraft hardware. Therefore, NASA science missions should be viewed as integrated systems of hardware and software. The trade-offs among capabilities that are inevitable in missions and programs with fixed budgets must include not only the funding for new missions, the development of new capabilities, and the fabrication of spacecraft instrumentation, but also the funding for software development for mission operations, data distribution, and data analysis. In cases where hardware cost overruns occur, maintaining an adequate investment in software and scientific analysis may well require reducing the capabilities of the mission itself. Ground and flight systems should be designed in conjunction in order to achieve cost-effective data acquisition and analysis.

Recent program solicitations from both the Earth Science and Space Science Enterprises require the PIs to prepare budgets for the total mission cycle cost—from mission definition to data processing, publication, and archiving. The task group encourages the continuation of this practice.

Recommendation. Budgets for mission operations and data analysis should be included as an integral part of mission and/or program funding. Reviews, including NASA's nonadvocate review, which is required to authorize project funding, should include assessment of the data analysis elements, including archiving and timely provision of data to users. While reviews of some projects already follow this recommendation, its implementation is not uniform across all NASA programs. The appropriate balance between hardware and software investment is best determined jointly by NASA managers and the user communities involved in the mission.

The prime mission phase includes the development, launch, data collection, and analysis for a fixed period of time that is estimated to be sufficient to answer the minimum set of scientific questions that must be addressed in order for the mission to be judged a success. However, for many missions and many scientific problems, the value of data extends well beyond the termination of the prime mission phase. Missions are extended, calibrations are improved, novel

uses of the data are made that were neither foreseen nor planned by the original mission investigation team, and many significant discoveries occur only after a variety of heterogeneous data sets are integrated and studied. The peak publication rate for a mission often occurs 4 to 5 years after launch. All of these factors argue for continuation of support for scientific analysis after the prime mission phase is completed. Mechanisms (e.g., proposal pressure and advisory committees) exist for setting priorities within a discipline. However, NASA, in consultation with the scientific community, will have to develop mechanisms for addressing issues of balance across disciplines or between new missions, extended missions, and postmission data analysis within or between programs. Whatever mechanism it chooses should be carried out on a regular and systematic basis.

LONG-TERM MAINTENANCE OF DATA

NASA currently provides a data center—the National Space Science Data Center (NSSDC)—for long-term maintenance of space science data. However, the NSSDC faces tremendous challenges in serving current users as well as future generations of scientists. Many scientifically valuable data sets are not archived in the center, and those that are may not be sufficiently well documented or formatted to be readily accessible. Declining budgets and rapidly growing volumes of holdings will only exacerbate these problems.

A permanent storage facility is not even available for most of NASA's earth science data. Instead, these data are to be transferred to the U.S. Geological Survey and the National Oceanic and Atmospheric Administration 15 years after collection. Even if adequate resources can be found, transferring petabytes of data from those familiar with them to organizations with little knowledge of the data entails a risk. Because NASA data sets are a national resource and because the value of many of them increases in direct proportion to the time interval covered by them, it is important to preserve the data indefinitely. The care of the data must be accomplished so as to maximize their knowledge-enhancement possibilities, scientific impact, and discovery potential.

Recommendation. NASA should assume formal responsibility for maintaining its data sets and ensuring long-term access to them to permit new investigations that will continue to add to our scientific understanding. In some cases, it may be appropriate to transfer this responsibility to other federal agencies, but NASA must continue to maintain the data until adequate resources for preservation and access are available at the agency scheduled to receive the data from NASA.

FEDERATED DATA SYSTEMS

Many of the important scientific problems of the 21st century in both space and earth science will require the ability to explore and integrate data obtained from different spacecraft and different instruments. Rather than creating a single information system to meet the evolving needs of a wide range of users, it is now possible, and may even prove to be more cost-effective, to create a federation of distributed databases with universal standards for archiving and to provide common and easily used visualization tools. Federations capitalize on bottom-up

decision making and local, custom solutions to specific user needs. A prototype federation of Earth Science Information Partners, which has been operating for 3 years, has demonstrated the ability of different NASA-funded organizations to cooperate, provide system operability at the catalog level, and produce specialized data products. The astrophysics community has developed a plan called the National Virtual Observatory (NVO), which would provide common access tools for their multiwavelength databases; development of the overall architecture and establishing of metadata standards have been funded at a level of \$10 million over the next 5 years by the National Science Foundation (NSF). These and other grass-roots efforts to establish multimission data sets and data products in support of interdisciplinary or cross-cutting approaches should be nurtured, although they may not be the best solution in every case. A challenge for the future will be to develop methods for making complex queries of these federated databases.

Recommendation. NASA should encourage efforts by the scientific community to develop plans for federations of data centers and services that would enable complex querying, mining, and merging of data from different instruments and missions in order to answer complex, large-scale scientific questions.

- **The National Virtual Observatory, an astrophysics project funded recently by the National Science Foundation (NSF), will develop the architecture, standards, and so forth for creating a distributed system of data centers that can be cross-accessed and queried in a transparent manner by users. NASA should coordinate with the NSF-funded work on the NVO, which is predicated on seamless joint access to ground- and space-based data, to ensure that space data are compliant with NVO standards.**
- **NASA should encourage close communications among the groups operating or developing federated systems in order to transfer best practices among its various scientific programs.**
- **The successful implementation of methods for making complex queries of multiple databases is likely to be technically challenging and costly. The level of appropriate investment by NASA in federated data systems should be evaluated at regular intervals and should be based on (1) the importance of the scientific questions that can be addressed through the simultaneous mining of multiple databases, (2) demonstrated scientific return from past investments, and (3) the readiness of computational and communications technology to support data mining.**

EARTH SCIENCES DATA SYSTEM

The earth science community has a particular need to generate and access data within a unified framework that integrates data sets and data centers in a seamless way. The Earth Observing System (EOS) Data and Information System (EOSDIS) Core System (ECS) software was intended to provide “one-stop shopping” access to multidisciplinary data in a timely manner. This goal was not, and probably could not have been, achieved with the technology available at the time the ECS was designed. A restructured ECS with fewer capabilities will be used for a subset of EOS missions, and data processing and distribution for the remainder will be handled by active archives or PI facilities.

NASA recognizes the problems associated with EOSDIS and is developing a strategy for the evolution of the network of data systems and service providers that support the Earth Science Enterprise. The next-generation system is called SEEDS (Strategic Evolution of ESE Data Systems). SEEDS is intended to support all phases of the data management life cycle: (1) acquisition of sensor, ancillary, and ground validation products necessary for processing; (2) processing of data; (3) generation of value-added products via subsetting, format translation, and data mining; (4) archiving and distribution of products; and (5) search, visualization, subsetting, translation, and order services to assist users in identifying, selecting, and acquiring products of interest. Study teams drawn from the user community will be engaged to identify options, define scope, and establish schedule requirements. SEEDS is intended to be managed and implemented as an open and distributed information system architecture under a unifying framework of standards, core interfaces, and levels of service. SEEDS is a work in progress; details about the implementation plan were not available at the time this task group concluded the current report.

Recommendation. The ECS (the EOSDIS Core System) software should be placed in a maintenance mode with no (or very limited) further development until a concrete plan for the follow-on system, SEEDS (Strategic Evolution of ESE Data Systems), has been formulated, its relationship to ECS defined, and the plan reviewed by an external advisory group. This plan should be measured against the lessons learned from EOSDIS and from the experience in other disciplines, and should include provisions for rapid prototyping and an evolutionary and distributed approach to implementing new capabilities, with priorities established by the scientific and other user communities.

USERS OF NASA DATA

NASA currently regards scientists as the end users of data from its missions. While scientists are a major user segment, there are many others, including project and program managers, engineers, educators, the general public, and decision makers. These users need information, rather than data, in order to design and operate missions and to make policy decisions.

Recommendation. NASA planning and project funding should continue to include provisions for the timely generation and synthesis of data into information and the dissemination of this information to the diverse communities of users. This plan should take into account the needs—and the contribution to information generation—of end users, including other federal and state agencies, educational organizations, and commercial enterprises. The plan should include provisions for ongoing assessment of the effectiveness of data transfer and its educational value.

STRATEGIES FOR MEETING THE REQUIREMENTS OF THE RESEARCH COMMUNITY

The task group has identified several elements that appear to be common to those overall data management systems that best meet the requirements of the science communities that they

serve. These elements are listed below and should be included in planning for future missions and facilities:

- Archives and data centers should have (1) scientists on staff with a strong background in the scientific discipline being supported and (2) scientific working groups to help set priorities for acquiring, managing, and discarding data.
- Prelaunch funding should be provided for software development to ensure the timely development of pipelines for processing newly acquired data.
- Multiyear funding should be provided for research, including research using archived data, on the basis of the quality of the proposals received. A recent senior review (the highest level of peer review within the Space Science Enterprise) of extended planetary missions, for example, noted the success of the archival research programs maintained in astrophysics and suggested that these programs might profitably be emulated by the Planetary Data System.
- Guest investigator programs should be established to allow the community to conduct research not planned by the initial project teams.
- Early and open access to data should be provided to permit follow-on proposals to take advantage of new discoveries.
- A mechanism should be established (such as the senior reviews in space science) for making trade-offs among operations of long-lived missions and operations of active archives and data centers in a way that reflects the scientific merit of the range of possible investments.

The importance of managing data and information from NASA's space missions will only continue to grow in the coming years. Maintaining the increasing volumes of data in forms that are readily accessible and that meet the needs of very diverse user communities presents intellectual challenges that are at least the equal of the challenges of building and launching hardware into space. NASA is well positioned to become a leader in developing the techniques and tools for querying and mining large nonproprietary data sets. However, doing so will require a new emphasis on software management; rigorous review of the balance between investments in software and hardware to optimize the science return from both individual missions and suites of missions; and development of new techniques for exploring and intercomparing data contained in a distributed system of active archives, data centers, and data services located both in the United States and abroad.

1

NASA: A Knowledge Agency

INTRODUCTION

The mission of the National Aeronautics and Space Administration (NASA) is—

- To advance and communicate scientific knowledge and understanding of the Earth, the solar system, and the universe.
- To advance human exploration, use, and development of space.
- To research, develop, verify, and transfer advanced aeronautics and space technologies.¹

NASA's program is divided into five strategic enterprises: (1) Aerospace Technology, (2) Biological and Physical Research, (3) Earth Science, (4) Human Exploration and Development of Space, and (5) Space Science.² This report is concerned with the Earth Science and Space Science Enterprises. Both enterprises collect large volumes of data from spaceborne instruments, either to study changes in the oceans, atmosphere, and land surface of the Earth or to explore the universe and search for life beyond the Earth. Managing the data collected from these missions in order to further scientific understanding now and in the future is an enormous challenge.

The Task Group on the Usefulness and Availability of NASA's Space Mission Data was charged by NASA's associate administrators for earth science and space science to (1) evaluate the availability and accessibility of data from earth and space science missions, (2) determine the usefulness of NASA's data collections for supporting scientific studies, and (3) assess whether the balance between attention to mission planning and implementation versus data analysis and utilization is appropriate. (The complete charge is presented in Appendix A.) This report reviews the data systems, services, and strategies for managing earth and space science data collected from space. (The stages in collecting data, from planning a mission to long-term maintenance of data, are described in Appendix B.) Chapter 1 explores the goals of several of the earth and space science disciplines that rely on NASA missions, and it describes how data are used to achieve important science objectives. Chapter 2 describes how the data are currently managed and evaluates the effectiveness of these management strategies. The focus is on the 16 major data facilities and data services that have significant holdings (e.g., at least 1 terabyte) or budgets (e.g., more than \$1 million), or are intended to operate for many years. The major data facilities include active archives, which hold data that are being used intensively for research, and data centers, which maintain data that will continue to be used in the future. (Information asked of

¹National Aeronautics and Space Administration, 2000, *NASA 2000 Strategic Plan*, Washington, D.C., 72 pp.

²See <<http://www.nasa.gov>>.

these facilities in a questionnaire is listed in Appendix C.) The satisfaction of the users, who ultimately judge the success of the system, is discussed in Chapter 3. Chapter 4 then examines some new approaches for increasing the availability and usefulness of earth and space science data, discusses the balance between mission operations and data analysis, and makes some recommendations about how to meet the data challenges of the next decade. Background information, including biographical information on task group members (Appendix D), meeting agendas (Appendix E), and an acronym list (Appendix F) appear at the end of the report.

SPACE SCIENCE ENTERPRISE

The science objectives of NASA's Space Science Enterprise are to "solve mysteries of the universe, explore the solar system, discover planets around other stars, search for life beyond Earth from origins to destiny, chart the evolution of the universe and understand its galaxies, stars, planets, and life."³ The Space Science Enterprise, managed by the Office of Space Science (OSS), is divided into four science themes: (1) origins, which seeks to understand where we come from and whether we are alone; (2) the structure and evolution of the universe; (3) the Sun-Earth connection; and (4) solar system exploration. Examples of the science programs and their interactions with data sets are described below.

Astrophysics: Origins and the Structure and Evolution of the Universe

NASA missions have opened up new windows on the universe, vastly increasing our knowledge about the world around us. Astrophysical sources, collectively, radiate across the spectrum: from gamma rays and X-rays, through the visible and infrared, all the way to microwaves and long-wavelength radio waves. Much of this radiation does not penetrate the Earth's atmosphere and can be studied only from space. NASA's scientific priorities for future missions, developed in coordination with the research community,⁴ include:

- Understand the structure of the universe, from its earliest beginnings to its ultimate fate;
- Explore the ultimate limits of gravity and energy in the universe;
- Learn how galaxies, stars, and planets form, interact, and evolve.⁵

Even modest success in achieving these goals would constitute a spectacular advance in human understanding, and NASA has become an acknowledged leader in this exciting venture. The program seeks to address "the most fundamental questions that science can ask: how the universe began and is changing, what are the past and future of humanity, and whether we are alone. In taking up these questions, researchers and the general public—for we are all seekers in this quest—will draw upon all areas of science and the technical arts."⁶

³National Aeronautics and Space Administration, 2000, *The Space Science Enterprise Strategic Plan*, Washington, D.C., 127 pp.

⁴Review of NASA's Office of Space Science Strategic Plan 2000, letter to Edward J. Weiler, Associate Administrator for NASA's Office of Space Science, National Research Council, Washington, D.C., June 1, 2000.

⁵National Aeronautics and Space Administration, 2000, *The Space Science Enterprise Strategic Plan*, Washington, D.C., 127 pp.

⁶National Aeronautics and Space Administration, 2000, *The Space Science Enterprise Strategic Plan*, Washington, D.C., 127 pp.

The goals outlined above require that data be accessible in a form useful to the science community, that is, calibrated and maintained in accessible data facilities, along with tools for analyzing and visualizing the data. As stated in the 2000 OSS Strategic Plan:

Vast amounts of data are returned from space science missions. The volume, richness and complexity of the data, as well as the need to integrate and correlate data from multiple missions into a larger context for analysis and understanding, present growing opportunities. Exploration and discovery using widely distributed, multi-terabyte databases will challenge all aspects of data management and rely heavily on the most advanced analysis and visualization tools. The design and implementation of the next generation of information systems will depend on close collaboration between space science and computer science and technology.⁷

To achieve its objectives, NASA is flying or plans to fly an ambitious suite of missions (see Table 1.1), with still more to come (e.g., Next Generation Space Telescope, Space Interferometry Mission, and Constellation X). The missions illustrate the diversity of fields that will contribute to the goals of the strategic plan (cosmic rays, nature of high-energy sources, star formation in galaxies, dark matter, cosmology). The diversity of the science and the associated experimental approach lead to a wide range in types of data (time-tagged event logging, multispectral images, and spectroscopy, among others), and each data set and its archive will naturally have different characteristics and requirements.

With the launch of new missions, the volume of astrophysics data will increase substantially, and the demand to compile federated data sets—that is, data sets that can be accessed, intercompared, and queried simultaneously—from different missions will increase. For example, the Galaxy Evolution Explorer (GALEX) is designed to measure the ultraviolet light emitted directly from populations of hot, young stars in galaxies. Some of this ultraviolet light is absorbed by dust grains in interstellar space in the galaxies and is re-emitted as infrared radiation. One of the goals of the Space Infrared Telescope Facility (SIRTF) is to measure that reradiated energy. Thus, a combination of GALEX and SIRTF observations is needed to obtain a comprehensive picture of the cycling of interstellar gas through stars. That information, in turn, is needed to achieve an understanding of how galaxies were formed and how they have evolved. It is clear that the science will require databases that facilitate combining not just GALEX and SIRTF data, but data from other ultraviolet and infrared missions as well as data at other wavelengths.

The Hubble Space Telescope (HST) is one of the most powerful tools ever built for astronomy, and it continues to produce spectacular results. Several generations of instruments on HST will have been deployed during its expected 20-year lifetime. Data are calibrated and held by the Space Telescope Science Institute (see Chapter 2), along with data from several other past and current missions and ground-based surveys. With the accumulation of new observations, research based on mining the HST active archives—often for studies quite different from those originally conceived—has increased at a rapid rate. Archival research now accounts for a substantial fraction of all HST research. Data are now retrieved from the HST active archive at a rate four times higher than that at which new data are put into the archive (see Figure 1.1).

The growing number of data sets from diverse missions makes it possible to tackle important scientific problems in new ways, both by combining measurements from different missions and by taking advantage of the time baselines covered by the data (see Box 1.1).

⁷National Aeronautics and Space Administration, 2000, *The Space Science Enterprise Strategic Plan*, Washington, D.C., p. 90.

TABLE 1.1 Selected U.S.-Led Astrophysics Missions

| Mission | Objective |
|---|--|
| <i>Current Missions</i> | |
| Chandra X-ray Observatory (CXO) | Observes X-rays from high-energy regions of the universe, such as the remnants of exploding stars. |
| Far-Ultraviolet Spectroscopic Explorer (FUSE) | Explores the universe using high-resolution spectroscopy in the far-ultraviolet spectral region. |
| High Energy Transient Explorer 2 (HETE-2) | Detects and localizes gamma-ray bursts. |
| Hubble Space Telescope (HST) | Provides detailed images of celestial objects at high resolution. |
| Microwave Anisotropy Probe (MAP) | Measures the temperature of the cosmic background radiation over the full sky. |
| Submillimeter Wave Astronomy Satellite (SWAS) | Measures the amount of water, molecular oxygen, carbon monoxide, and atomic carbon in interstellar clouds. |
| <i>Upcoming Missions</i> | |
| Advanced Cosmic Ray Composition Experiment for the Space Station (ACCESS) | Study cosmic rays of very high energy to understand elementary particles in our galaxy. |
| Galaxy Evolution Explorer (GALEX) | Measure the ultraviolet light emitted directly from populations of hot, young stars in galaxies. |
| Space Infrared Telescope Facility (SIRTF) | Measure astrophysical phenomena at infrared wavelengths. |
| Swift Gamma Ray Burst Explorer (Swift) | Discover, detect, and study gamma-ray bursts. |

SOURCE: <<http://spacescience.nasa.gov/missions/index.htm>>.

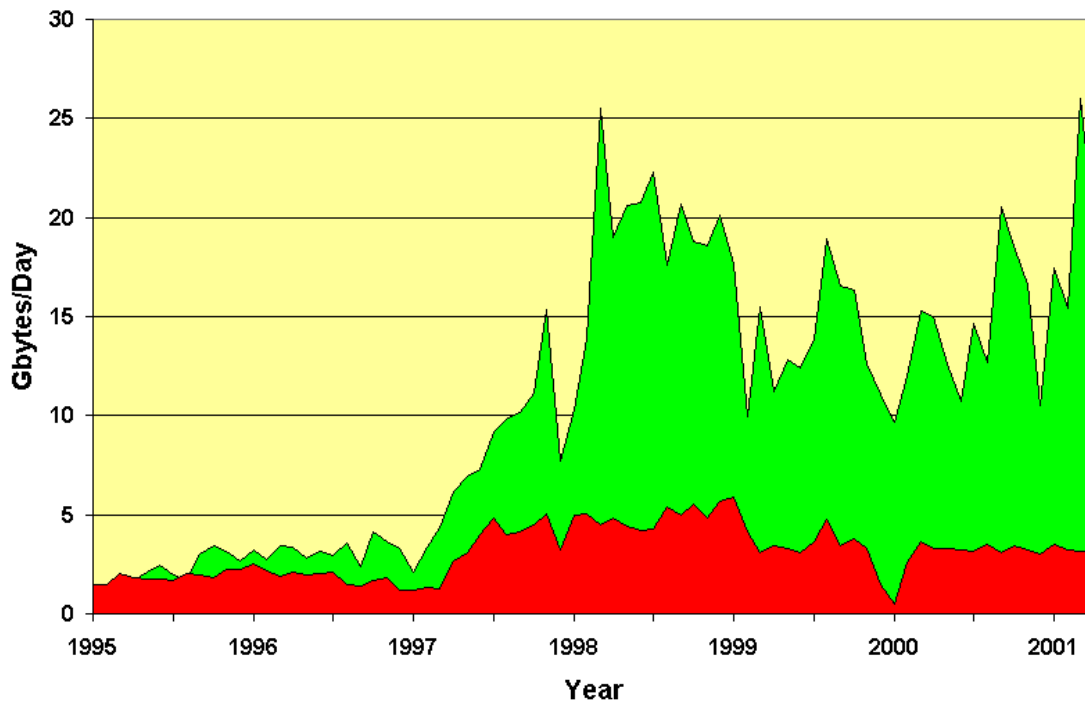


FIGURE 1.1 Data flow into (dark gray) and out of (light gray) the Hubble Space Telescope mission archive, 1995-2001. Note that data flow out of the archive at a rate about four times higher than that of ingest. The increase in this ratio over time is the result of a growth in archival research. If data were used only by the principal investigator, as was true in the first few years after the launch of the HST, the ratio of data retrievals to ingest rate would be close to 1. SOURCE: Ethan Schreier, Space Telescope Science Institute.

BOX 1.1

Importance of Astrophysics Archives

Examples of the role of NASA's astrophysics archives in advancing knowledge include the following:

- The Cosmic Background Explorer flew in 1989-1990 and was successful in detecting large-scale fluctuations in the microwave background radiation. The character of the fluctuations matched theoretical predictions for structure on those scales emerging from the Big Bang, thereby providing a keystone in the field of cosmology.¹ Two of the instruments, the Diffuse Infrared Background Experiment and the Far-Infrared Absolute Spectrometer, collected data that were subsequently mined from the archives for another purpose: to detect infrared light from galaxies at cosmological distances. This measurement demonstrated that substantial amounts of material had undergone nuclear processing inside massive stars and that substantial nucleosynthesis occurred at large redshift—that is, when the universe was very young. Much of the star-forming activity at large redshifts was shrouded behind dense clouds of interstellar dust. Thus important results were derived from archival research using data from an experiment designed for other purposes.

- The first evidence that the expansion of the universe is accelerating was reported in 1998.² The basic observation is that distant supernovas appear dimmer than expected for a uniform rate of expansion. Alternative explanations have been proposed, including the possibility that distant supernovas are dimmed by intervening dust that absorbs all wavelengths equally and that does not betray its existence by making distant objects look redder. In order to rule out this possibility, astronomers searched archives for, and found, the most distant known supernova in the image of longest exposure ever taken by the Hubble Space Telescope. They then found that this same supernova had been observed in several other archived HST images and were able to show that it was twice as bright as it would have been if intergalactic dust or evolutionary effects were responsible for the dimming. This result, which requires that the universe be filled with some kind of mysterious “dark energy,” is probably the most significant cosmological discovery since the detection of the cosmic microwave background radiation.

- A particularly important example of research based on data stored in the active archives is the work stimulated by the observations of the Hubble Deep Fields. Designed to obtain images of the faintest objects observable with HST, long exposures were obtained of two small patches of the sky, one in the Northern Hemisphere and one in the Southern Hemisphere. Some of the galaxies seen in these images are at a distance of 12 billion light years; they are being seen as they were when the universe was only about 10 percent of its current age. These data allow astronomers to probe the characteristics of galaxies when they were just coming into existence. The observations were made available to the community as soon as they were reduced, with no proprietary period. Additional observations have now been obtained, either by spectroscopy or measurements at other wavelengths, by every major observatory in the world, both in space (e.g., by the Chandra X-ray Observatory, X-ray Multi-Mirror Mission, and Infrared Space Observatory) and on the ground (e.g., Wm. Keck Observatory, Very Large Array, and James Clerk Maxwell Telescope), and more than 200 follow-on papers have been published.

¹C.L. Bennett et al., 1996, Four-year COBE DMR cosmic microwave background observations: Maps and basic results, *Astrophysical Journal Letters*, v. 464, p. L1, and references therein.

²A.G. Riess, 1998, Observational evidence for supernovae for an accelerating universe and a cosmological constant, *Astronomical Journal*, v. 116, p. 1009; S. Perlmutter, Measurements of omega and lambda from 42 high-redshift supernovae, *Astrophysical Journal*, v. 517, p. 565.

The Sun-Earth Connection

The “Sun-Earth Connection” is the name given to a broad NASA program that includes studies of the Sun, the processes that link the Sun to the Earth, and the space environments and upper atmospheres of other solar system bodies. Another area of study characterizes the properties of the solar wind as it moves through the solar system. The overall goal of the program is to understand how and why the Sun varies and how the Earth and other planets respond to those variations.

The Sun’s energy output varies on timescales from seconds to billions of years. This energy reaches the Earth in two forms: as electromagnetic radiation and charged atomic particles. The Earth responds to the Sun’s varying energy inputs in a number of ways. Growing evidence indicates that even small variations in the total energy emitted by the Sun can alter circulation in the Earth’s atmosphere and hence affect climate (e.g., the Maunder minimum in solar activity, which is associated with a little ice age in Europe in the 17th century). Ejections of mass from the corona, which are more frequent near the peak of the solar cycle, cause auroras and disturb the Earth’s ionosphere in such a way as to disrupt communications, disable power grids, and damage satellites and alter their orbits.

In order to explore the effect of the Sun on the Earth, NASA is developing a series of missions that will characterize the solar energy output and the mechanisms that control it; explore the Earth’s space environment; compare the space environment of the Earth with that of other planets; and assess the impact of space weather on humanity. Many of these investigations will require access to archived data (see Box 1.2). A sampling of solar physics missions is listed in Table 1.2.

Solar System Exploration

NASA’s planetary exploration program is focused on answering fundamental questions about how planets form, why they are different from one another, and what conditions lead to the development of life. The last half of the 20th century was an extraordinary age of exploration and discovery. All of the planets in our solar system except Pluto have now been visited by NASA spacecraft. Each has been transformed from a remote astronomical object into a unique world, clearly distinct from all of the other objects in the solar system. Comparative planetology can provide real clues as to how the Earth itself and its habitability will be affected by changes in the total energy output of the Sun, climate change, increasing abundance of greenhouse gases, asteroid impacts, and so on.

Planetary research has been one of the primary beneficiaries of the recent change in NASA philosophy to support a diverse set of missions of moderate scale. Flight opportunities have become more frequent; several comets and asteroids, in addition to the major planets, have now been visited and characterized; and the advent of modern detectors has greatly increased the volume of data from each mission. Table 1.3 presents a sampling of planetary missions.

BOX 1.2
Importance of Archives for Solar and Space Physics

Following are examples of the role of NASA archives in advancing solar and space physics:

- The Solar and Heliospheric Observatory (SOHO) mission archive has provided nearly continuous data since 1996. This uniquely consistent archive of solar data has been mined to produce several new insights into solar phenomena. One example is the realization that coronal mass ejections (CMEs)—eruptions of gas that disrupt the flow of the solar wind and produce disturbances that strike the Earth causing electrical power outages, damaging communications satellites, and producing auroral displays—involve an unexpectedly large portion of the solar surface and that several spatially separate regions participate in the process. This observation implies that CMEs are the result of large-scale reorganization of the solar magnetic field, rather than localized events. Another example is the discovery of a subsurface flow of plasma toward the solar equator that exists only at the north pole and that advances and retreats as the solar activity cycle evolves. This is the first time that a flow asymmetric with respect to the equator has been discovered, and it may hold a key to the reversal in polarity of successive solar cycles. However, the flow has been observed so far for only a quarter of a single solar cycle and needs to be followed over multiple cycles before its role in solar activity can be characterized. In both of these examples, multiple archival data sets from different instruments were combined to clarify the nature of the phenomenon.
- Data from the Transition Region and Coronal Explorer (TRACE) mission have revealed that the solar atmosphere is threaded by an enormous number of very thin channels of high heat conductivity created by the solar magnetic field. These channels may be the key to unlocking the long-standing problem of coronal heating in solar and stellar physics. TRACE data are currently being used in new calculations of coronal thermodynamics. Such discoveries and applications are fostered by the availability of this continuous data record immediately and without restriction from the archive.
- By combining X-ray images, infrared images, and particle detector data, solar researchers have discovered that high-speed solar wind streams emanate primarily from the boundaries of coronal holes. Monitoring the evolving positions of coronal holes permits researchers to estimate when a solar wind gust will hit the Earth and disrupt telecommunications. The prediction of space weather as well as the exploration of the solar activity cycle is facilitated by the Solar Data Analysis Center, which provides “one-stop shopping” of three decades of data from past NASA solar missions and several major ground-based observatories, along with the tools to analyze them.

SOURCE: Frank Hill, National Solar Observatory.

TABLE 1.2 Selected Solar and Space Physics Missions

| Mission | Objective |
|---|--|
| <i>Current Missions</i> | |
| Advanced Composition Explorer (ACE) | Samples low-energy particles of solar origin and high-energy galactic particles, and provides near-real-time solar wind information. |
| Fast Auroral Snapshot Explorer (FAST) | Probes the physical processes that produce auroras. |
| Genesis | Collects particles of the solar wind and returns them to Earth. |
| Imager for Magnetopause-to-Aurora Global Exploration (IMAGE) | Produces the first comprehensive global images of the plasma populations in the inner magnetosphere. |
| Interplanetary Monitoring Platform 8 (IMP-8) | Measures the magnetic fields, plasmas, and energetic charged particles of the Earth's magnetotail and magnetosheath and of the near-Earth solar wind. |
| International Solar Terrestrial Physics Global Geospace Science Program Polar (Polar) | Images the aurora and measures the fluxes of charged particles and ions, magnetic and electric fields, and electromagnetic waves over the poles. |
| Solar Anomalous and Magnetospheric Particle Explorer (SAMPEX) | Studies the composition of local interstellar matter and solar material and the transport of magnetospheric charged particles into the Earth's atmosphere. |
| Solar and Heliospheric Observatory (SOHO) | Studies the internal structure of the Sun, its outer atmosphere, and the origin of the solar wind. |
| Stardust | Collects dust from a comet's nucleus. |
| Transition Region and Coronal Explorer (TRACE) | Images the solar corona and transition region. |
| Ulysses | Explores interplanetary space at high solar latitudes. |
| Voyager Interstellar Mission (VIM) | Searches for the heliopause boundary, the outer limits of the Sun's magnetic field, and the outward flow of the solar wind. |
| International Solar Terrestrial Physics Global Geospace Science Program Wind (Wind) | Samples the upstream interplanetary medium, a principal region of geospace where energy and momentum are transported and stored. |
| <i>Future Mission</i> | |
| Two Wide-angle Imaging Neutral-atom Spectrometers (TWINS) | Provide a new capability for stereoscopically imaging the magnetosphere. |

SOURCE: <<http://spacescience.nasa.gov/missions/index.htm>>.

TABLE 1.3 Selected U.S.-led Planetary Missions

| Mission | Objective |
|------------------------------|--|
| <i>Current Missions</i> | |
| Cassini | Makes observations of Jupiter and its moons (atmospheric dynamics and composition, Jupiter's magnetic environment, the interactions between Jupiter and its moons) on its way to Saturn. |
| Galileo | Studies Jupiter and its moons in more detail than any previous spacecraft. |
| Mars Global Surveyor (MGS) | Measures surface features, atmosphere, and magnetic properties of Mars. |
| 2001 Mars Odyssey | Maps the amount and distribution of chemical elements and minerals that make up the Martian surface. |
| <i>Future Missions</i> | |
| Mars Exploration Rover | Analyze rocks and soils for evidence of liquid water that may have been present in Mars's past. |
| Mars Express | Explore the atmosphere, structure, and geology of Mars to search for subsurface water from orbit and deliver a lander to the Martian surface. |
| Comet Nucleus Tour (CONTOUR) | Image two comet nuclei, and collect and analyze dust to reveal the comet's composition. |

SOURCE: <<http://spacescience.nasa.gov/missions/index.htm>>.

EARTH SCIENCE ENTERPRISE

Characterize, understand, and predict—these are the themes of NASA's Earth Science Enterprise (ESE). The goal is “to develop a scientific understanding of the Earth system and its response to natural or human-induced changes to enable improved prediction capability for climate, weather, and natural hazards.”⁸ The research program is organized around a set of scientific questions aimed at understanding how the Earth is changing and the consequences of those changes for life on Earth.⁹ Some of the questions being addressed by the ESE program are as follows:

- How is the global Earth system changing?
- What are the primary causes of change in the Earth system?
- How does the Earth system respond to natural and human-induced changes?
- What are the consequences of change in the Earth system for human civilization?
- How well can we predict future changes in the Earth system?¹⁰

⁸National Aeronautics and Space Administration, 2000, Exploring Our Home Planet: The Earth Science Enterprise Strategic Plan, May 25, 2000, draft.

⁹National Aeronautics and Space Administration, 2000, Understanding Earth System Change: NASA's Earth Science Enterprise Research Strategy for 2000-2010, January 2001, 46 pp.; National Research Council, 2000, *Review of NASA's Earth Science Enterprise Research Strategy for 2000-2010*, National Research Council, Washington, D.C., 33 pp.

¹⁰National Aeronautics and Space Administration, 2000, Understanding Earth System Change: NASA's Earth Science Enterprise Research Strategy for 2000-2010, January 2001, 46 pp.

In order to answer these questions, the Earth Science Enterprise is currently conducting research in the following areas:

- Oceans and ice in the climate system;
- Biology and biogeochemistry of ecosystems and the global carbon cycle;
- Atmospheric chemistry, aerosols, and solar radiation;
- Global water cycle; and
- Solid Earth science.

These research topics also address major subproblems of the U.S. Global Change Research Program¹¹ to which a space-based observational system is uniquely capable of making a significant contribution.¹² Current and upcoming ESE missions are listed in Table 1.4.

Space-based data collected by the ESE address three classes of problems: (1) characterization of physical and biological processes, (2) monitoring status and changes, and (3) analysis of feedback mechanisms. “Characterizing and understanding a process” involves measurements to examine a specific process that operates in the Earth system, with the aim of developing physical models and model parameterizations. An example of this type of mission is the Tropical Rainfall Measurement Mission (TRMM),¹³ which measures the spatial and temporal variations in the tropical region (-35° to 35° latitude). The goals of this three-year mission are to study the frequency distributions of rainfall intensity and areal coverage and to relate the timing of heaviest rainfall to such factors as the nocturnal intensification of large mesoscale convective systems over the oceans and the diurnal intensification of orographically and sea-breeze-forced systems over land. TRMM data will potentially improve estimates of latent heating,¹⁴ which in turn will improve the prediction of rainfall events from global climate models. Recent results from TRMM, for example, show that windblown desert dust can choke rain clouds, cutting rainfall hundreds of miles away.¹⁵

Many of the instruments developed by the NASA ESE are used for systematic monitoring. An example of this class of instrument is the Total Ozone Mapping Spectrometer (TOMS).¹⁶ This class of instrument has been flown in four spacecraft with data extending back to November 1978 and has been used to monitor the amount of stratospheric ozone. A major result from the use of this and other instruments was the discovery of the growth of the Southern Hemisphere Ozone Hole,¹⁷ which led to the nearly worldwide phasing out of the use of the chlorofluoro-

¹¹The U.S. Global Change Research Program was established in 1989 to develop and coordinate a research program to understand, assess, predict, and respond to natural and human-induced global change. Nine federal agencies, including NASA, and the Executive Offices of the President participate in the program. See Subcommittee on Global Change Research, *Our Changing Planet, The FY2002 U.S. Global Change Research Program*, Washington, D.C., 74 pp.

¹²National Research Council, 2000, *Review of NASA's Earth Science Enterprise Research Strategy for 2000-2010*, National Research Council, Washington, D.C., 33 pp.

¹³See <<http://trmm.gsfc.nasa.gov>>.

¹⁴Energy from the Sun is stored in the form of water vapor. Condensation of water vapor in clouds releases this latent heat, causing the atmosphere to warm locally.

¹⁵See <<http://www.gsfc.nasa.gov/gsfc/earth/dust/rainfall.htm>>.

¹⁶<<http://toms.gsfc.nasa.gov/>>.

¹⁷The loss of ozone was first detected by the British Antarctic Survey, which was monitoring the atmosphere using a network of ground-based instruments. The TOMS data confirmed that the ozone loss was real and that it extended over most of the Antarctic continent. See G. Carver, 1988, *The ozone hole tour, Part 1. The history behind the ozone hole*, University of Cambridge, <<http://www.atm.ch.cam.ac.uk/tour/part1.html>>.

carbons (CFCs).¹⁸ The history of the total ozone measurements is composed of results from multiple instruments flown on different spacecraft; consequently, calibration between the results from these different instruments is critical to understanding the long-term evolution of total ozone. More importantly, new data are needed to determine if the mitigation steps (i.e., reducing the amount of CFCs released into the atmosphere) are effective. By their nature, long-term monitoring programs need to be able to relate measurements from different instruments, and the original data need to be available so that improved calibration and reduction algorithms can be applied.

The final category of problem—analysis of feedback mechanisms—is the most challenging for any data system, because understanding cause and effect requires comparison of different data sets collected from different satellites with different types of instruments. One of the fundamental questions to be addressed in this class of problem is the role that clouds play in relation to the effects of increasing carbon dioxide and other greenhouse gases. Clouds both reflect sunlight (which cools the Earth) and trap heat in the same way as greenhouse gases (thus warming the Earth). Different types of clouds do more of one than the other. The net effect of clouds on climate change depends on which cloud types change, and whether they become more or less abundant, thicker or thinner, and higher or lower in altitude.¹⁹ Different instruments measure different characteristics of the clouds, and determination of the full impact of clouds requires that these measurements be merged. Understanding of the evolution of cloud-type cover, how cloud types are being affected by climate change, and how they in turn affect climate change requires access to long-time histories of space-based and ground data and the ability to apply new algorithms to original data to extract data relevant to cloud types. This type of synthesis of results is the most challenging for any data system, but it is also the area where the most significant results from the NASA ESE are likely to come.

Many of the important science questions being addressed by NASA investigators require long-term, continuous measurements to detect and monitor environmental change. Consequently, data centers providing accessible, usable long-term data are essential in the earth sciences (see Box 1.3).

¹⁸The breakdown of ozone by CFCs in the presence of high-frequency UV light was demonstrated in 1974 (M.J. Molina and F.S. Rowland, 1974, Stratospheric sink for chlorofluoromethanes: chlorine atom catalyzed destruction of ozone, *Nature*, v. 249, p. 810-812). International negotiations to reduce CFC levels began in 1983, but the Montreal Protocol was not signed until 1987, after the existence of the Antarctic ozone hole was confirmed and linked to CFCs.

¹⁹See <http://www.giss.nasa.gov/research/intro/delgenio_03>.

TABLE 1.4 Selected ESE Missions

| Mission | Objective |
|---|---|
| <i>Current Missions</i> | |
| Active Cavity Radiometer Irradiance Monitor III (ACRIM III) | Measures total solar irradiance from the Sun. |
| Earth Radiation Budget Satellite (ERBS) | Investigates how energy from the Sun is absorbed and re-emitted by the Earth, and determines the effects of human activities on the Earth's radiation balance. |
| Landsat 7 | Provides multispectral, moderate-resolution digital images of the Earth's continental and coastal areas, with global coverage on a seasonal basis. |
| SeaSTAR | Measures bio-optical properties of the global ocean. |
| Terra | Provides global data on the state of the atmosphere, land, and oceans, as well as their interactions with solar radiation and with one another. |
| Total Ozone Mapping Spectrometer Earth Probe (TOMS-EP) | Provides daily global measurements of the total column ozone. |
| Tropical Rainfall Measurement Mission (TRMM) | Monitors tropical rainfall and the associated release of energy that helps to power global atmospheric circulation. |
| Quick Scatterometer (QuikSCAT) | Records sea-surface wind speed and direction for global climate research, weather forecasting, and storm warning. |
| <i>Future Missions</i> | |
| Advanced Earth Observing Satellite (ADEOS)-II | Measure near-surface wind velocity under all weather and cloud conditions over the Earth's oceans. |
| Aqua | Measure clouds, precipitation, atmospheric temperature and moisture content, terrestrial snow, sea ice, and sea-surface temperature. |
| Ice, Clouds, and Land Elevation Satellite (ICESat) | Determine decadal variation of ice sheet thickness over Greenland and Antarctica, altitude and thickness of clouds, vegetation heights, land topography, and ocean surface and sea ice altimetry. |
| Meteor | Monitor the global distribution of aerosols, ozone, and other trace gases in the Earth's atmosphere. |
| Solar Radiation and Climate Experiment (SORCE) | Provide total irradiance measurements (ultraviolet, extreme ultraviolet, and the visible to near infrared) required by climate studies. |

SOURCE: <http://gaia.hq.nasa.gov/ease_missions/default.cfm?transaction=Enter_ESE_Missions>.

BOX 1.3

Applications of Earth Science Archives

Archived data have proven to be extremely important for investigations of changes in the Earth's atmosphere, oceans, and land cover. Examples include:

- Global ocean surface topography has been measured by Ocean Topography Experiment (TOPEX)/Poseidon, a joint NASA-French Space Agency mission, since October 1992. The unprecedented accuracy (2 cm) and precision (4 mm) of the data allowed sea-level change in the Pacific to be monitored and predicted during the large 1997-1998 El Niño event.¹ El Niño events disrupt the ocean-atmosphere system in the tropical Pacific, with global consequences for weather and climate. Analysis of TOPEX data has also augmented coastal tide gauge records, revealing a long-term global mean sea-level rise of 3.2 mm/yr, which can be completely explained by the thermal expansion of seawater. The relationship of the Pacific Decadal Oscillation to El Niño events or its effect on fisheries, coral bleaching, or ocean eddies has yet to be determined, owing to the relative shortness of the TOPEX data record. Such questions may be addressed as the current Jason-1 mission extends the record of sea-surface height another decade.
- Establishing the magnitude and causes of greenhouse warming requires access to accurate data over as long a period as possible. Harries and others recently used satellite interferometry data from NASA and Japan to compare the outgoing long-wave radiation spectra from the Earth in 1970 and 1997.² They showed experimental evidence of “a significant increase in the Earth's greenhouse effect that is consistent with concerns over radiative forcing of climate.” Their investigations were hampered by the poor quality of older data tapes, which had deteriorated over time.³ Considerable effort was required to rescue the data and make them usable, illustrating the importance of routine migration of data to new media and working with archived data to ensure their long-term scientific value.

¹C. Cabanes, A. Cazenave, and C. Le Provost, 2001, Sea level changes from TOPEX/Poseidon altimetry for 1993-1999, and warming of the southern oceans, *Geophysical Research Letters*, v. 28, p. 9-12.

²J.E. Harries, H.E. Brindley, P.J. Sagoo, and R.J. Bantges, 2001, Increases in greenhouse forcing inferred from the outgoing longwave radiation spectra of the Earth in 1970 and 1997, *Nature*, v. 410, p. 355-357.

³Richard Goody, Professor Emeritus, Harvard University, personal communication to J. Purdom, fall 2001.

THE CHANGING PARADIGM FOR NASA

With the adoption of the scientific goals of the Earth Science and Space Science Enterprises, NASA can no longer be viewed primarily as a technology-demonstration agency. Instead, NASA has defined itself as a knowledge-generating agency, with missions at the front end of the information pipeline. NASA data are a national resource; the stewardship and exploitation of NASA data are necessarily a *national* responsibility. The care of the data, including the tasks of archiving and distribution, must be accomplished so as to maximize knowledge enhancement, scientific impact, and discovery potential. The chapters following describe and evaluate the strategies adopted by NASA to date and make recommendations to enhance the usefulness and accessibility of the growing databases obtained from NASA missions.

2

Accessibility of Data: The Architecture of the Archives

Over the last two decades, major changes have taken place in the way that NASA's data are archived and distributed. These changes have resulted in more data being more accessible more rapidly to a larger number of users. Prior to the 1980s, most data were processed and interpreted by principal investigators (PIs), working either individually or as teams. Mailing data tapes to the PIs was slow, and data were sometimes lost because instrument failures were not discovered in a timely manner.¹ Other data were lost because PIs had strong incentives to publish, but fewer incentives to archive and distribute the data or to send properly documented data to established archives. Even if data were archived, they were not always in convenient formats or on usable media.² The primary facility for storing and maintaining data was the National Space Science Data Center (NSSDC), which had been operating since 1966.

The 1980s saw the introduction of data systems that would process and archive data centrally and provide a variety of services. Nevertheless, a 1982 National Research Council (NRC) report found that "the distribution, storage, and communication of data currently limit the efficient extraction of scientific results from space missions."³ These problems were expected to worsen as data volumes continued to grow exponentially. A 1985 NRC report recommended the establishment of a network of geographically distributed data centers and active archives for dealing with the data.⁴ Data that require long-term maintenance because of the likelihood of future use would be held in data centers, and data being used intensely in research would be held in active archives. NASA adopted the idea and established 10 active archives by the early 1990s. Today, there are 16 major data archives, data centers, and services (see Table 2.1), which disseminate most of the data from the Earth Science and Space Science Enterprises.⁵

¹National Research Council, 1982, *Data Management and Computation. Volume 1: Issues and Recommendations*, National Academy Press, Washington, D.C., 167 pp.

²For example, only paper records of the Viking data were sent to NSSDC.

³National Research Council, 1982, *Data Management and Computation. Volume 1: Issues and Recommendations*, National Academy Press, Washington, D.C., p. 5.

⁴National Research Council, 1985, *Issues and Recommendations Associated with Distributed Computation and Data Management Systems for the Space Sciences*, National Academy Press, Washington, D.C., 111 pp.

⁵A number of the active archives in existence today have their roots in the systems developed in the 1970s or 1980s. For example, the Goddard Space Flight Center Distributed Active Archive Center (DAAC) was created from the NASA Climate Data System and Pilot Land Data System, and the physical oceanography DAAC (PO.DAAC) was created from the NASA Ocean Data System. On the space science side, the Astronomical Data Center grew out of a stellar data center operating in Strasbourg, France, and the Solar Data Analysis Center grew out of the Solar Maximum Mission Data Analysis Center.

TABLE 2.1 Earth and Space Science Archives and Data Centers

| Facility | Year Established | Host Institution | Scientific Specialty |
|----------------------|---------------------|--|--|
| <i>Earth Science</i> | | | |
| ASF DAAC | 1990 | Alaska SAR Facility, University of Alaska | Sea ice, polar processes |
| EDC DAAC | 1992 | EROS Data Center, U.S. Geological Survey | Land processes |
| GSFC DAAC | 1993 | Goddard Space Flight Center, NASA | Upper atmosphere, atmospheric dynamics, global biosphere, hydrologic processes |
| LaRC DAAC | 1989 | Langley Research Center, NASA | Radiation budget, aerosols, tropospheric chemistry |
| NSIDC DAAC | 1991 | National Snow and Ice Data Center, University of Colorado | Snow and ice, cryosphere |
| ORNL DAAC | 1993 | Oak Ridge National Laboratory, U.S. Department of Energy | biogeochemical fluxes and processes |
| PO.DAAC | 1991 | Jet Propulsion Laboratory, NASA-Caltech | Ocean circulation, air-sea interaction |
| SEDAC | 1994 | CIESIN, Columbia University | Socioeconomic data and applications |
| <i>Space Science</i> | | | |
| ADC | 1977 | Goddard Space Flight Center, NASA | Astronomy, astrophysics, photometry, spectroscopy |
| HEASARC | 1990 | Laboratory for High-Energy Astrophysics, Goddard Space Flight Center, NASA | High-energy astrophysics |
| IRSA | 1999 | Infrared Processing and Analysis Center, CalTech | Infrared science |
| MAST | 1997 | Multi-mission Archive, Space Telescope Science Institute | Optical/UV science |
| NED | 1989 | Infrared Processing and Analysis Center, CalTech | Extragalactic astronomy and cosmology |
| NSSDC | 1966 | Office of the Space Science Directorate, Goddard Space Flight Center, NASA | Space physics data and long-term maintenance of all space science data |
| PDS | 1991 | Jet Propulsion Laboratory, NASA-Caltech | Planetary and space science |
| SDAC | 1991 | Goddard Space Flight Center, NASA | Solar and heliospheric physics |

NOTE: ADC = Astronomical Data Center; ASF = Alaska Synthetic Aperture Radar Facility; CIESIN = Consortium for International Earth Science Information Networks; DAAC = Distributed Active Archive Center; EDC = EROS Data Center; EROS = Earth Resources Observations Systems; GSFC = Goddard Space Flight Center; HEASARC = High Energy Astrophysics Science Archive Research Center; IRSA = Infrared Science Archive; LaRC = Langley Research Center; MAST = Multi-mission Archive at Space Telescope; NED = NASA/IPAC Extragalactic Database; NSIDC = National Snow and Ice Data Center; NSSDC = National Space Science Data Center; ORNL = Oak Ridge National Laboratory; PDS = Planetary Data System; PO.DAAC = Physical Oceanography DAAC; SAR = synthetic aperture radar; SDAC = Solar Data Analysis Center; SEDAC = Socioeconomic Data and Application Center.

Data have never been as plentiful as they are now. The widespread availability of desktop computing and the ability to transfer data via the Internet have made a wide range of data quickly and easily accessible to all. Both the Earth Science and Space Science Enterprises have policies of full and open access (i.e., data are available without restriction, for no more than the cost of filling a user request), which encourages data use by the broader community.⁶ Proprietary periods differ by discipline, but in all cases, data are to be made available to the broader community within two years. This policy encourages rapid data processing and publication. Finally, plans for documenting and archiving data are now required of every mission. Of course, compliance with these policies varies, and some data systems and services operate more effectively than others do.

This chapter summarizes strategies for making data available in several earth and space science disciplines and identifies the approaches that appear to be most effective. The space science active archives are discipline-specific and operate independently of one another, using standards and formats developed for their specific holdings. The earth science active archives are also discipline-specific, but they use common standards and formats to permit data from multiple centers to be located and integrated. Such integration is essential for studying complex environmental processes. Space science research problems have traditionally not required the integration of data from multiple centers. However, as described in Chapter 4, this is starting to change in some disciplines.

SPACE SCIENCE DATA SYSTEMS

NASA has supported the creation of a number of data centers for astrophysics, planetary science, and solar science. Information about the active archives, data centers, and data services is summarized in Table 2.2. There is wide disparity in budgets, but it is not the size of the holdings that determines the costs of operating a data center. Instead, cost drivers include (1) the complexity of the holdings and the number of unique data sets that must be acquired, quality controlled, and maintained, with planetary science being a prime example of a discipline that collects very different types of data; (2) the demand for user services compared with automated data delivery; (3) the need to repackage the data in formats suitable for particular types of research; (4) the investment in user interfaces, visualization programs, and querying tools; and (5) the overhead imposed by the host institution.

Astrophysics Data Systems

NASA has supported the successful development of an end-to-end system for managing and distributing astrophysics data. The overall architecture of the astrophysics data system is shown in Figure 2.1. Each mission has an associated science center or data facility, which is responsible for the acquisition, characterization, and documentation of the data. In a few cases, the PI may be responsible for data processing. After the initial proprietary period, if there is one, the data are

⁶For example, see NASA Earth Science Enterprise Statement on Data Management, April 1999, <<http://globalchange.gov/policies/agency/nasa.html>>.

TABLE 2.2 Characteristics of Space Science Data Facilities and Services

| Center | Number of Users ^a | <u>Budget (\$M)</u> | | <u>Holdings (TB)</u> | | Number of Staff |
|----------------------|---------------------------------|---------------------|---------|----------------------|---------|--------------------|
| | | FY 2000 | FY 2005 | FY 2000 | FY 2005 | |
| <i>Data Facility</i> | | | | | | |
| HEASARC | 8,887 | 1.5 | 1.8 | 2 | 6 | 14 |
| IRSA | 4,022 | 1.2 | 1.3 | 18 | 23 | 2.5 |
| MAST | 3,300 | 0.6 | 1.1 | 11 | 13 | 4.6 |
| NSSDC | Unknown | 5.9 | 5.9 | 20 | 35 | 58 |
| PDS | 6,000 | 4.8 | 6.1 | 1 | 76 | 27 |
| SDAC | Unknown | 0.6 | 1.1 | 3 | 5-15 | 2.5 |
| <i>Data Service</i> | | | | | | |
| ADC | 59,418 | 0.6 | 0.6 | 18 GB | 23 GB | 5 |
| NED | 18,382 | 0.9 | 1.5 | 0.2 | 2 | 11 |

^a Unique users who received data in FY 2000. "Unknown" indicates that the facility counts only the number of Web site hits.

NOTE: Budgets and holdings for FY 2005 are estimated.

SOURCE: Managers of the data facilities and services (see questionnaire in Appendix C).

placed in an active archive, from which they can be downloaded by the community. In some cases, the active archives are maintained by the mission-specific science center. In other cases, they are transferred to one of the wavelength-oriented centers: the Multi-mission Archive at Space Telescope (MAST) for optical/ultraviolet data, the Infrared Science Archive (IRSA), or the High Energy Astrophysics Science Archive Research Center (HEASARC). These centers take advantage of the economies of scale associated with providing a common archive and distribution infrastructure, and they maintain staff who are sufficiently knowledgeable about the data to assist community users. Standard algorithms are developed and made available to the community for performing functions such as extracting sources from images and classifying sources to determine whether they are stars or galaxies. Algorithm development is science-driven, with priorities determined by the astrophysics community. Long-term maintenance of the data is the responsibility of the NSSDC.

The standard policy is to make all data openly available; for some facilities there is an initial, usually brief, proprietary period. For example, Hubble Space Telescope (HST) data become available one year after they are obtained for the investigator who proposed the specific observations. Support for calibration, documentation, archiving, and distribution makes the policy effective; HST data are used extensively by scientists other than those who submitted the original observing proposals.

The Space Infrared Telescope Facility (SIRTF) Legacy Science program illustrates another approach to providing early and open access to data. SIRTF is a cryogenically cooled telescope with a finite lifetime, probably about three years. The usual sequence followed by an observing program (i.e., submit a proposal, observe, analyze, publish, interpret, and then submit a new proposal based on what was learned) is too long for a short mission, especially one that offers orders-of-magnitude gains in sensitivity and that will undoubtedly discover unexpected phenomena. The Legacy Science program will move data into the public domain immediately in order to guide subsequent proposals from the community.⁷ Funding is being made available to

⁷The six science teams supported by the Legacy Science program were chosen through peer review. A description of the projects is given at <http://sirtf.jpl.nasa.gov/SSC/A_GenInfo/SSC_A1_Legacy_Selection.html>.

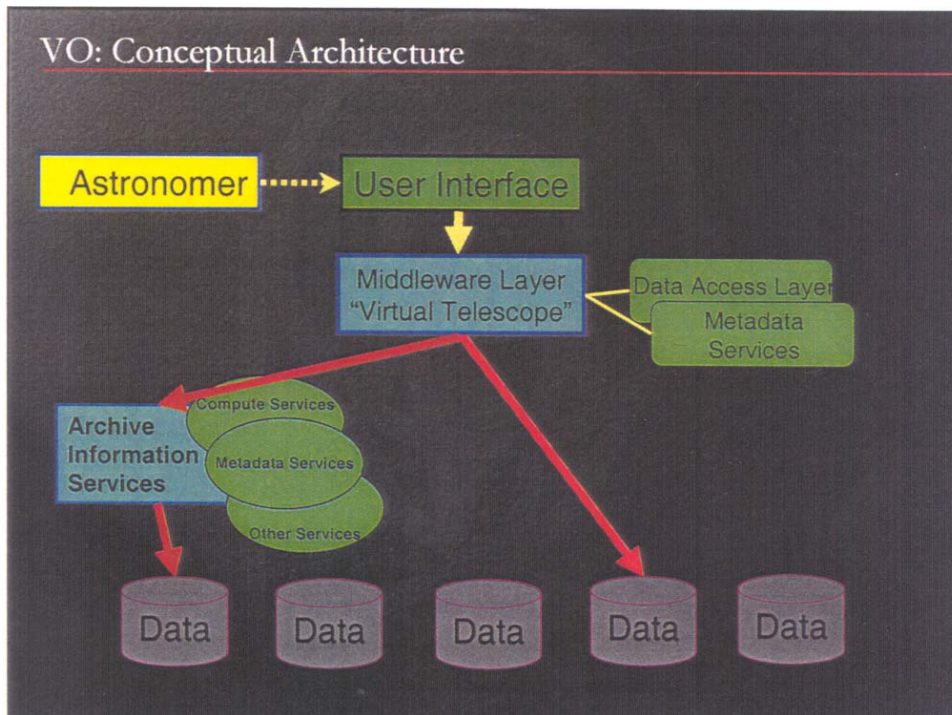


FIGURE 2.1 The architecture of the astrophysics data centers. Data are initially calibrated and stored at mission-specific centers, then transferred to centers organized by wavelength—IRSA for infrared data, MAST for optical/ultraviolet data, and HEASARC for high-energy data. These wavelength centers maintain most of the data online so that they can be readily accessed by the user community. The NSSDC provides long-term maintenance and backup storage of data. In addition, a number of services facilitate access to data. NED, for example, makes it possible to locate data on individual galaxies; SIMBAD (Set of Identifications, Measurements, and Bibliography for Astronomical Data) performs a similar service for stellar data; and the ADS (Astrophysics Data System) provides online access to most of the astronomical literature.

SOURCE: Ethan Schreier, Space Telescope Science Institute.

six science teams prior to launch in order to support the planning of large, coherent SIRTf investigations that will provide data of general and lasting importance to the astronomy community. These science teams will also collect ancillary data if they are required, and will develop postpipeline processing algorithms and software in time to be applied as soon as the SIRTf data become available.

In addition to collecting data, astronomers have used NASA support to develop a number of integrating services that facilitate research. The Astronomical Data Center (ADC) provides Internet access to bibliographic information and abstracts for most of the published papers in space science and full articles from many journals. For an astronomer looking for relevant material in the published literature, a computer terminal, not a library, is likely to be the first stop. The NASA/Infrared Processing and Analysis Center Extragalactic Database (NED) provides online access to information on galaxies, quasars, and extragalactic radio, X-ray, and infrared sources. The database contains positions, redshifts, photometry, images, other basic data, and associated physical quantities as well as a comprehensive catalog of the published literature. NED has become, according to the most recent senior review (see Box 2.1), “an irreplaceable tool for observational and archival extragalactic research.”⁸

The active archives maintained by MAST, IRSA, and HEASARC are seeing heavy and growing use for research (see Figure 1.1 in Chapter 1). In the case of long-lived missions such as HST, grant support is available for research that makes use of data stored in the active archives. Awards for both new observations and for use of older data are made through peer review.

NASA’s Astrophysics Senior Review panel, which met in June 2000, found that the astrophysics active archives and data services are generally serving the community well.⁹ However, the panel recommended that greater attention be paid to increasing interoperability of data sets and active archives. Plans for creating such a system are described in Chapter 4.

Planetary Data Systems

Planetary science receives its data from ground-based telescopes, Earth-orbiting telescopes, and space missions to solar system objects. In addition, some complex and expensive modeling studies are viewed as community resources, and the data from these calculations are made available to the wider community. The structure and character of data from these sources varies greatly. Data from ground-based telescopes are under the control of the PI, who is responsible for data reduction, analysis, interpretation, and dissemination of results. No guidelines exist for making these data available to a wider community. On occasion they are placed in the planetary active archive (i.e., the Planetary Data System, described below), but this is the exception. Data from Earth-orbiting observatories make use of the same facilities as for astrophysics data and are handled in the same way. Data from planetary missions are handled in a variety of ways.

⁸National Aeronautics and Space Administration, 2000, Report of the Senior Review of Origins and Structure and Evolution of the Universe: Mission Operations and Data Analysis (MO&DA) Programs, June 27-29, 17 pp.

⁹National Aeronautics and Space Administration, 2000, Report of the Senior Review of Origins and Structure and Evolution of the Universe: Mission Operations and Data Analysis (MO&DA) Programs, June 27-29, 17 pp.

BOX 2.1

The Senior Review Process

The senior review, held every 2 years by an ad hoc panel of researchers active in the field being reviewed, is the highest level of peer review within the Space Science Enterprise. Senior review panels consider operating missions, data analysis from current and past missions, and supporting science and data facilities. Scientific merit is the primary evaluation criteria. The panels are chartered to carry out these tasks:

- Rank the scientific merit of the expected returns of the programs (or scientific usefulness of the data facilities) over the following two years;
- Assess cost efficiency, technology development and dissemination, and education/outreach; and
- Recommend an implementation strategy that considers continuing programs as originally planned or with enhancements or reductions, extending missions beyond prime phase, and terminating programs.

The senior review process has only recently been implemented in each of the space science programs. The astrophysics program has held six senior reviews since 1998, the Sun-Earth connection program held reviews in 1997 and 2001, and the first planetary science review was held in 2001.

SOURCE: G. Riegler, NASA Office of Space Science, white paper on the “Senior Review” Process, January 2002

Data from early planetary missions were disseminated in an ad hoc manner. No formal archives were kept, standards and formats varied widely, and in-depth and detailed knowledge of instrument and spacecraft operations was often required to use the data. Frequently, a strong working relationship with the instrument team was necessary. Many early planetary missions were exploratory, and the ability to independently browse, examine, and process large and comprehensive data sets was not a priority.

With the advent of modern instruments and the development of missions that obtain comprehensive measurements of solar system bodies, the planetary science community recognized the need for an established data system for archiving and distributing data. The Planetary Data System (PDS) has been in place for approximately 8 years, and data from all current and planned missions are required to be stored there.

The PDS consists of eight distributed discipline nodes, maintained at university or research centers across the country (see Table 2.3). The nodes were chosen by a competitive proposal process. Most of them are headed by a scientist actively working in the subject area of the node, and most have an advisory committee that meets regularly to review performance, goals, and developments in their area. Some of the discipline nodes (e.g., the Small Bodies Node) consist of several subnodes. A central PDS node at the Jet Propulsion Laboratory links the discipline nodes.¹⁰

The PDS facilitates access to planetary data from both ongoing and previous planetary missions. For example, users can access either original experimental data records or derived imaging products from the PDS Imaging Node over the Internet. The data can be searched either by spacecraft mission or by planetary target.¹¹ Although the bulk of its inventory consists of

¹⁰See <<http://pds.jpl.nasa.gov>> for descriptions and links to individual nodes.

¹¹See <<http://www-pdsimage.jpl.nasa.gov/PDS/public/jukebox.html>>.

TABLE 2.3 Planetary Data and Image Facilities

| Facility | Location |
|---|--|
| <i>PDS Nodes</i> | |
| Central Node | Jet Propulsion Laboratory, Pasadena, Calif. |
| Planetary Atmospheres Node | New Mexico State University, Las Cruces, N. Mex. |
| Geosciences Node | Washington University, St. Louis, Mo. |
| Imaging Node | Jet Propulsion Laboratory, Pasadena, Calif., and U.S. Geological Survey, Flagstaff, Ariz. |
| Navigation and Ancillary Information Facility | Jet Propulsion Laboratory, Pasadena, Calif. |
| Planetary Plasma Interactions Node | University of California, Los Angeles, Calif. |
| Rings Node | NASA Ames Research Center, Moffett Field, Calif. |
| Small Bodies Node | University of Maryland, College Park, Md. |
| <i>U.S. RPIFs</i> | |
| Center for Information and Research Services | Lunar and Planetary Institute, Houston, Tex. |
| Northeast Regional Planetary Data Center | Brown University, Providence, R.I. |
| Pacific Regional Planetary Data Center | University of Hawaii, Honolulu, Hawaii |
| Regional Planetary Image Facility | National Air and Space Museum, Washington, D.C. |
| Regional Planetary Image Facility | Washington University, St. Louis, Mo. |
| Regional Planetary Image Facility | Jet Propulsion Laboratory, Pasadena, Calif. |
| Regional Planetary Imaging Facility | U.S. Geological Survey, Flagstaff, Ariz. |
| Space Imagery Center | University of Arizona, Tucson, Ariz. |
| Space Photography Laboratory | Arizona State University, Tempe, Ariz. |
| Spacecraft Planetary Imaging Facility | Cornell University, Ithaca, N.Y. |
| <i>RPIF Centers in Other Countries</i> | |
| Israeli Regional Planetary Image Facility | Ben-Gurion University of the Negev, Beer-Sheva, Israel |
| Phototheque Planetaire | Universite Paris-Sud, Orsay, France |
| Nordic Regional Planetary Image Facility | University of Oulu, Oulu, Finland |
| Planetary and Space Science Centre | University of New Brunswick, Fredericton, Canada |
| Regional Planetary Image Facility | University College London, London, United Kingdom |
| Regional Planetary Image Facility | Institute of Space Sensor Technology and Planetary Exploration, Berlin, Germany |
| Regional Planetary Image Facility | Institute of Space and Astronomical Sciences, Sagami-hara-Shi, Kanagawa, Japan |
| Southern Europe Regional Planetary Image Facility | Consiglio Nazionale delle Ricerche Istituto de Astrofisica Spaziale, Area Ricerca di Roma Tor Vergata, Rome, Italy |

spacecraft data, the PDS also stores some ground-based telescope data, and even some theoretical model output.

The recent change in NASA's approach to planetary missions—from large, expensive, and infrequent missions such as Voyager, Galileo, and Cassini, to smaller and more frequent missions such as those in the Mars and Discovery programs—implies that the number of missions contributing data to the PDS will increase substantially in the near future. Moreover, because of advances in instrument technology, the new, smaller missions may yield larger volumes of data than those from the historic flagship missions. For these reasons, demands on the PDS are anticipated to grow exponentially in the near future (see Table 2.2).

The standard policy is to make all planetary data openly available both to scientists and to the public. In general, the proprietary period during which new data are only available to the science team members has decreased with time. The large planetary missions that typified the 1970s (e.g., Viking and Voyager) had proprietary periods of up to 18 months, which often led to considerable frustration among members of the science community who were not part of a flight instrument team. In contrast, the more frequent, smaller planetary missions have short or no proprietary periods. For example, the Mars Global Surveyor (MGS) spacecraft, in orbit around Mars since 1997, has no proprietary period; instead, there is a brief data validation period during which the science teams verify data quality prior to data releases at roughly 6-month intervals. The large quantity of data generated by the MGS instruments is released via the Internet, either at the appropriate PDS nodes or at a dedicated site maintained by the instrument science team. CD-ROMs of the same data are available a few months later from the PDS node. However, the steadily increasing data volumes will soon make it impractical to distribute all planetary data on CD-ROMs (or even DVDs).

The PDS nodes have evolved with the increasingly sophisticated needs of both researchers and the general public and with the rapidly growing volume of planetary data. The distributed nature of the PDS has both advantages and disadvantages. On the one hand, each node is tailored to the specific requirements of its research community and in that sense is highly responsive. Some nodes even distribute ancillary data (for example, absorption cross sections) and software that is particularly useful to its community. On the other hand, the existence of a large number of nodes does mandate continued oversight to ensure coordination and minimal redundancy.

The PDS has fundamentally changed the manner in which NASA planetary data are distributed. With the advent of this data system, fully calibrated and documented data can be retrieved remotely by researchers who have no relationship with the instrument PI. Moreover, the entire system inventory can be searched to discover data on a particular object or topic. The increase in availability and ease of retrieval provided by the PDS is a substantial benefit. However, along with automated distribution comes a decrease in the degree of guidance and interaction on research questions between data users and senior scientists associated with each spacecraft mission.

Another type of resource available to planetary scientists is the Regional Planetary Image Facility (RPIF). A network of 10 RPIFs was established in the United States in the early 1980s to help scientists obtain planetary data required for their research projects. NASA provides the RPIFs with copies of all planetary imaging data, along with annual support for data storage and maintenance. There are also 8 RPIFs in other countries, which only receive data (Table 2.3). In addition to serving scientists, the RPIFs serve as a resource to the local press, students, teachers, and the general public looking for information on planetary imaging data. Helping interested individuals (both scientists and nonspecialists) to find and obtain data appropriate for their needs

is a growing role for the RPIFs. Although these facilities have not yet been reviewed, NASA's Planetary Geology and Geophysics program has initiated a rotating schedule of reviews that will evaluate the performance of each RPIF every 5 years.

Solar and Space Physics Data Systems

Data related to the Sun and its influence on the interplanetary and Earth environment are managed by a variety of NASA-sponsored organizations, including the Solar Data Analysis Center (SDAC), PI and mission facilities, the National Solar Observatory, and the Stanford Solar and Heliospheric Observatory data center.¹² The NSSDC is both the active archive for space physics data (through the Space Physics Data Facility) and the permanent data center for U.S. solar and space physics data.

Data from these organizations as well as from other observatories and facilities around the world are increasingly available via the Internet. Proprietary periods are decreasing, and most solar and space physics data are now available a year or less after they were collected. Some observations, such as images from the Solar and Heliospheric Observatory, are even provided to scientists and the general public in real time.

The use of the Internet to disseminate solar and space physics data has led to a significant improvement in the ability of scientists to access data from different instruments and ground stations. However, many valuable data sets, particularly those held by individual PIs, remain offline. Moreover, as pointed out by a recent NRC report, searching across centers remains problematic, particularly for researchers who need to combine data from several archives.¹³ Although systems such as the Space Science Data System have been proposed to address this problem,¹⁴ the systems have largely lapsed, and users must rely on Web links provided by the individual centers to find data.

Solar Data Analysis Center

The Solar Data Analysis Center at Goddard Space Flight Center is the active archive for solar physics. It serves as the distribution center for a large and growing solar database and provides network access to data and images from such missions as the Solar and Heliospheric Observatory, Yohkoh, and the Transition Region and Coronal Explorer. Much of the data is distributed via network-attached servers with no interactive operating system. According to the archive manager, this approach is necessary for staying within a small budget (see Table 2.2). A senior review held in August 2001 found that the SDAC is an excellent example of a small discipline active archive that operates very cost-effectively and provides major services to the solar physics community.¹⁵

¹²NOAA centers, such as the National Geophysical Data Center and the World Data Center for Solar Terrestrial Physics, also manage U.S.-collected solar physics data.

¹³National Research Council, 1998, *Ground-Based Solar Research: An Assessment and Strategy for the Future*, National Academy Press, Washington, D.C., 47 pp. + 11 appendixes.

¹⁴Final Report of the Task Group on Science Data Management to the Office of Space Science, NASA, Jeffrey Linsky, chair, October 23, 1996, 61 pp.

¹⁵Senior Review of the Sun Earth Connection Missions Operations and Data Analysis Programs, August 29, 2001, <[http://spacescience.nasa.gov/admin/divisions/ss/SEC Senior Review 2001.pdf](http://spacescience.nasa.gov/admin/divisions/ss/SEC_Senior_Review_2001.pdf)>.

The PI team historically has been responsible for all aspects of handling space physics data derived from the instrument that they built. In the early years of space science, proprietary rights to the data lasted for 2 years after their receipt, and data analysis funding officially was planned for 2 years after launch. Data quality and level of processing varied greatly across the various PI data nodes. PI teams were encouraged to submit their data to the NSSDC for long-term maintenance. Generally, no time interval, medium, or format was specified for this submission.

Because of NASA encouragement and support in the more than 40 years since the beginning of the space age, more standardization has been introduced into the data management process. Instrument teams remain the focal point of all data-processing requirements and their implementation, and they are responsible for processing the data, developing higher-level data products, storing data, and maintaining accessibility. The data quality and level of processing are now not only more uniform across PI data nodes, but the data products are much more refined and sophisticated. PI teams also contribute processed low-resolution, “quick-look” data to missionwide databases or PI Web sites that are accessible in near-real time to the research community. However, high-resolution data, which are needed to study fundamental processes governing space plasmas, are not always widely available, owing to lack of funds. For example, a number of high-resolution data sets from the International Solar Terrestrial Program are available on neither NSSDC nor PI Web sites.¹⁶ Investigators are required to submit the full data set to NSSDC, although this requirement has not always been enforced, and resources have generally not been made available to do this job adequately. In some instances, an extended version of quick-look data is held in other archival systems (e.g., Galileo particles and field data are submitted to and held in the PDS).

Although some prelaunch support is available for planning and development of data-handling software, it is generally insufficient to provide fully usable data production immediately after launch. Postlaunch data processing and analysis are usually funded for 2 years after launch. While initial results and discoveries appear during this period, the primary scientific return occurs in the following several years, after confidence has been established in the data-processing software.

EARTH OBSERVING SYSTEM DATA AND INFORMATION SYSTEM

Because many of the important research problems studied by earth scientists are multidisciplinary in nature, the active archives of NASA’s Earth Science Enterprise were designed to be interoperable at the outset. The Earth Observing System (EOS) Data and Information System (EOSDIS) was built to process, disseminate, and archive data from the entire EOS program, with the goal of creating “one-stop shopping” for researchers interested in studying the Earth as a system.¹⁷ The objectives of this ambitious program include the following:

¹⁶For example, high-resolution data sets from the 3DP plasma instrument on the Wind spacecraft, the energetic particles instrument on the Geotail spacecraft, and the Hydra Plasma and Energetic Particles instruments on the Polar spacecraft are not available from NSSDC or PI Web sites. See <<http://nssdc.gsfc.nasa.gov/space/>>.

¹⁷For a history of EOSDIS, see National Research Council, 1998, *Review of NASA’s Distributed Active Archive Centers*, National Academy Press, Washington, D.C., 233 pp.

- Facilitate the creation of standard data products, thereby permitting the immediate scientific goals of the science teams to be realized.
- Catalyze the preparation of a wide range of secondary data sets and information products that combine information from different satellites and in situ sources, thereby stimulating collaborative, multidisciplinary research.
- Make such products readily accessible to the broader scientific community.
- Preserve data in usable forms for future generations of scientists.

As originally conceived, EOSDIS had two main elements: (1) the EOSDIS Core System (ECS), which was intended to perform a variety of functions—from spacecraft command and control to data acquisition, processing, distribution, and archiving; and (2) a network of eight distributed active archive centers (DAACs) to manage the data and provide user services (see Table 2.1). However, delays in the ECS and problems with the system design led to the adoption of back-up plans for processing data and creating data products. Data from most current Earth Science Enterprise (ESE) missions are being processed by science computing facilities (SCFs) using software designed and implemented for the task at hand, not by the ECS (see Table 2.4).¹⁸

The DAAC and SCF components of the system are working well. Users can obtain a wide range of data and products, and the use of common formats and standards permits the integration of different types and scales of data. In general, the production of data sets from all the currently operational missions (e.g., Landsat 7, Terra, the Tropical Rainfall Measurement Mission) is being performed in a timely fashion, including both level 1 and higher data products from each of the instruments. Each day, more than a terabyte (10^{12} bytes) of data is added to the EOSDIS archive, and 2 terabytes of products are distributed to the community through the DAACs. In addition to fulfilling the needs of scientific users, the DAACs are producing a variety of data products for use by nonscientists, including farmers and urban planners. These data products have already garnered a large and growing user community (see Table 2.5).

¹⁸Current plans call for the ECS to be used only for EOS missions: Terra, Aqua, Aura, ICESat, SOURCE, SAGE III, and ACRIMSAT. The ECS will provide the full suite of services for the largest EOS missions (Terra, Aqua, and Aura), including satellite control and data downlink, and data distribution, processing, and archiving. For the other missions, the ECS will provide only data distribution, processing and archiving capabilities.

TABLE 2.4 Processing Summary for EOSDIS Instruments

| Mission | Instrument | Level 0 Processing ^a | Level 1 Processing ^a | Level 2 Processing ^a |
|--------------------------|--|---|--|---|
| <i>Current Missions</i> | | | | |
| ERBS | ERBS SAGE | LaRC Instrument SCF | LaRC Instrument SCF | LaRC Instrument SCF |
| TOMS-EP | TOMS | Instrument SCF | Instrument SCF | Instrument SCF |
| TOPEX/Poseidon | NASA ALT | Instrument SCF | Instrument SCF | Instrument SCF |
| UARS | All | CDPF | CDPF | CDPF |
| TRMM | TIM PR VIRS CERES LIS | TSDIS TSDIS TSDIS LaTIS LIS SCF | TSDIS TSDIS TSDIS LaTIS LIS SCF | TSDIS TSDIS TSDIS LaTIS LIS SCF |
| SeaStar | SeaWiFS | Instrument SCF | Instrument SCF | Instrument SCF |
| Landsat 7 | ETM+ | LPGS | LPGS | N/A |
| Terra | MODIS CERES MOPITT MISR ASTER | EDOS EDOS EDOS EDOS EDOS | GSFC DAAC/ECS LaTIS MOPITT SIPS LaRC DAAC/ECS ERSDAC Japan | MODAPS LaTIS MOPITT SIPS LaRC DAAC/ECS EDC DAAC/ECS |
| ACRIMSAT | ACRIM | Instrument SCF | Instrument SCF | Instrument SCF |
| QuikSCAT | Sea Winds | Instrument SCF | Instrument SCF | Instrument SCF |
| <i>Upcoming Missions</i> | | | | |
| Meteor | SAGE III | Instrument SCF | Instrument SCF | Instrument SCF |
| ADEOS II | SeaWinds | Instrument SCF | Instrument SCF | Instrument SCF |
| Jason | Poseidon-2/ DORIS/JMR | Instrument SCF | Instrument SCF | Instrument SCF |
| Aqua | MODIS AIRS, HSB, AMSU AMSR-E CERES | EDOS EDOS EDOS EDOS | GSFC DAAC/ECS GSFC DAAC/ECS NASDA LaTIS | MODAPS GSFC DAAC/ECS Instrument SCF LaTIS |
| SORCE | SOLSTICE | Instrument SCF | Instrument SCF | Instrument SCF |
| ICESat | GLAS | EDOS | Instrument SCF | Instrument SCF |

^a Level 0 = Reconstructed unprocessed instrument data with all communications artifacts removed; Level 1 = Level 0 data that have been calibrated, time referenced, and annotated with ancillary information; Level 2 = Higher-level data products, e.g., derived geophysical variables at the same resolution and location as the Level 1 data. Definitions modified from G. Asrar and R.

Greenstone, eds., 1995, *MTPE/EOS Reference Handbook*, National Aeronautics and Space Administration, NP-215, Washington, D.C., 276 pp.

NOTE: ACRIM = Active Cavity Radiometer Irradiance Monitor; ACRIMSat = ACRIM Satellite; ADEOS = Advanced Earth Observation Satellite (Japan); AIRS = Atmospheric Infrared Sounder; ALT = Altimeter; AMSU = Advanced Microwave Sounding Unit; ASTER = Advanced Spaceborne Thermal Emission and Reflection Radiometer; CDPF = Central Data Processing Facility (UARS); CERES = Clouds and the Earth's Radiant Energy System; DORIS = Doppler Orbitography and Radiopositioning Integrated by Satellite; EDOS = EOS Data Operations Systems; EP = Earth Probe; ERBE = Earth Radiation Budget Experiment; ERBS = Earth Radiation Budget Satellite; ERSDAC = Earth Remote Sensing Data Analysis Center; ETM = Enhanced Thematic Mapper; GHRC = Global Hydrology Resource Center (Huntsville AL); GLAS = Geoscience LASER Altimeter System; HSB = Humidity Sounder for Brazil; ICESat = Ice Clouds and land Elevation Satellite; JMR = Jason-1 Microwave Radiometer; Landsat = Land Satellite; LaTIS = Langley TRMM Information System) (LaRC DAAC V0); LIS = Lightning Imaging Sensor; LPGS = Landsat Product Generation System; MISR = Multi-angle Imaging SpectroRadiometer; MODAPS = MODIS Adaptive Production System; MODIS = Moderate Resolution Imaging Spectroradiometer; MOPITT = Measurement of Pollution in the Troposphere; PR = Precipitation RADAR; QuikSCAT = Quick Scatterometer; SAGE = Stratospheric Aerosol and Gas Experiment; SCF = Science Computing Facility; SeaWiFS = Sea-Viewing Wide-Field-of-View Sensor; SIM = Spectral Irradiance Monitor; SOLSTICE = Solar Stellar Irradiance Comparison Experiment; SORCE = Solar Radiation and Climate Experiment; TIM = Total Irradiance Monitor; TOMS = Total Ozone Mapping Spectrometer; TOPEX = Topography Experiment; TRMM = Tropical Rainfall Measuring Mission; TSDIS = TRMM Satellite Data and Information System; UARS = Upper Atmospheric Research Satellite; V0 = Version 0 DAAC Developed System; V1 = Version 1GSFC DAAC Developed System (for TRMM); VIRS = Visible Infrared Spectroradiometer; XPS = XUV Photometer System.

SOURCE: V. Griffen, Science Operations Manager, Goddard Space Flight Center, August 2001.

TABLE 2.5 Characteristics of DAACs

| Center | Number of Users ^a | Budget (\$M) | | Holdings (TB) | | Number of Staff |
|---------|------------------------------|--------------|---------|---------------|---------|-----------------|
| | | FY 2000 | FY 2005 | FY 2000 | FY 2005 | |
| ASF | 736 | 13.3 | 6.8 | 239 | 712 | 68 |
| EDC | 20,004 | 4.1 | 11.2 | 74 | 3148 | 87 |
| GSFC | 47,144 | 5.4 | 12.7 | 154 | 1465 | 131 |
| LaRC | 3,570 | 10.5 | 12.5 | 39 | 610 | 105 |
| NSIDC | 1,225 | 3.5 | 5.2 | 5 | 72 | 39 |
| ORNL | 1,973 | 2.4 | 3.0 | 0.3 | 3 | 13 |
| PO.DAAC | 15,657 | 5.4 | 6.1 | 8 | 42 | 30 |
| SEDAC | 17,000 | 3.0 | 4.0 | 0.1 | 0.2 | 27 |

^a Unique users who received data in FY 2000.

NOTE: Budgets and holdings for FY 2005 are estimated.

SOURCE: Managers of the DAACs (see questionnaire in Appendix C).

In contrast to the DAAC and SCF components, the capabilities of the ECS component of EOSDIS fall short of those originally envisioned. Early operational problems included: (1) processing delays or failures were caused by bit flips in data and system outages and anomalies; (2) data gaps and missing data files hindered the ability to process the science data routinely; (3) the DAACs and instrument teams promoted new science algorithms, which contributed to the processing backlog; (4) the need for reprocessing was greater than anticipated; and (5) commercial-off-the-shelf (COTS) and system tuning issues decreased system stability.¹⁹ NASA has worked diligently to correct these issues, but the capacity of EOSDIS to process and distribute data has not been sufficient to meet all of the expectations of the earth science community. As noted by the Office of the Inspector General, "The ECS contract has been

¹⁹Report of NASA's Earth Systems Data and Information Systems and Services Advisory Subcommittee, April 27-28, 2000, Washington, D.C.

problematic with significant delays. The entire ECS as originally envisioned is no longer affordable.’²⁰ Accordingly, Goddard Space Flight Center issued a request for proposal in 1998 to restructure the contract. The restructuring defers and/or eliminates some lower-level-data processing functionality; provides less user support; reduces production capacity by 25 percent; discards interim products after 6 months; reduces distribution capacity to users by one-third; reduces timeliness of data distribution; and permits DAACs and SCFs to take on some ECS functions. The increase in the estimated cost of the ECS contract is \$98.8 million for 3 years, which includes the reduced requirements; inclusion of a new flight segment approach for Terra; the addition of the control center requirement for Aqua; and the addition of science data management for Aqua, Aura, and ICESat.²¹ The total award fee to the ECS contractor was decreased 12 percent owing to poor performance in both cost and technical management.

The primary reason for the shortcomings in the ECS capabilities is probably that the ECS software is far too complicated ever to achieve a high degree of reliability. Discussions with DAAC managers and ECS developers suggest that although the ECS software has become increasingly stable over the 22 months since its initial release, it remains fragile. For example, the Moderate Resolution Imaging Spectroradiometer ECS data flow had been running with a 90 to 92 percent uptime prior to the release of the ECS 6A04 software in the summer of 2001.²² After the new software was installed, uptime dropped to only 84 percent, but gradually returned to previous levels as software patches were implemented. Such fragility is symptomatic of a system that is too large (there are currently over 1.2 million lines of code and more than 40 COTS packages) and too complex to be properly tested, maintained, and extended. For instance, testing of the ECS release 6A04 software was incomplete, partly because of the prohibitive expense of testing the performance of the system and partly because of the requirement to rush software to operations to meet the schedule.

It is not clear what should be done with the ECS software in the future. Data streams that are currently captured or processed using the ECS software will continue for several more years, so this software will have to be maintained. On the other hand, a number of tasks handled by the ECS software could possibly be performed more reliably and/or cost-effectively using other existing software.²³ Similarly, capabilities not currently part of the ECS could be provided by other software. For example, the Land Rapid Response Project is producing level 1B MODIS products within three to five hours of receiving level 0 granules.²⁴ Since the focus of the data pipeline is to produce level 3 fire products for use by the National Oceanic and Atmospheric Administration (NOAA) and the U.S. Forest Service, level 0 granules corresponding to portions of the Earth covered with water are currently discarded. However, according to a PI on the project, the addition of a small increment of computing capability (a few more nodes) could enable the pipeline to produce level 1B MODIS data sets for the entire Earth with the same

²⁰Office of Inspector General, 1999, Performance Evaluation Plan for the Earth Observing System Data and Information System Core System Contract, IG-99-038, September 8.

²¹Martha Maiden, Code YF Data Network Manager, personal communication, February 2002. The \$100 million was allocated to the ECS contractor and the Science Computing Facilities that wished to process data.

²²Steven Kempler, Manager, Goddard DAAC, personal communication, August 2001 and March 2002.

²³For example, according to the Langley DAAC manager, the LaTIS software, which is already being used to handle data from the Clouds and the Earth’s Radiant Energy System instruments on Terra and TRMM, could have been used for the Multi-angle Imaging Spectroradiometer and Measurements of Pollution in the Troposphere instruments.

²⁴In contrast, the Goddard DAAC normally requires 24 to 48 hours. See <<http://rapidfire.sci.gsfc.nasa.gov/index.html>>.

degree of delay.²⁵ Further modifying the software to permit receipt of broadcasts directly from the Terra satellite could eventually lead to a worldwide network of sites generating MODIS (and other) products in near real time. These concepts have yet to be tested, and it remains to be seen whether the system architecture and operations plan of the Land Rapid Response Project would be scalable. Nevertheless, systems that grow from small, focused efforts on the part of many individuals and organizations are commonly more successful than top-down, centralized approaches because they are simpler and more flexible.²⁶ A recent NRC report laid out the following principles for creating small, evolvable information systems:²⁷

- Because the analysis of long-term data sets must be supported in an environment of changing technical capability and user requirements, any data system should focus on simplicity and endurance.
- Adaptability and flexibility are essential for any information system if it is to survive in a world of rapidly changing technical capabilities and science requirements.
- Experience with actual data and actual users can be acquired by starting to build small, end-to-end systems early in the process. EOS data are available now for prototyping new data systems and services. . . .

The task group agrees with these principles and encourages NASA to adopt them in future data and information systems.

STRATEGIC EVOLUTION OF ESE DATA SYSTEMS

NASA recognizes the problems associated with the ECS and is developing a strategy for the evolution of the network of data systems and service providers that support the Earth Science Enterprise.²⁸ The next-generation system is called SEEDS (Strategic Evolution of ESE Data Systems, formerly known as NewDISS). SEEDS is intended to support all phases of the data management life cycle: (1) acquisition of sensor, ancillary, and ground validation products necessary for processing; (2) processing of data; (3) generation of value-added products via subsetting, format translation, and data mining; (4) archiving and distributing products; and (5) providing search, visualization, subsetting, translation, and order services to assist users in identifying, selecting, and acquiring products of interest. Study teams drawn from the user community are being engaged to identify options, define scope, and establish schedule requirements. It is intended that SEEDS will be managed and implemented as an open and distributed information system architecture under a unifying framework of standards, core interfaces, and levels of service.

SEEDS faces a number of major challenges, including determining how to organize and manage a distributed system and achieving a balance between providing science teams with the

²⁵Jacque Descloitres, Goddard Space Flight Center and PI of the Land Rapid Response Project, personal communication to D. DeWitt, September 2001.

²⁶A lesson learned from the modernization of the Internal Revenue Service was that complex systems should be developed by making incremental changes to small, successful projects, rather than by building all components of the system simultaneously (National Research Council, 1996, *Continued Review of the Tax Systems Modernization of the Internal Revenue Service: Final Report*, National Academy Press, Washington, D.C., 101 pp.).

²⁷National Research Council, 2000, *Ensuring the Climate Record from the NPP and NPOESS Meteorological Satellites*, National Academy Press, Washington, D.C., p 3.

²⁸Briefing to the task group by Steven Wharton, NewDISS program formulation manager, July 30, 2001.

appropriate levels of freedom in developing and operating science data systems while maintaining NASA agency accountability for data stewardship and accessibility.

The preformulation phase of SEEDS was initiated in 1998, and the formulation phase is scheduled to conclude in 2003. The SEEDS program has solicited lessons learned from EOSDIS, which NASA summarized for the task group as follows:²⁹

- *Information technology outpaces the time required to build large, operational data systems and services.* Technology is now changing at such a rapid pace that it is impossible to predict technological solutions even 2 years into the future. And, in contrast with 10 to 15 years ago, government information systems no longer drive the development of hardware and software; NASA is now just another customer trying to capture the attention of the vendors.
- *Data systems and services should leverage off emerging information technology and not try to drive it.* Since NASA can no longer drive commercial hardware and software development, SEEDS must be open to the infusion of new technologies developed by industry. A few years ago many of these industries were completely unassociated with digital information management but are now leaders in the field.³⁰
- *A single data system should not attempt to be all things to all users.* The ESE research and applications community is extraordinarily diverse, ranging from scientific researchers to for-profit companies, policy makers, government operations, and the general public. The standards and practices governing the acquisition, archiving, documentation, distribution, and analysis of earth science data vary by user group as well as by scientific discipline. SEEDS must recognize and embrace this tapestry of disciplines and subcommunities; there is no one-size-fits-all solution to the myriad data management needs of the community as a whole.
- *A single, large design- and development-contract stifles creativity.* Given the complexity of the required systems and services, the volatility of the technology, and the potential for changes in scientific priorities, centralized development is too inflexible and increases the risk that large portions of the data system will be vulnerable to single-point failures. Such an approach is also prone to “monopolistic” tendencies and does not encourage the kind of diversity and variety found in a competitive marketplace.
- *Future information systems will be distributed and heterogeneous in nature.* Management tools and practices must encourage a flexible, distributed, and loosely coupled network of data providers, even if this requires a fundamentally new management approach within the NASA culture.

The task group agrees with these conclusions and notes that many of these “lessons learned” describe the more evolutionary approach implemented successfully in the development to date of the Astrophysics Data System.

NASA had not yet completed its plans for SEEDS prior to completion of this report. Therefore, the task group cannot comment on whether or not SEEDS will meet the needs of the earth science community. Also, no information was available about what role the ECS software will play in the SEEDS effort. Will it be replaced? Evolved? Or simply maintained in its current state? The task group is concerned, however, by the timelines that were provided for the SEEDS

²⁹Briefing to the task group by Steven Wharton, NewDISS program formulation manager, July 30, 2001. The lessons learned were lightly edited for conciseness.

³⁰For example, the banking, entertainment, and retail industries.

effort.³¹ The timeline specifies five years of planning and seven years of implementation. This extended time for both phases is inconsistent with the rapid timescales for the evolution of relevant technologies and appears inconsistent with the first of the EOSDIS lessons listed above.

Recommendation. The ECS (the EOSDIS Core System) software should be placed in a maintenance mode with no (or very limited) further development until a concrete plan for the follow-on system, SEEDS (Strategic Evolution of ESE Data Systems), has been formulated, its relationship to ECS defined, and the plan reviewed by an external advisory group. This plan should be measured against the lessons learned from EOSDIS and from the experience in other disciplines, and should include provisions for rapid prototyping and an evolutionary and distributed approach to implementing new capabilities, with priorities established by the scientific and other user communities.

LONG-TERM MAINTENANCE OF DATA

The growing body of NASA data is becoming an increasingly powerful tool for identifying and monitoring long-term changes in objects as nearby as rain forests and as distant as supernovae near the edge of the visible universe. Long-term maintenance requires much more than just making sure that the data are preserved and that the storage media are kept up to date.³² In order to ensure that archived data sets can continue to be used in the future, they must be properly documented, stored with data access and processing software, and migrated regularly to new media, operating systems, and so on. Only by continually reprocessing all data sets and data products can one ensure that the data will be viable 50 years from now.

NASA data are federal records and thus must comply with standards developed by the National Archives and Records Administration (NARA). NARA provides guidance to federal agencies on the management of records, the retention and disposition of records, and the storage of records in centers from which agencies and their agents can retrieve them.³³ In addition, NARA collaborates with other federal agencies and universities to develop new archiving approaches. Examples include the Persistent Archive Initiative, and the Methodologies for Preservation and Access of Software-dependent Electronic Records, which are being carried out by the San Diego Supercomputing Center with NARA funding.³⁴ The goals of the Persistent Archive Initiative are to develop an information architecture that can evolve with changes in technologies into the indefinite future. Work on maintaining the ability to discover and access digital objects while the supporting hardware and software systems evolve is of particular importance to NASA, since so many NASA mission data are dependent upon software systems. The Methodologies project is concerned with developing software-independent tools for

³¹Briefing to the task group by Steven Wharton, NewDISS program formulation manager, July 30, 2001.

³²A number of NRC reports have discussed the rationale and provided principles for the long-term maintenance of scientific data. For example, see National Research Council, 1982, *Data Management and Computation: Volume I: Issues and Recommendations*, National Academy Press, Washington, D.C., 167 pp.; National Research Council, 1995, *Preserving Data on Our Physical Universe*, National Academy Press, Washington, D.C., 67 pp.; National Research Council, 2000, *Ensuring the Climate Record from the NPP and NPOESS Meteorological Satellites*, National Academy Press, Washington, D.C., 51 pp.

³³See <<http://www.nara.gov/records/>>.

³⁴See <<http://www.sdsc.edu/NARA/>>.

archiving and accessing data.³⁵ Central to this work is the development of criteria for infrastructure-independent representations of electronic documents (including spatial data), which are key to providing access to complex scientific data over time. This project is also contributing to major grid projects such as the National Virtual Observatory (see Chapter 4) and NASA's Information Power Grid.

Attention must also be paid to international standards, because many countries collect data used in U.S. earth and space science studies. An example of such standards is the International Organization for Standardization reference model for long-term maintenance of data sets, which was recently developed by the Consultative Committee for Space Data Systems.³⁶ The members of this committee included representatives from NASA and space agencies in Europe and Japan. A 2000 NRC report found that the OAIS (Open Archival Information System) model is "important for digital preservation standards and strategies because it defines the functions and requirements for a digital archive through an international standard that vendors and producers of digital information can reference."³⁷

The OAIS reference model addresses a full range of archival information-preservation functions, including ingest, archival storage, data management, access, and dissemination.³⁸ It covers the migration of digital information to new media and forms, the data models used to represent the information, the role of software in information preservation, and the exchange of digital information among archives. Both internal and external interfaces to the archive functions are identified, as well as a number of high-level services at these interfaces. Finally, the reference model defines a minimal set of responsibilities for an archive to be called an OAIS and an optimum archive in order to provide a broad set of useful terms and concepts.

The earth and space sciences have taken different approaches to long-term maintenance of data. Space science data are maintained indefinitely at the NSSDC. In contrast, earth science data will be transferred to agencies mandated to archive data—the U.S. Geological Survey (USGS) and NOAA—15 years after collection. Both approaches entail a risk to the usefulness of data to future generations of scientists, as detailed below.

Space Science Data and the National Space Science Data Center

The mission of the NSSDC is "to provide data and information from space flight experiments for studies beyond those performed by the principal investigators."³⁹ The NSSDC acts as the active archive for most space physics data and selected long-wavelength astrophysics data. Much of its current emphasis is on serving the heliospheric, magnetospheric, and ionospheric communities. The NSSDC also serves as the data center for long-term maintenance of data from all other space science missions. It receives data directly from spacecraft project data facilities or their PIs as well as from other space science active archives (e.g., PDS nodes, HEASARC). However, as noted above, only a fraction of data from these missions is actually contributed to the NSSDC; scientifically important data are commonly held by the PIs or active archives.

³⁵See <<http://www.sdsc.edu/NHPRC>>.

³⁶See <<http://www.ccsds.org/>>.

³⁷National Research Council, 2000, *LC21: A Digital Strategy for the Library of Congress*, National Academy Press, Washington, D.C., pp. 112.

³⁸See <<http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf>>.

³⁹NSSDC Charge and Service Policy, <http://nssdc.gsfc.nasa.gov/nssdc/cands_policy.html>.

Traditionally, NSSDC has archived only processed data, but it is increasingly being asked to include raw data and software.

It has been required since 1993 that every project data management plan specify what data will be maintained in the long term, when they will be sent to the data center, and in what easily usable format.⁴⁰ Yet, some scientific data are not reaching the data center because the PIs do not have the resources to prepare the material for archiving. Those data sets that do reach the NSSDC are not always formatted for convenient use for downstream users.⁴¹ For example, data contributed from past space physics missions are typically processed at a low level and are not well enough documented to be used for purposes and by investigators outside the original project. Planetary science data from previous missions are in a variety of formats, typically a different formatting scheme and processing software for each instrument, although the PDS has since developed standards for documenting planetary science data.

In the past decade, NSSDC has taken a number of steps to improve both its data center functions and the services it offers to the scientific community. Data are now held in climate-controlled conditions, and back-up copies are stored in commercial facilities that are compliant with NARA standards. However, the recent destruction of thousands of historic images because of water damage⁴² illustrates the need to devote additional attention to the safety of the holdings, particularly the nondigital records.

The operations of the NSSDC have been addressed by two recent Office of Space Science senior reviews. The 2000 astrophysics senior review concluded that the NSSDC archives data satisfactorily and with apparent care. However, the review recommended that the NSSDC work more closely with other active archives in terms of connectivity and active linking and with the goal of sorting out overlapping functions in order to streamline the agency's overall data storage, archiving, and handling functions.⁴³ The 2001 senior review of the Sun-Earth Connection program expressed concern about the long-term availability of solar data and recommended that the current informal agreements concerning the transfer of data from SDAC, the active archive, to NSSDC be formalized.⁴⁴ This review also noted that the NSSDC had incorporated value-added services that have greatly facilitated accessibility and research in space physics. Finally, the senior review encouraged the NSSDC to complete planning for how to archive raw data and software.

NASA has substantially increased its budget for archive activities (including the active archives) over the last 10 to 15 years. However, funding for NSSDC has declined by 6 percent since the late 1990s.⁴⁵ NSSDC budgets are projected to remain flat or decline further over the next 5 years, even though holdings are projected to increase by 30 to 40 percent, resulting in a substantial decrease in the number of real dollars available for archival activities. Activities

⁴⁰National Aeronautics and Space Administration, 1993, Guidelines for Development of a Project Data Management Plan (PDMP), Office of Space Science and Applications.

⁴¹National Research Council, 1993, *1992 Review of the World Data Center-A for Rockets and Satellites and the National Space Science Data Center*, National Academy Press, Washington, D.C., 80 pp.; Final Report of the Task Group on Science Data Management to the Office of Space Science, NASA, Jeffrey Linsky, chair, October 23, 1996, 61 pp.

⁴²Burst pipe inundates NASA photo archives, *Washington Post*, May 10, 2001.

⁴³National Aeronautics and Space Administration, 2000, Report of the Senior Review of Origins and Structure and Evolution of the Universe: Mission Operations and Data Analysis (MO&DA) Programs, June 27-29, 17 pp.

⁴⁴National Aeronautics and Space Administration, 2001, Final Report of the Senior Review of the Sun-Earth Connection Mission Operations and Data Analysis Programs, 27 pp.

⁴⁵Joe King, director of the National Space Science Data Center, personal communication, December 2001 and March 2002, and written response to a task group questionnaire in April 2001.

meant to serve the general public have been reduced or eliminated to accommodate the budget cuts, but if these trends continue, the NSSDC may not meet the needs of the scientific community in the future.

Earth Science Data

The earth science active archives are meant to hold data until 15 years after the mission. At that time, responsibility for the data will be transferred to federal agencies with a long-term data maintenance mission—NOAA and the USGS. The USGS has obtained funding for the long-term maintenance of Landsat data, but funding is still not available for archiving the majority of the data at NOAA. The 1989 NASA/NOAA Memorandum of Understanding specifies that NASA will “transfer to NOAA, at a time to be determined, responsibility for active long-term archiving and appropriate science support activities for atmosphere and oceans data.”⁴⁶ NASA and NOAA are responsible for making “joint presentations to NASA, DOC [U.S. Department of Commerce], NOAA, OMB [Office of Management and Budget], and the Congress, as necessary, to explain the essential role of each organization and funding needs.” These efforts have been largely unsuccessful, although the president’s budget for Fiscal Year 2003 includes \$3 million to begin archiving NASA EOS data at NOAA’s National Climatic Data Center. However, as noted by a 2000 NRC report, “even if this work is fully funded by Congress, it should be recognized that substantially greater investments will be required to develop the [data center].”⁴⁷

The uncertainty over the ultimate fate of EOS data has long been a concern of scientific researchers and science agencies.⁴⁸ For example, some are concerned that data will be transferred from scientists and data managers who work with the data and thus understand their usefulness and limitations to data managers without similar experience. This is not an issue for the Landsat holdings, which are already collocated with the Landsat data center. A similar solution, in which a NOAA data center is built at Goddard Space Flight Center, is being considered for atmosphere and oceans data. In 1998, NASA and NOAA sponsored a workshop to develop guiding principles for long-term maintenance of Earth observation data and for assessing lessons learned from current and past experience (see Box 2.2). In 2000, an NRC report outlined the initial steps that should be taken to ensure the continuity of the climate record in the transition, including the following:⁴⁹

⁴⁶Memorandum of Understanding between the National Aeronautics and Space Administration and the National Oceanic and Atmospheric Administration for Earth observations remotely sensed data processing, distribution, archiving, and related science support, July 1989.

⁴⁷National Research Council, 2000, *Ensuring the Climate Record from the NPP and NPOESS Meteorological Satellites*, National Academy Press, Washington, D.C., pp. 26. The National Climatic Data Center estimates that it will require an additional \$13 million to \$20 million per year to handle the increase in data volume and provide user services.

⁴⁸National Research Council, 1994, *Panel to Review EOSDIS Plans: Final Report*, National Research Council, Washington, D.C., 88 pp.; National Research Council, 1998, *Review of NASA’s Distributed Active Archive Centers*, National Academy Press, Washington, D.C., 233 pp.; National Research Council, 2001, *Enhancing NASA’s Contributions to Polar Science: A Review of Polar Geophysical Data Sets*, National Academy Press, Washington, D.C.; National Research Council, 2000, *Ensuring the Climate Record from the NPP and NPOESS Meteorological Satellites*, National Academy Press, Washington, D.C., 51 pp.

⁴⁹National Research Council, 2000, *Ensuring the Climate Record from the NPP and NPOESS Meteorological Satellites*, National Academy Press, Washington, D.C., 51 pp.

- NOAA should begin now to develop and implement the capability to preserve in perpetuity the basic satellite measurements (radiances and brightness temperatures);
- NASA, in cooperation with NOAA, should support the development and evaluation of climate data records, as well as their refinement through data reprocessing;
- NOAA and NASA should define and develop a basic set of user services and tools to meet specific functions for the science community, with NOAA assuming increasing responsibility for this activity as data migrates to the long-term archive; and
- NASA and NOAA should develop and support activities that will enable a blend of distributed and centralized data and information services for climate research.

NASA and NOAA should not address these issues in isolation. A number of efforts underway, such as those sponsored by NARA, are developing technologies and approaches to supporting long-term preservation and access to data. Consultation with NARA should be very useful in planning the transition from NASA to NOAA data centers, once adequate funding is secured.

BOX 2.2

Findings from the Report of a Workshop: *Global Change Science Requirements for Long-Term Archiving*

According to a 1998 workshop sponsored by NASA and NOAA, data centers should be supported by two guiding principles:

1. A data center must be established and operated in the simplest way possible to meet user needs and program goals, and
2. A data center is not only for today's generation of users, but also for the next generation of scientists and citizens whose needs have yet to be expressed but must be provided for.

Specific findings include the following:

- The data center must be actively engaged with its user community, including scientists, observing-system managers, private-sector users, and data experts.
- The data center must develop procedures and criteria for determining what data are to be included, excluded, and removed from the data center. The center should be driven by present science priorities, scientific assessments, general public needs, and national interests.
- The data center must ensure that the archived data sets and products are accompanied by complete, comprehensive, and accurate documentation so that they are useful for users. Information about the physical location and access paths to the data must be easily available.
- Data and documentation from operational or research sources must be verified, stored, cataloged, and made available as soon as possible to meet user needs. The ability to access data for re-analysis when improvements are made in data-processing algorithms is required.
- The data must be preserved and maintained in perpetuity. Integrity checks during the migration of data from one type of media to another must occur on a routine basis to prevent the data from becoming inaccessible or deteriorating beyond repair.
- Customer service and technical representatives are required for user support and to ensure that users' access needs are met. Research points of contact are also required. Near-real-time data should be accessible with minutes or hours from the time of acquisition, and other archived data should be accessible within hours or days of processing.

SOURCE: Adapted from U.S. Global Change Research Program, 1999, *Global Change Science Requirements for Long-Term Archiving*, Report from a workshop, National Center for Atmospheric Research, Boulder, Colorado, October 28-30, 1998, 78 pp.

Conclusions

NASA should have in place both a strategy and funding for long-term maintenance that will preserve data in usable forms. Since the data are a national resource, their preservation is an appropriate federal responsibility and should not be left solely to contractors or principal investigators. If resources are inadequate for preservation, NASA should establish a process involving the scientific community to examine the priorities between acquiring new data and preserving existing data for ongoing scientific uses.

Recommendation. NASA should assume formal responsibility for maintaining its data sets and ensuring long-term access to them to permit new investigations that will continue to add to our scientific understanding. In some cases, it may be appropriate to transfer this responsibility to other federal agencies, but NASA must continue to maintain the data until adequate resources for preservation and access are available at the agency scheduled to receive the data from NASA.

3

The Users of NASA Data

Chapters 1 and 2 of this report discuss the importance of NASA data in advancing our knowledge of the world around us and describe the structure of the active archives that distribute the data. This chapter looks at the question of who uses NASA data and assesses user satisfaction with the systems that are in place. The assessment is based on input from relevant NRC standing committees, the chairs of three NASA advisory committees that have addressed data usage issues, interviews with colleagues, briefings from NASA data system and education program representatives, and data collected by this task group on NASA's major data facilities and services (see Appendix C for the questionnaire used).

USER PROFILE

The mission of NASA's Earth Science and Space Science Enterprises is to conduct science and communicate the resulting knowledge to the public.¹ Scientists, who design the experiments, analyze the returned data, and publish results in scientific journals, are the key link in the knowledge-generation chain, and meeting their needs is given the highest priority in both enterprises. Scientific users can be divided into two categories: discipline scientists who have an in-depth understanding of a particular instrument or observatory and multidisciplinary scientists who need to access and integrate data from a variety of sources. The latter category is particularly prevalent in the earth sciences and is becoming more common in some of the space science subdisciplines. It is important for these users to have easy access to the data, adequate and standardized metadata and data formats, and tools to organize the data.

Other major user groups include NASA engineers and managers, who use data from previously flown spacecraft to plan future missions (see Appendix B); the general public, which is particularly interested in images, movies, and popular science features; and the education community. Although some of these users need to combine data from a variety of sources, most require small data sets, packaged for their particular application.

In addition to the user groups mentioned, the data systems of the Earth Science Enterprise (ESE) serve commercial companies, which use remotely sensed data to create value-added products targeted to specific customer groups; federal, state, and local government agencies, which use NASA data for operational purposes, such as predicting weather patterns or making land use plans; and policy makers, who need to make decisions on subjects such as managing the

¹National Aeronautics and Space Administration, 2000, *NASA 2000 Strategic Plan*, Washington, D.C., p. 60.

Earth's resources. These groups typically require custom data sets and comprehensive user services. Some of these users are served by the Distributed Active Archive Centers (DAACs), but increasing numbers are served by short-term, focused programs such as Earth Science Information Partners (ESIPs) and Regional Earth Science Applications Centers (RESACs).

Tables 2.2 and 2.5 in Chapter 2 indicate that both the earth and space science active archives serve a large user community. The DAACs supplied data to more than 104,000 unique users in FY 2000,² greatly exceeding even NASA's original expectations. The number of space science users is more difficult to determine, because some of the active archives only count Web site hits, which are considerably higher than the number of data requests. Nevertheless, it is clear that there are many tens of thousands of space science users. Using the membership of the American Geophysical Union as a proxy for the size of the earth and space science community (38,000 members from 115 countries),³ it is apparent that the data facilities and services serve more than just NASA investigators.

As noted above, the user community—especially of the DAACs—is quite diverse and thus is challenging to characterize in detail. All DAACs track electronic address extensions (e.g., .com, .edu, and so on) as a general measure of who accesses a site or obtains data. Figure 3.1 shows that scientists and government agencies—the highest-priority users—make up only a small fraction of total DAAC users. This observation highlights the importance of paying significant attention to the needs of the nonscientific community. Although the DAACs are aware of the broad characteristics of their user communities, a recent National Research Council (NRC) report found that few DAACs have a detailed understanding of their user profiles.⁴ The task group's survey (see Appendix C) found the same still to be true.

The task group was unable to obtain user statistics on most of the space science active archives. However, based on the response of these archives to the task group's survey, they appear to use similarly inadequate metrics to characterize their user communities. Some centers (e.g., Solar Data Analysis Center [SDAC]) do not even keep track of IP addresses and thus do not know the size or composition of their user community. All of the centers should know this basic information. A better understanding of their user profile would help them know whether it is scientifically necessary to expand the user base or provide new specialized services. The task group recognizes, however, that obtaining this information would increase operational costs and would require a re-evaluation of priorities by the supporting NASA office.

In the earth sciences, some specialized products and services are provided by ESIPs. The ESIPs were created in 1998 in part to develop value-added products from EOS and related data and to provide data services that are not being provided by the DAACs.⁵ Like the DAACs, the ESIPs track users by electronic address extensions, which provides only limited information about users. Using this categorization, the breakdown of users in the first quarter of 2001 was as follows: 33 percent education, 17 percent government, 14 percent commercial, 0.2 percent

²The actual number of users might be slightly lower, since many scientists use more than one DAAC (see the results of a user survey in National Research Council, 1998, *Review of NASA's Distributed Active Archive Centers*, National Academy Press, Washington, D.C., p. 215-229).

³American Geophysical Union members represent the fields of atmospheric and ocean sciences, solid-earth sciences, hydrologic sciences, and space sciences. See <<http://www.agu.org>>.

⁴National Research Council, 1998, *Review of NASA's Distributed Active Archive Centers*, National Academy Press, Washington, D.C., 233 pp.

⁵For a list of ESIP products and services, see <http://www.esipfed.org/data_center/ps_brochure.pdf>.

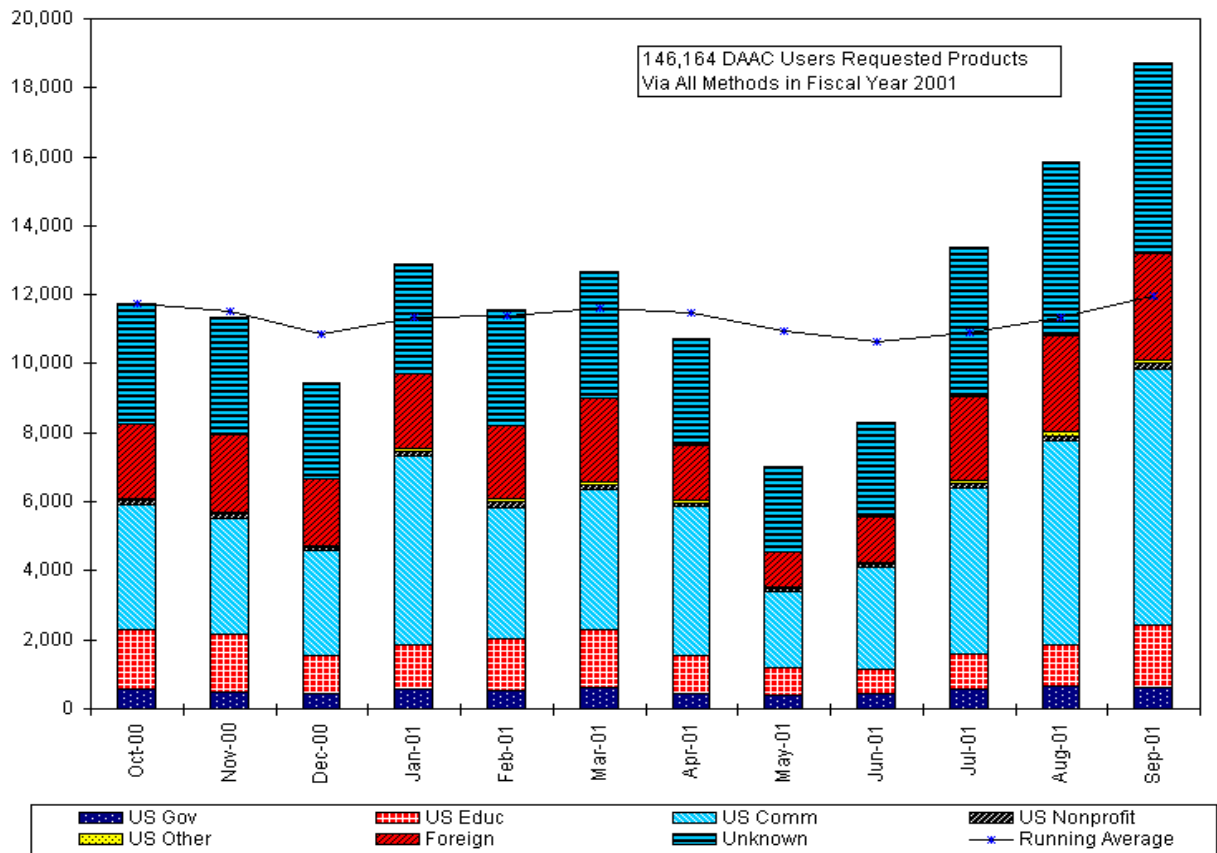


FIGURE 3.1 Profile of DAAC users in FY 2001, as determined by electronic address extensions.
 SOURCE: See <<http://ulabibm.gsfc.gov/~stats/charts/fy01>>.

military, and 35 percent other.⁶ Since their establishment, the ESIPs have created more than 200 new information products and served more than 16,000 users each quarter.

In response to the task group's survey, only three centers identified potential new user groups: Goddard Space Flight Center (GSFC) DAAC (students and disaster-warning organizations); Socioeconomic and Applications Center (SEDAC) DAAC (journalists, librarians, and policy advisors); and SDAC (heliospheric scientists). The other centers appear to seek only marginal changes in their user base, either by developing and distributing specialized materials and sampler data sets (e.g., High Energy Astrophysics Science Archive Research Center [HEASARC], Planetary Data System [PDS], and Langley Space Flight Center DAAC), or by "advertising" their holdings and services at scientific conferences and through newsletters (e.g., NASA/Infrared Processing and Analysis Center Extragalactic Database, National Space Science Data Center, and National Snow and Ice Data Center DAAC). A number of centers (e.g., Infrared Science Archive [IRSA] and EROS Data Center DAAC) rely on users to find them by word of mouth, citations in journal articles, or Web search engines. New NASA policies in some space science disciplines (e.g., astrophysics) have also broadened the user base over the past several years. These policies include minimal proprietary periods for data, guest investigator programs on nearly all satellites that allow scientists to design their own observing programs, the coupling of support for analyzing data with the award of observing time, online access to data, and support for archival research. Finally, advances in data storage and distribution technologies, coupled with the growth of the World Wide Web, have broadened the user base by increasing the amount of online data and by making data easier for everyone to find.

AVAILABILITY OF EARTH AND SPACE SCIENCE DATA

All earth science data held at the DAACs (including, eventually, some scientific products created by the ESIPs) are available through the EOSDIS Data Gateway, and a considerable amount of space science data is available electronically through the individual active archives or mission Web sites.⁷ Most of the remaining space science data is available on media or in various forms from the PIs. Such data, particularly in solar and space physics, are commonly less accessible than are data held in active archives, because distribution and user services are not explicitly supported, and rewards to scientists come from publishing papers, not from depositing organized data sets into national data centers.⁸ Consequently, many PI data sets are not fully exploited, and much new information remains to be uncovered. Moreover, PI data sets are very much at risk when mission resources end.

⁶Briefing to the task group by John Townshend, past president of the ESIP Federation, University of Maryland, July 30, 2001.

⁷The amount of space science data that is network-accessible depends on the archive. For example, only 16 percent of NSSDC holdings were network-accessible by December 31, 2000, although some of these are available through the active archives (National Aeronautics and Space Administration, 2000, Annual Statistics and Highlights Report for the National Space Science Data Center. NSSDC/WDC-SI 2001-01, 33 pp.).

⁸Final Report of the Task Group on Science Data Management to the Office of Space Science, NASA, Jeffrey Linsky, chair, October 23, 1996, 61 pp. Recent OSS solicitations require that PI-held data be deposited in an internationally accessible data bank within two months of collection (e.g., see Announcement of Opportunity for the Solar Dynamics Observatory and Related Missions of Opportunity, <http://research.hq.nasa.gov/code_s/nra/current/AO-02-OSS-01/index.html>).

A task group review of Web sites suggests that all of the active archives are usable by scientists. However, the data are not all easily accessible. Determining which data are appropriate for a specific need can be difficult. The Web pages offer detailed instructions, but it takes time and practice to learn how to navigate them to obtain the right data and tools. Those scientists who are or expect to be frequent users will invest the time in learning the system, but casual users may not. Similarly, descriptions of the data and tools may be insufficient for users who are not closely involved with a particular project. Indeed, insufficient documentation of satellite data has been a widespread and persistent problem.⁹ In general, obtaining data does not seem to be a problem, although users may have to retrieve larger volumes of data than are needed or convert the data into more useful formats. These observations echo those of a recent NRC report, which indicates that a majority of science users found access to be “somewhat easy” to “very easy.”¹⁰ Enhancements in bandwidth, better documentation, and increases in online data sets held in publicly accessible data facilities should improve access to scientific data.

In addition to providing access to scientific data, most centers provide easily found links to education and outreach materials that are readily understood by nonscientists (see Table 3.1). For example, HEASARC provides information about black holes and supernovas in nontechnical terms, and the Physical Oceanography DAAC (PO.DAAC) provides a tutorial and time-series data on El Niño and La Niña events. Given that a mission of all the active archives is to serve educators and the general public, it would be valuable if sites that are designed for professionals, such as the Multi-mission Archive at Space Telescope (MAST), the Alaska SAR Facility (ASF) DAAC, and IRSA would also provide prominent links to nonspecialist information.

SCIENTIFIC COMMUNITY USERS

As noted above, the active archives primarily serve the scientific community, and their success depends largely on how useful scientists find the data. The active archives measure user satisfaction through customer feedback via Web sites, user services, and comment cards; citations of data in journals; and user surveys (see Table 3.2). They also infer user satisfaction from informal feedback at conferences and from increases in the number of users and repeat customers. In addition to these metrics, the DAACs have begun to track (1) kudos and complaints and (2) errors, in response to recommendations from a 1998 NRC report.¹¹ The task group notes that these are reasonable measures of satisfaction but that most active archives adopt only some of these measures. For example, only about half of the centers survey their users, and even fewer (Astronomical Data Center [ADC], ASF, PDS, PO.DAAC) were able to provide the results of a recent survey to the task group. The most comprehensive of these surveys was conducted by the ASF DAAC, which queries users about user expectations, data quality, receipt of correct data in a timely manner, quality of user services, and usefulness of interfaces and software. The PDS survey focused on characterizing the user community, data set content and availability, and data analysis languages tools. In both cases, the majority of users describe

⁹See, for example, National Research Council, 1982, *Data Management and Computation. Volume 1: Issues and Recommendations*, National Academy Press, Washington, D.C., 167 pp.; National Research Council, 1995, *Preserving Data on Our Physical Universe*, National Academy Press, Washington, D.C., 67 pp.

¹⁰National Research Council, 1998, *Review of NASA's Distributed Active Archive Centers*, National Academy Press, Washington, D.C., 233 pp.

¹¹National Research Council, 1998, *Review of NASA's Distributed Active Archive Centers*, National Academy Press, Washington, D.C., 233 pp.

TABLE 3.1 User Groups Served Online by Data Facilities

| Data Facility | Home Page | Data for Scientists | Data or Links for Education/General Public |
|----------------------|---|---------------------|--|
| <i>Space Science</i> | | | |
| ADC | http://adc.gsfc.nasa.gov | x | x |
| HEASARC | http://heasarc.gsfc.nasa.gov | x | x |
| IRSA | http://irsa.ipac.caltech.edu | x | |
| MAST | http://archive.stsci.edu | x | |
| NED | http://nedwww.ipac.caltech.edu | x | |
| NSSDC | http://nssdc.gsfc.nasa.gov | x | x |
| PDS | http://pds.jpl.nasa.gov | x | x |
| SDAC | http://umbra.nascom.nasa.gov | x | x |
| <i>Earth Science</i> | | | |
| ASF | http://www.asf.alaska.edu | x | |
| EDC | http://edcdaac.usgs.gov | x | x |
| GSFC | http://daac.gsfc.nasa.gov | x | x |
| PO.DAAC | http://podaac.jpl.nasa.gov | x | x |
| LaRC | http://eosweb.larc.nasa.gov | x | x |
| NSIDC | http://nsidc.org | x | x |
| ORNL | http://www.daac.ornl.gov | x | x |
| SEDAC | http://sedac.ciesin.columbia.edu | x | |

themselves as satisfied with the center. These results agree with a 1998 NRC survey of DAAC users, which found that the majority of respondents are satisfied with the DAACs and judge their performance to be above average.¹²

User surveys are valuable not only because they gauge user satisfaction, but because they identify products, tools, and services needed for improving the usefulness of the data. ASF users identified a long list of tools for making the data more useful, and cited improvements in documentation, interfaces, and technical and scientific support for making the data easier to find and use. PDS users would like to see a better Web interface, finer-grained searching, and the ability to choose nonstandard data volumes. Such information would be beneficial for all of the centers and their users, and the task group strongly encourages all centers to collect this information through regular, comprehensive user surveys.

An indirect measure of user satisfaction is growth in the number of users, although this measure may not be appropriate to centers that serve small, highly specialized user communities. Most of the centers report significant growth in the number of users from FY 1995 to FY 2000. The number of users more than doubled at ADC, HEASARC, and the DAACs and increased by an order of magnitude at MAST and the NASA/Infrared Processing and Analysis Center Extragalactic Database. The number of PDS users stayed constant. Changes in the size of the user community could not be determined for IRSA, which was created in 1999, and the National Space Science Data Center (NSSDC) and SDAC, which do not track numbers of unique users. The numbers referred to here are not directly comparable, because the space centers count users differently from each other and from the DAACs. Nevertheless, it is clear that the user community is growing, suggesting that users are increasingly finding the holdings useful to their work.

¹²National Research Council, 1998, *Review of NASA's Distributed Active Archive Centers*, National Academy Press, Washington, D.C., 233 pp.

User working groups, NASA advisory committees, and senior reviews provide a means of gauging the satisfaction of scientific users, as well as identifying improvements needed by the scientific community. Science advisory committees make specific suggestions for obtaining new data sets, setting priorities for data processing, monitoring data and image quality, developing new tools, and improving user services in the individual active archives. Because advisory committees are dominated by working scientists, they provide an effective mechanism for improving the usefulness of the active archives to the scientific community. Each center has an advisory committee (usually designated user working groups in the earth sciences), and the most effective ones operate independently of the center rather than under it, providing ongoing, critical reviews of archive operations.¹³

The space science senior reviews also provide an important assessment tool for comparing the usefulness of the active archives within a discipline program. The 2000 astrophysics senior review urged strong support for the active archives and commented that the reviewers “saw no other way of assuring that the enormous data troves now being gathered by increasingly sophisticated astronomical missions in space be made rapidly accessible for scientific analysis.”¹⁴ The 2001 senior review of solar physics found that the SDAC is serving its users well but that many of the NSSDC holdings are still poorly documented,¹⁵ thus limiting their usefulness. However, the report notes that the quality of the data and services provided by the NSSDC has greatly improved in recent years and that it remains a cost-effective data resource for the Sun-Earth Connection community. Both the astrophysics and solar physics senior reviews mentioned the importance of finding ways to combine resources from different space missions and different active archives. In order to facilitate interoperability of the active archives, both reviews recommended streamlining data analysis tools, minimizing duplication of effort, and, in the case of astrophysics, providing transparent access to all of NASA’s databases currently being handled by the different centers.

Finally, most of the active archives track journal citations as a measure of scientific usefulness (see Table 3.2), either by consulting scientists directly or by using literature search or abstract services such as the Astrophysics Data System. The number of citations they report varies, ranging from tens to hundreds of citations per year. However, these numbers can be difficult to interpret, because publication counts depend on several factors, including (1) the usefulness of data in scientific research; (2) the aggressiveness of the center in obtaining the information, since many authors do not cite data sets; and (3) the age of the data, with recently collected data being cited more frequently than older data.

Although many of these measures of satisfaction are difficult to interpret in isolation, taken together they suggest that the active archives are serving their science users well. Such a result agrees with the positive feedback that the task group received from NRC standing committees and with the task group’s own experience with the data facilities.

¹³National Research Council, 1998, *Review of NASA’s Distributed Active Archive Centers*, National Academy Press, Washington, D.C., 233 pp.

¹⁴National Aeronautics and Space Administration, 2000, Report of the Senior Review of Origins and Structure and Evolution of the Universe: Mission Operations and Data Analysis (MO&DA) Programs, June 27-29, 17 pp.

¹⁵Senior Review of the Sun-Earth Connection Missions Operations and Data Analysis Programs, August 29, 2001, <http://spacescience.nasa.gov/admin/divisions/ss/SEC_Senior_Review_2001.pdf>.

TABLE 3.2 Methods Used by Data Facilities for Determining User Satisfaction

| Center | Customer Feedback | Advisory Committee ^a | Journal Citations | User Surveys | Meetings | Increase in Users |
|----------------------|-------------------|---------------------------------|-------------------|--------------|----------|-------------------|
| <i>Space Science</i> | | | | | | |
| ADC | x | x | x | x | x | x |
| HEASARC | x | x | x | | x | x |
| IRSA | x | x | | | x | x |
| MAST | x | x | x | x | | x |
| NED | x | x | x | | | x |
| NSSDC | x | x | x | x | x | x |
| PDS ^b | x | x | x | x | x | x |
| SDAC | | x | | | | |
| <i>Earth Science</i> | | | | | | |
| ASF | x | | x | x | | |
| EDC | x | x | x | | | |
| GSFC | x | | | | x | |
| PO.DAAC | x | x | x | x | x | |
| LaRC | x | | x | | | x |
| NSIDC | x | | | | | x |
| ORNL | x | | x | x | | |
| SEDAC | x | x | x | | x | x |

^a Includes user working groups (DAACs), senior reviews (space science facilities), and NASA advisory committees.

^b These methods are not used by all PDS nodes.

NONSCIENTIST USERS

As noted above, nonscientists require information rather than data. The active archives and the education and applications programs provide a wide variety of value-added data products tailored to specific applications. In addition, several offer user support services and search tools geared toward less-sophisticated users. For example, both the SEDAC and GSFC DAAC convert data into geographic-information-system-compatible formats, and the GSFC DAAC offers subsetting and data-mining capabilities so that users can obtain small, manageable chunks of data.

Education Community

NASA's Strategic Plan contains a mandate "to involve the education community in our endeavors to inspire America's students, create learning opportunities, enlighten inquisitive minds," and to "communicate widely the content, relevancy, and excitement of NASA's mission and discoveries to inspire and to increase understanding and the broad application of science and technology."¹⁶ The education community, both formal and informal, is served by the active archives, flight projects, ESIPs, and education and outreach programs within the Space Science and Earth Science Enterprises.

¹⁶National Aeronautics and Space Administration, 2000, *NASA 2000 Strategic Plan*, Washington, D.C., 72 pp.

The key to preparing useful science education products is packaging. In order to serve the education community, NASA has made—and must continue to make—a substantial investment in packaging specific data products to meet the complex needs of a very diverse community. Educators are normally not trained—nor do they have the time—to make use of research data. Both the research data and the software needed to access them vary with mission architecture and science goals, and effective use of mission-specific tools requires training and practice. Common, intuitive, affordable, and easily used visualization tools that work with a wide variety of data sets are needed to serve the education community.

Formal Education

The national science education standards developed by the National Research Council specify age-appropriate content goals for the teaching of science in grades K-12.¹⁷ These standards emphasize the teaching of science through inquiry-based methods. Engaging students in the active process of inquiry can help them develop a deeper understanding of both scientific concepts and of *how* we know what we know about science. Both earth and space science data constitute a rich resource for inquiry-based curricula. An increasing number of schools is able to connect to the Internet, which makes data, images, and tools widely available. Using these resources, students have discovered a supernova, dozens of novae in the Andromeda galaxy, and a new Kuiper Belt object through examination of ground-based data. In the earth sciences, students are using NASA data and images to monitor environmental change at local (e.g., Boreal Forest Watch) to global scales (e.g., Global Learning and Observations to Benefit the Environment Program [GLOBE]).¹⁸ Virtual research expeditions in which students use satellite observations and climate models to study the Earth as a system of interacting components are offered by the Planet Earth Sciences ESIP.¹⁹ Many other types of earth and space data, properly packaged, have at least comparable potential for bringing the thrill of discovery into the classroom.

Informal Education

Earth and space science data play a key role in programs and exhibits at museums, science centers, and planetariums. The audience is large; for example, approximately 28 million visits are made to the planetariums in the United States each year. Astronomical and earth science data offer a rich variety of images that both illustrate important scientific advances and are aesthetically pleasing. Astronomy has long had an obvious appeal, in large part because of the kinds of questions it addresses: Where did we come from? What will be our ultimate fate? Are we alone in the universe? Earth sciences have a great deal of practical importance: characterizing

¹⁷National Research Council, 1996, *National Science Education Standards*, National Academy Press, Washington, D.C., 262 pp.

¹⁸The Boreal Forest Watch (<<http://www.bfw.sr.unh.edu/>>) is the outreach program for NASA's Boreal Ecosystem-Atmosphere Study. In this program, students in grades 9 to 12 collect and analyze data on a Canadian boreal forest and contribute their data to a national archive. The GLOBE Program (<<http://www.globe.gov/>>) is a worldwide education program in which NASA is a partner. Students make a wide range of environmental measurements according to scientific standards, contribute their data to a student data archive, analyze their data and create maps and graphs through the interactive Web site, and collaborate with scientists and other GLOBE students around the world.

¹⁹See <<http://www.planearthsci.com/>>.

the past history of the Earth, mapping current resources, and making predictions about the impact of current decisions on the environment. Space-based data can be used to enrich inquiry-based exploration of the world in which we live in informal settings; to develop interactive “kiosk” applications that enable unguided exploration of space data sets; and to create kits useful in hands-on demonstrations, image-rich presentations, and easily replicated exhibits. For example, the “Dynamic Earth” exhibit at the Discovery Science Center in Santa Ana, California, features topics such as plate tectonics, the sun’s influence on convection within the Earth’s atmosphere and oceans, and the impact of humans on the Earth’s atmosphere.²⁰ The exhibit is supplemented with a formal education component, supplying corresponding materials to students and cultivating teachers. Such museums need not be physical. A digital museum created by the Museums Teaching Planet Earth ESIP in association with the Houston Museum of Natural History provides interactive displays of earth and space science data.²¹

Usefulness of Data to the Education Community

Numerous data products have been created by the active archives and flight projects to serve the education community. However, making interesting data products does not guarantee that anyone will use them. Both the Office of Space Science (OSS) and the ESE are developing metrics to evaluate the success of their education programs, but it will take at least a decade to obtain results. Moreover, both enterprises have commissioned external reviews of parts of their programs. For example, researchers at Cornell University are reviewing the ESE’s informal education program, and University of Arizona researchers have assessed the usefulness of equipment grants for working with remote-sensing data for teaching and training. A trio of reports from Lesley University evaluates the infrastructure and activities of the OSS education and public outreach activities. Among the findings is that even if teachers have access to the Internet, they have difficulty finding information they can understand; most of the material on the Web appears to be aimed at those who are already familiar with space science.²² This kind of external evaluation should be an ongoing part of the education and outreach activities of the active archives and flight projects to ensure the best use of limited resources.

Online Outreach

Since the Web has become an essential source of information for the education community and general public, the first place many people turn when they wish to explore a topic is no longer to the library but rather to online sources of information. Many of the NASA centers, including the Space Telescope Science Institute and Jet Propulsion Laboratory, have developed very effective sites that have won national awards.²³ In addition, the OSS Education and Public

²⁰National Aeronautics and Space Administration, 2000, Earth Science Enterprise 2000 Education Catalog, <<http://earth.nasa.gov/education>>.

²¹See <<http://mtpe.com/mtpe/mtpe.html>>.

²²S.B. Cohen, J. Griffith, J. Gutbezahl, and M. Lynch, 2000, The Office of Space Science Education/Public Outreach Evaluation Report, November 1998-December 1999, Program Evaluation and Research Group, Lesley College, 45 pp. + appendixes.

²³For example, the Chandra X-ray Observatory’s site (<http://www.chandra.harvard.edu/>) was given the Griffith Observatory’s Star Award and was named one of the San Francisco Exploratorium’s Ten Cool sites and a USA Today Hot Site. Other award-winning sites include the Solar Max 2000 Web site (<http://sunearth.gsfc.nasa.gov/max/index.html>), the Cosmic and Heliospheric Learning Center

Outreach Program plans to implement a coordinated electronic dissemination system that ensures that all NASA educational activities and data products are available through appropriate networking technologies (i.e., the Internet and satellite or cable television). It is the nature of Web-based interactions that people expect timely information with frequent updates and changes to the material presented. Responding to this expectation requires an ongoing investment, with resources already in place when especially newsworthy events are likely to occur.

News Media

Discoveries in astronomy and solar system exploration are well covered by the media, and the earth sciences have a similar but underexploited potential for attracting the general public. Some lessons in how to work with the media can be derived from the successful experience of the Space Telescope Science Institute, which has made the Hubble Space Telescope (HST) a household name. Key elements of their success are (1) making a long-term investment in establishing good relations with the media by providing easy access to scientists who can explain the data obtained from NASA missions in understandable terms; (2) developing confidence on the part of the media that the science highlighted in press releases, for example, is accurate; (3) using sophisticated visualization techniques to present data in ways that are suitable for video as well as print and Internet publications; and (4) becoming recognized as a resource for background information and commentary on results in astrophysics, whether or not they come from HST.

In addition to reporting discoveries, the news media itself can develop value-added products and disseminate them to the general public. Meteorologists in the broadcasting industry already provide such services using NOAA data, and the StormCenter ESIP is adding high-resolution NASA imagery to network and local broadcasts.²⁴ A goal of this ESIP is to educate fellow media professionals about the usefulness of certain NASA data.

Commercial Users and Decision Makers

As with the education community, commercial users and decision makers require value-added products tailored to their specific needs. None of the major data facilities focuses on serving these users, so NASA has initiated a number of short-term projects aimed at creating specialized products for a wide range of applications. Chief among them are the ESIPs, RESACs, Infomarts, A Remote Sensing Product Development Partnership for Agriculture (A20/20), and Food and Fiber Applications of Remote Sensing (FFARS) (see Box 3.1 and Table 3.3). The overall program responsibility for all but the ESIPs rests with the Applications Directorate of the Earth Science Enterprise.

(<http://helios.gsfc.nasa.gov>), the Science Education Gateway (<http://cse.ssl.berkeley.edu/segway/>), and the SIRTf Multi-Wavelength Messier Gallery (sirtf.jpl.nasa.gov/Education/Messier/tie.html).

²⁴See <<http://www.stormcenter.com/>>.

BOX 3.1
Selected Programs for Creating Data Products for Nonscientists

Type 3 Earth Science Information Partners (ESIPs) are primarily commercial companies engaged in developing tools for practical applications of earth science data (see Table 3.3). Since their establishment in 1998, the ESIPs (including those that develop science products) have served more than 16,000 users each quarter, including 2,000 unique users from the education community.¹ The Type 3 ESIPs, which received only half of their \$13 million funding from NASA, are expected to become self-sustaining within five years.

The nine Regional Earth Science Applications Centers (RESACs) are operated by academic/industry/government consortia, with the goal of leveraging scientific results, technologies, and data products from the Earth Science Enterprise to address regional resource management and economic policy issues (see Table 3.3). Some of the RESACs are also ESIPs. NASA invested \$14 million in this three-year program in 1999.

A Remote Sensing Product Development Partnership for Agriculture (Ag20/20) is run jointly by NASA and the U.S. Department of Agriculture (USDA); projects are aimed at farming applications. Funding for the 15 projects began in 2001. All of the products are intended to be used to validate and demonstrate remote-sensing solutions that can be used as benchmarks for operational agricultural systems. Another joint NASA-USDA program is Fiber Applications of Remote Sensing (FFARS), which is aimed at creating products relevant to agriculture, forestry, ranges, and natural resource management. There are 16 FFARS projects, each of which receives three years of funding.

Infomarts were created to demonstrate the applicability of Earth Observing System data to the nonresearch community, particularly to policy makers. Initiated in 2000, Infomarts are partnerships between Raytheon (the developer of the EOSDIS core system) and universities and state and local governments. The 11 Infomarts are developing products related to the protection of natural resources, precision agriculture, water resources management, urban planning, and disease management (see Table 3.3). Organizations that host Infomarts were selected by Raytheon and approved by NASA. The annual budget and duration of funding is set by Congress. Funding for the Infomarts grew from \$6 million in FY 2000 to \$23 million in FY 2002.

¹Bruce Caron, President of the ESIP Federation, personal communication, January 2002.

TABLE 3.3 Selected Programs Serving Commercial Users and Decision Makers

| Facility | Applications |
|--|--|
| <i>Regional Earth Science Applications Centers</i> | |
| Wildlands Fire Hazard ^a California State University, Long Beach | Management of fire hazards at the urban-wildlands interface in Southern California |
| California Water Resources Lawrence Berkeley National Laboratory | Water resource management in the western United States |
| Great Plains University of Kansas | Agroecosystem development and planning in the Great Plains |
| Mid Atlantic ^a University of Maryland | Management of land use, coasts, and watersheds in the mid-Atlantic states |
| Midwest Center for Natural Resource Management University of Wisconsin | Management of forest and agricultural resources in the upper Midwest |
| Northeast Applications of Useable Technology in Land Planning for Urban Sprawl ^a University of Connecticut | Land use decision making in four watersheds in the Northeast |
| Northern Great Plains ^a University of North Dakota | Farming and ranching in North Dakota, South Dakota, Montana, Idaho, and Wyoming |
| Southwest Earth Science Applications Center University of Arizona | Use and management of water resources in the Southwest |
| Upper Great Lakes University of Minnesota | Natural resource management in the upper Great Lakes region |
| <i>Type 3 Earth Science Information Partners</i> | |
| Bay Area Shared Information Consortium Mountain View, California | Various applications of earth science and geographic information in the San Francisco and Monterey Bay areas |
| California Land Science Information Partnership California Resources Agency | Various applications, including real-time response and long-term monitoring and planning |
| Earth Data Analysis Center University of New Mexico | Resource management projects focusing on land economics, regional hydrology, and air quality in the upper Rio Grande Basin |
| Environmental Legal Information System University of Maryland, Baltimore County | Legal applications related to the environment |
| Reading Information Technology, Incorporated Reading, Massachusetts | Improving the efficiency of marine operations |
| Scientific Fishery Systems Anchorage Alaska | Improving the efficiency of fisheries |
| TERRA-SIP University of Minnesota | Land and environmental management |
| Terrain Products from EOS Sensor Data Veridian MRJ Technology Solutions, Incorporated | Various applications of mapping and geospatial information |

continues

TABLE 3.3 *continued*

| Facility | Applications |
|--|---|
| <i>Infomarts</i> | |
| University of Arizona, College of Agriculture | Management of natural resources, particularly as they apply to elk population and density |
| University of Arizona, College of Engineering | Management of water resources through snow mapping |
| University of Idaho | Management of water resources by estimating evapotranspiration |
| University of North Dakota | Farming and ranching applications based on near-real-time plant assessments |
| University of Missouri, Columbia | Management of soybean production |
| University of Missouri, Columbia | Local urban development planning |
| University of Hawaii | Disaster management in the Pacific and Indian Ocean regions |
| University of Texas, Austin | Management of Texas droughts and coastal hazards |
| Towson University | Management of the Chesapeake Bay and Maryland coastal bays watershed |
| Interagency Research Partnership for Infectious Diseases | Monitoring of environmental and ecological variables that trigger epidemics |
| Missouri Resource Assessment Partnership | Sustainable forestry |

^a This RESAC also operates as an ESIP 3.

SOURCE: <<http://www.esad.ssc.nasa.gov/resac/>>, <<http://www.esipfed.org/>>, <<http://earth-outlook.east.hitc.com:1500/>>.

NASA is primarily a research and development agency. Consequently, all of these applications programs are expected to become self-sustaining after their initial three to five years of funding.²⁵ How successful they will be in transitioning to the marketplace remains to be seen. It is likely that many products and services will be discontinued because they are not commercially viable. Those that are maintained may become less accessible, since commercial entities must often control access to make a profit.²⁶

Products created by these programs are currently accessible via the Web site of the facility. No comprehensive list of products or metadata exists, although the Applications Directorate has begun to catalog the metadata. To facilitate the process, future solicitations will require that data products comply with Federal Geographic Data Committee (FGDC) metadata standards and register in the FGDC clearinghouse.²⁷ These steps should help future users find information from the applications programs.

By making value-added products and tools more usable for specialized user groups, the ESIPs, RESACs, Infomarts, and similar programs increase the value of NASA's existing

²⁵A joint solicitation for follow-on RESAC, Type 3 ESIP, and Infomart programs is currently being formulated.

²⁶National Research Council, 2001, *Resolving Conflicts Arising from the Privatization of Environmental Data*, National Academy Press, Washington, D.C., 99 pp.

²⁷Ronald Birk, director of the NASA Applications Directorate of the Earth Science Enterprise, personal communication, March 2002.

holdings. How much value they will add is unknown, because all of these programs are still underway and none has been formally evaluated. However, based on the large number of users reported by the ESIPs (see Box 3.1), the value could be significant.

CONCLUSIONS

The task group concludes that NASA has done a good to excellent job in making data available to the research communities that it serves. User committees, other advisory and oversight committees, and a variety of formal and informal contacts with the researchers provide NASA with the information it needs to monitor the effectiveness of its programs to distribute scientific data and to make any needed midcourse corrections. NASA has also made a major commitment to enhancing science literacy and understanding at all levels of the educational system. External evaluation has been initiated by both the Office of Space Science and the Earth Science Enterprise to evaluate the effectiveness of this investment. In general, the space science community sees researchers and educators, including the media, as the primary “customers” of NASA data. Earth science data have potentially a much broader customer base, including commercial users, policy makers, and others. The DAACs do not appear to have either the mandate or the resources to provide extensive custom data sets or user services to nonscientists. Rather, such services are mainly provided by other ESE programs. While evaluation is needed to determine whether this distributed approach is the most effective strategy, it is clear that specific investment in meeting the unique requirements of each of a diverse set of end users will be required in order to maximize the usefulness of NASA data.

Recommendation. NASA planning and project funding should continue to include provisions for the timely generation and synthesis of data into information and the dissemination of this information to the diverse communities of users. This plan should take into account the needs—and the contribution to information generation—of end users, including other federal and state agencies, educational organizations, and commercial enterprises. The plan should include provisions for ongoing assessment of the effectiveness of data transfer and its educational value.

Strategies for Managing Earth and Space Science Data

As shown in Chapters 1 through 3, NASA's space missions are a primary source of data and information for a wide range of earth and space research and applications. However, the following trends pose challenges for managing the data effectively:

1. *The user community is growing in size and diversity.* The increase in users is a measure of success of NASA's active archives, but it also poses a challenge because new user groups commonly require data sets tailored for specific applications. Providing calibrated data is no longer sufficient for meeting the needs of NASA's customer base.
2. *The volume of data is increasing rapidly, with increases of one to two orders of magnitude expected over the next 5 years* (see Figures 4.1 and 4.2). The challenges and the costs of preserving and enabling access to these growing data sets will also grow with time. The larger data volumes will place increasing demands on developing tools for finding data and for extracting small subsets.
3. *Research questions and practical applications increasingly require the integration of a wide variety of data sets, which are commonly stored in differing data models and formats, with lack of agreement on metadata, different resolutions, and different data quality requirements, and in different locations.* The holdings must be well documented with standard formats and metadata standards and available through a common set of querying tools so that it becomes possible for users to integrate data across centers and to combine data from the active archives with data from long-term data centers in order to identify patterns (e.g., environmental influences on galaxy evolution) and monitor long-term variations and trends (e.g., in land use or climate).
4. *Data relevant to NASA-supported research programs may be held by other federal agencies or by other countries. It will be necessary to establish agreements to ensure that these data are also properly curated and made accessible and that the formats and metadata standards are compatible.*

Dealing with these management challenges will require more than simply increasing funding to the active archives, although providing increased funding to the centers is reasonable in many cases. To get the most out of its holdings, NASA will have to reexamine its overall strategy for collecting and managing data. This chapter focuses on the need for a comprehensive strategy for managing earth and space science data; the balance between acquiring, analyzing, and archiving data; and usefulness of federated approaches to managing databases.

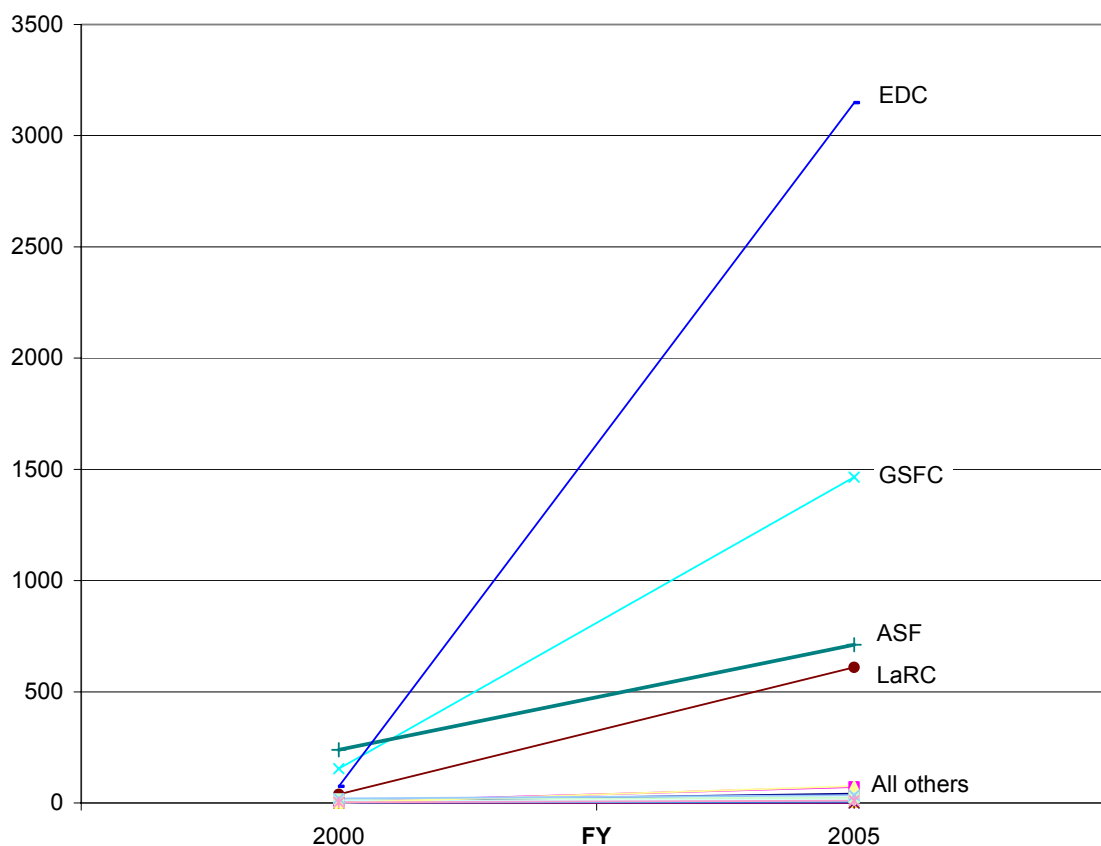


FIGURE 4.1 Projected growth in the volume of data at all active archives and data centers evaluated in this report (see Table 2.1), FY 2000 to FY 2005. Most of the data are held in earth science centers, particularly the Goddard Space Flight Center DAAC (GSFC), EROS Data Center DAAC (EDC), Alaska SAR Facility DAAC (ASF), and Langley Research Center DAAC (LaRC). SOURCE: Data provided by managers of the active archives and data centers.

FIGURE 4.2 Projected growth in the volume of data at all active archives and data centers evaluated in this report except the four centers with the largest holdings (LaRC, GSFC, EDC, and ASF DAACs), FY 2000 to FY 2005. SOURCE: Data provided by managers of the active archives and data centers.

A COMPREHENSIVE APPROACH TO INFORMATION MANAGEMENT

Corporate America has long recognized the importance of its data to both daily operations and long-term corporate viability. Every major corporation has a chief information officer (CIO) who is responsible for all of the corporation's data sets and who frequently has substantial power and budgetary authority. Whereas the collection and exploitation of data are critical to the operation of any modern business, it is not the primary focus of most. However, the task group contends that the collection and exploitation of data are NASA's *main* business. Although NASA has a CIO, that person's primary responsibility is for the business systems maintained by NASA.¹ The enterprises are responsible for overall program planning, including scientific data management, in their disciplines. No NASA-wide mechanism exists for (1) advocating appropriate investment in data management, (2) ensuring that best practices are communicated across the scientific enterprises,² (3) overseeing and evaluating the development of strategic plans for major data initiatives such as the National Virtual Observatory (NVO) and Strategic Evolution of ESE Data Systems (SEEDS), and (4) ensuring the preservation and accessibility of NASA's valuable information resources. The task group believes that this set of responsibilities would be most effectively carried out through the leadership of a single individual with a very high level of training both in a science field related to NASA missions and in information science. For convenience, the person(s) assigned to manage these responsibilities is referred to here as the chief science information officer, or CSIO. This CSIO would carry out the following tasks:

- Provide strategic planning, oversight, and advice concerning the collection, processing, archiving, and dissemination of data and information collected by NASA's space missions.
- Be the advocate for the appropriate balance of investment in data analysis.
- Ensure the preservation and accessibility of valuable space mission data and information.
- Require a data management plan for each mission and monitor its implementation.
- Provide oversight for design and implementation of software, hardware, and database systems for processing and storing NASA's massive data sets.
- Develop a long-term software plan.
- Require interenterprise communication and sharing of successful methods and systems for data management.
- Work out the memorandums of understanding governing access to data from those missions that are carried out cooperatively with other countries.
- Determine how information generated by the space programs of other countries can be accessed and effectively used by U.S. scientists and institutions.

¹The general responsibilities of an agency's CIO are (1) to ensure that information technology is acquired and information resources are managed effectively; (2) to develop, maintain, and facilitate sound and integrated information-technology architecture; and (3) to promote the effective and efficient design and operation of all major information-resources management processes for the agency, including improvements to work processes. See <<http://www.hq.nasa.gov/office/cio/>>.

²NASA has developed procedures and guidelines for reviewing and applying lessons learned from previous missions to avoid the recurrence of mistakes and to share best practices. According to a recent General Accounting Office report, however, NASA managers do not routinely identify, collect, or share lessons. See General Accounting Office, 2002, Better Mechanisms Needed for Sharing Lessons Learned, GAO-02-195, Washington, D.C., 51 pp.

Some of these responsibilities relate to cross-NASA issues, while others are more specific to individual program offices. While a single CSIO is referred to here, it may be that each of the enterprises requires its own CSIO. However, regardless of the administrative structure that is selected, it should be one that supports cross-enterprise communication and cooperation. If there is a single CSIO, that person might appropriately report to the NASA administrator. If there are CSIOs for each enterprise, they should report to the heads of the respective enterprises. The important point is that the CSIO(s) must report to a level within NASA that will provide the support and authority needed to ensure that the CSIO is effective in carrying out the functions identified here.

Just as a major corporation assigns substantial budgetary authority to its CIO, the CSIO(s) should have budgetary authority for end-to-end management of data: collection, analysis, distribution, and long-term maintenance. The issue of balance between the funding for designing and deploying a piece of hardware and funding for collecting, analyzing, and storing the data sets produced by a mission has been addressed in earlier National Research Council (NRC) reports.³ When the cost for the hardware exceeds a particular mission's budget, funds for data analysis may be reduced, particularly in programs with cost caps. As a consequence, data analysis may have to be funded through research and analysis programs, which are also tightly funded and already oversubscribed. The task group proposes an alternative model in which the CSIO(s) would either have the budget for designing the data collection/analysis/dissemination/archiving function for each mission or would have the right of refusal for projects or programs that do not handle the required balance adequately, or both. When trade-offs must be made between hardware and data components, the CSIO(s) would be responsible for ensuring that the mission investment in data management remained adequate for optimizing the scientific return.

Recommendation. NASA should assign the overall responsibility for oversight and coordination of NASA's data assets to a chief science information officer (CSIO) (or alternatively to multiple science officers). The CSIO(s) would provide leadership; long-term strategic planning; and advice on the collection, processing, archiving, and dissemination of data and information collected by NASA's space missions to ensure the preservation and accessibility of these valuable resources. If a single CSIO is named, then this individual should report to the NASA administrator. Alternatively, CSIOs might be appointed for each of the enterprises and report to the heads of the enterprises, but in this case a mechanism should be established to ensure cross-enterprise coordination and communication of best practices.

This recommendation does not imply that NASA should centralize all data aspects of all missions. Rather, a combination of distributed and centralized activities would best serve NASA's scientific programs. For example, a distributed approach to developing software for managing data has proven to be the most cost-effective means for delivering usable software on the timescales required for scientific missions. Similarly, analysis and production of data products should continue to be performed in a distributed manner by scientists, whereas long-term maintenance is probably best handled centrally. The CSIO(s) would be responsible for overseeing the planning for the production of data products and assessing the outcomes, while leaving the actual production to the scientists. One of the charges to the NASA CSIO(s) should

³National Research Council, 2000, *Assessment of Mission Size Trade-offs for NASA's Earth and Space Science Missions*, National Academy Press, Washington, D.C., p. 14, and references therein.

be, however, to facilitate interenterprise sharing of methods and systems for data management. NASA has accumulated a wealth of space and earth science data that are archived, managed, processed, and distributed in a variety of methods with different levels of success. With an agency overview, the NASA CSIO(s) could seek out data management successes from one mission and apply them to future activities. It is possible that the Space Science and Earth Science Enterprises could benefit by emulating each other's successes.

The CSIO(s) will face many challenges, but none so daunting as the design and implementation of software, hardware, and database systems for processing and storing NASA's massive data sets. Corporate CIOs have a range of choices of suitable database systems, analysis software, and so on, but there is minimal commercial interest in producing software specifically for use by NASA. However, creating custom software tailored to meet very specific requirements also presents problems, as NASA and other federal agencies such as the Federal Aviation Administration and Internal Revenue Service have discovered (see Chapter 2). Consequently, one of the first tasks that the CSIO should undertake is the development of a long-term software strategic plan. To the maximum extent possible, NASA should make use of commercial off-the-shelf software in executing its mission in order to maximize cost-effectiveness. To assist with evaluating options, the CSIO(s) should create an advisory panel composed of instrument scientists, computer scientists, an electronic-records expert from the National Archives and Records Administration, and CIOs from major corporations and government organizations with very large and complicated data sets (e.g., Wal-Mart, Sears, Sabre, and USGS). The importance of including corporate CIOs on the panel cannot be overemphasized. In order to be successful financially, corporations today rely on their CIOs to acquire and exploit their data sets to the maximum extent possible. The techniques they use would be invaluable to the success of the NASA CSIO's mission.

ISSUES OF BALANCE: ACQUISITION, ANALYSIS, AND ARCHIVING

The goal of a scientific mission is to obtain the greatest scientific yield for a fixed amount of resources. The tasks that must be supported within a fixed budget are the following:

- A. Pre-mission science and technology definition;
- B. Mission development, flight, and operations; and
- C. Analysis of observations, modeling, archive, and education and public outreach.

Optimizing the scientific return from a mission necessarily involves optimizing the relative investment in these three broad categories of mission activities. The current distribution—about 75 percent is spent on category B and 25 percent is spent on categories A and C together (see Table 4.1)—yields much good science. However, in order to optimize the science per dollar, the relative fraction of funds spent in each category will necessarily depend on the mission. As noted earlier, even after the fractions are fixed, cost overruns during mission development may threaten the investment in data analysis. It is critically important that trade-offs among capabilities that are inevitable in missions and programs with a fixed budget result in a balanced investment in hardware and software that optimizes the overall scientific yield from the mission.

TABLE 4.1 Funding for Mission Development, Operations, and Data Analysis

| Activity | OSS Budget (\$M) | | | ESE Budget (\$M) | | |
|---------------------------------|------------------|---------|----------------------|------------------|---------|----------------------|
| | FY 1995 | FY 2000 | FY 2005 ^a | FY 1995 | FY 2000 | FY 2005 ^b |
| Development | 1,411 | 967 | 1,425 | 737 | 643 | 451 |
| Operations | 67 | 79 | 384 | 71 | 48 | 251 |
| Research | 141 | 250 | 320 | 269 | 371 | 439 |
| Data Analysis | 214 | 291 | 513 | | | |
| EOSDIS | | | | 221 | 279 | 69 |
| Other ^a | 199 | 608 | 1,173 | 46 | 102 | 69 |
| Total | 2,032 | 2,195 | 3,815 | 1,344 | 1,443 | 1,279 |
| A + C (percentage) ^c | 17 | 25 | 22 | 20 | 26 | 34 |
| B (percentage) ^c | 83 | 75 | 78 | 80 | 74 | 66 |

^a The Deep Space Network is scheduled to be transferred from the Office of Space Flight, greatly increasing the operations budget of the Office of Space Sciences.

^b EOSDIS will be split between operations and development in FY 2003, and the ground network activity will be transferred from the Office of Space Flight into mission operations, greatly changing the operations and EOSDIS budgets.

^c A = pre-mission science and technology definition; B = mission development, flight, and operations; C = analysis of observations, modeling, archive, and education and public outreach.

SOURCE: Joseph Bredekamp, Senior Science Program Executive/Information Systems, Office of Space Sciences, and Martha Maiden, Code YF Data Network Manager, Earth Science Enterprise.

Data Analysis Funding

The adequacy of data analysis funding for space missions has long been a concern of the scientific community.⁴ These concerns are summarized below.

1. Data analysis funding decreased throughout the 1990s.

A 1998 NRC report on NASA's research and data analysis programs found that the fraction of the total science-related budget that was allocated to research and data analysis fell by at least 30 percent over the period 1991 to 1998.⁵ In response, the Office of Space Science (OSS) planned to "reallocate current budgets and to seek funds for new projects that will provide selected increases in data analysis funding at an overall rate of 8% per year."⁶ Budget numbers provided to the task group showed that funding for space science data analysis has increased from about \$140 million in FY 1999 to over \$190 million in FY 2002.⁷ Moreover, projections to FY 2005 suggest that the trend of declining data analysis funding has been reversed in more recent years. The highly aggregated data in Table 4.1 show that OSS data analysis funding is projected to increase between FY 1995 and FY 2005, both in absolute and percentage terms. Data analysis funds grew from about 10 percent of the total budget in FY 1995 to about 13 percent in FY 2000 and increased from being equivalent to about 15 percent the size of mission

⁴For example, see National Research Council, 1982, *Data Management and Computation. Volume 1: Issues and Recommendations*, National Academy Press, Washington, D.C., 167 pp.

⁵National Research Council, 1998, *Supporting Research and Data Analysis in NASA's Science Programs: Engines for Innovation and Synthesis*, National Academy Press, Washington, D.C., pp. 51.

⁶Interim Assessment of Research and Data Analysis in NASA's Office of Space Science, letter to Edward Weiler, Associate Administrator for Space Science, National Research Council, September 22, 2000.

⁷Briefing to the task group by Gunter Reigler, director, Research Division, Office of Space Science, January 31, 2001.

development budgets in FY 1995 to about 30 percent as large as development budgets in FY 2000. The fractional size of the data analysis budget remains about the same in projections to FY 2005 of the growth in the total OSS funding. The trends in the earth sciences are not as clear, because research and data analysis are not separated in the ESE budget. However, the fraction of the budget devoted to research appears to be growing (Table 4.1).

The difficulty in interpreting these budget numbers underscores an important conclusion of a recent NRC report: “The fragmented budget structure for R&DA [research and data analysis] makes it difficult for the scientific community to understand the content of the program and for NASA to explain the content to federal budget decision makers”.⁸ The OSS developed plans in 2000 to establish a uniform procedure for collecting data,⁹ and the task group encourages them to continue this process. However, it is too soon to evaluate the results of the efforts so far.

2. Even if an adequate level of data analysis funding is planned, the funds are not always preserved to the end of the mission.

The generally tighter mission budgets of the past few years, coupled with the fact that data analysis typically comes at the end of a mission when project funds are near exhaustion, make it difficult to preserve funding for data analysis.¹⁰ The loss of data analysis funding can have a greater impact on small missions than on large missions. According to a recent RAND Corporation report, which analyzed a set of small science missions, the resources devoted to scientific analysis averaged only 1.6 percent of the total mission cost.¹¹ Given that targets for data analysis are generally an order of magnitude higher, it is unlikely that this level of funding achieved the maximum scientific return.

3. If extensions in data analysis are required, funding must be obtained from the science programs, which are already oversubscribed.

The period of data analysis often has to be lengthened because (1) software delays or unforeseen calibration problems prevent timely delivery of data to the user community; (2) unanticipated discoveries lead to new lines of research equal in importance to the initial goals; or (3) the mission is extended, sometimes for many years, because it continues to collect high-quality data at a small incremental cost or because longer-term monitoring proves important. Some funding to analyze data in this lengthened collection period is available through competitive grants from the science program offices. Two factors provide some guidance as to the adequacy of funding for these areas: (1) the fraction of proposals submitted that can be funded and (2) the quality of the rejected proposals. The task group’s experience is that a 3:1 oversubscription rate is about optimum.¹² If the oversubscription rate is significantly higher, many excellent proposals are rejected; if the oversubscription rate is much lower, choice is

⁸National Research Council, 1998, *Supporting Research and Data Analysis in NASA’s Science Programs: Engines for Innovation and Synthesis*, National Academy Press, Washington, D.C., pp. 67-68.

⁹Interim Assessment of Research and Data Analysis in NASA’s Office of Space Science, letter to Edward Weiler, Associate Administrator for Space Science, National Research Council, September 22, 2000.

¹⁰National Research Council, 2000, *Assessment of Mission Size Trade-offs for NASA’s Earth and Space Science Missions*, National Academy Press, Washington, D.C., 91 pp.

¹¹L. Sarsfield, 1998, *The Cosmos on a Shoestring*, RAND, p. 105.

¹²The oversubscription rate of OSS observing and data analysis proposals is 2:1 to 6:1. Presentation to the task group by G. Reigler, director of the Research Division, Office of Space Science, January 31, 2001.

limited. A more quantitative measure could be provided by the proposal-review panels. As part of their review, a panel could identify the division point between programs that are likely to yield excellent science and those that will lead only to modest gains. If the number of excellent proposals substantially exceeded the number that could be funded, an increase in funding would be warranted.

The senior reviews conducted in astrophysics, planetary science, and the Sun-Earth connection programs provide another mechanism for identifying where additional investments in spacecraft operations and data analysis after the prime mission phase are likely to yield important scientific returns. The senior reviews take a systems approach to evaluations. In the case of a recent senior review of the Sun-Earth Mission Operations and Data Analysis program, factors taken into account included (1) the health and status of each spacecraft and payload, (2) the scientific strengths of proposed programs, (3) the relevancy to other NASA missions, (4) the accessibility of scientific data products to principal investigator teams and outside investigators, and (5) the record for education and public outreach.

4. The task group could not identify a systematic procedure for determining the balance of funding between the flight programs and the associated research and data analysis, especially across science programs.

In a recent assessment of NASA's research and data analysis programs, a 1998 NRC report recommended that each science office do the following:

- Regularly evaluate the balance between the funding allocations for flight programs and the research and data analysis required to support them;
- Regularly evaluate the balance among various subelements of the R&DA program; and
- Use broadly based, independent scientific peer review panels to define suitable metrics and review the agency's internal evaluations of balance.¹³

In response, the OSS instituted a regular process of senior reviews of the research grants program. Senior reviews provide a mechanism for evaluating programs within a given discipline and within a fixed budget. They also provide a mechanism for terminating programs. Many astrophysics missions, for example, are long-lived, and the costs of operations and data analysis are substantial. Indefinite operation of *all* functioning satellites cannot be accommodated within available budgets. A mechanism already in place for considering issues of balance early in the development of individual missions is the non-advocate review, which takes place before a mission is funded and evaluates all aspects of the mission life cycle (Appendix B). As noted above, the CSIO(s) may have a role in shielding data analysis budgets from overruns that occur in missions after the non-advocate review is completed. While both senior reviews and non-advocate reviews play important roles within NASA, neither is designed to address issues related to the overall budget or issues of balance across disciplines, or between new missions, extended missions, and postmission data management. Senior reviews evaluate only missions that are underway and delivering data. The non-advocate reviews address only a single mission and do not provide program-wide direction. Whatever process NASA chooses for addressing these balance issues, it should be one that (1) is open and engages the research community, (2) is

¹³National Research Council, 1998, *Supporting Research and Data Analysis in NASA's Science Programs: Engines for Innovation and Synthesis*, National Academy Press, Washington, D.C., pp. 3-4.

carried out on a regular and systematic basis, and (3) is conducted early enough in the planning cycle to effectively influence mission and program priorities.

Timeliness of Data Analysis

When a new mission begins collecting useful data, it is essential that these data be analyzed quickly to discover errors or scientific results needed for follow-on missions. Usually these data are unique and lead to important insights that require rapid follow-up, especially in the case of short-lived missions. To accomplish this, a software system must be in place at launch that is reasonably mature and can provide high-quality, calibrated data products. This goal in turn requires that adequate resources be devoted to the development of the data system, beginning at very early stages in the project.

Unforeseen problems often arise after launch (e.g., calibration changes), but such issues can be addressed more quickly if the basic data-processing package has already been developed and tested. The timely development of software is so critical that it should be properly funded even if it leads to a reduction in capabilities of the flight hardware. Budgets for mission operations and data analysis are usually separated from those for mission development. This makes it difficult to make trade-offs that will optimize the overall knowledge return. However, as suggested by recent program solicitations in both the earth and space sciences, this situation may be changing. For example, proposals submitted to the OSS Explorer program and the ESE Earth System Science Pathfinder program must encompass all mission phases, including concept study, definition and preliminary design, detailed design, development, mission operations, data analysis, data publication, and delivery of data and metadata to an appropriate archive.¹⁴ The task group encourages NASA to adopt this approach for all its earth and space science programs.

Recommendation. Budgets for mission operations and data analysis should be included as an integral part of mission and/or program funding. Reviews, including NASA's nonadvocate review, which is required to authorize project funding, should include assessment of the data analysis elements, including archiving and timely provision of data to users. While reviews of some projects already follow this recommendation, its implementation is not uniform across all NASA programs. The appropriate balance between hardware and software investment is best determined jointly by NASA managers and the user communities involved in the mission.

FEDERATED DATABASES

The ongoing revolution in data collection, storage, and analysis of large data sets will challenge scientists by presenting new opportunities to combine the results from different types of measurements to analyze complex problems from a systems point of view. Disciplines ranging from earth science to astrophysics are actively exploring techniques for providing (1) fast access to geographically distributed data sets through standard, easy-to-use interfaces; (2) seamless interoperability of large data archives; and (3) a usable base of information for scientific

¹⁴For example, see Explorer Program Medium-class Explorers (MIDEX) and Missions of Opportunity, <http://research.hq.nasa.gov/code_s/nra/current/AO-01-OSS-03/index.html>; Earth System Science Pathfinder (ESSP) Missions, Announcement of Opportunity, <<http://essp.gsfc.nasa.gov/opportunity.html>>.

explorations. Federations of data systems are a possible mechanism for addressing these requirements. Federated data systems are most likely to succeed if they are guided by the principles that have proven sound in the context of federated corporations:

- Power should be placed at the lowest possible point in the structure;
- Interdependence distributes power and avoids the risks of a central bureaucracy;
- An effective federation needs a common language and laws, and a uniform way of doing business; and
- Participants in a federation must recognize their dual citizenship—members in the overall federated structure, but with substantial local autonomy.¹⁵

Governance—the mechanisms by which the participants share in the design, implementation, management, and operation of the federation—is the most important function for the organization’s future.¹⁶

Some federated data systems already exist: for example, the Planetary Data System and the Earth Science Information Partners (ESIPs). Other ambitious efforts, such as the NVO, have received initial funding. And others, such as SEEDS, are in the planning stages.

The National Virtual Observatory

Astrophysics is entering the era of “precision cosmology.” Over the next decade or two, astronomers expect to be able to characterize the size and evolution of the primordial fluctuations that formed the seeds of the structure in the universe, observe galaxies in the earliest stages of formation and test models of how they formed, determine the nature and distribution of both baryonic and dark matter, and characterize the dark energy as a function of the age of the universe. Achieving these objectives will require the collection and integration of petabytes of data from space and ground surveys, each measuring different variables and observing different regions. The NVO, one of the highest priorities of the recent astronomy and astrophysics survey,¹⁷ will develop mechanisms to federate collections of data and information for an entire scientific discipline.

The leadership of the astrophysics community in developing new techniques for data mining has been recognized by the National Science Foundation (NSF), which recently funded (\$10 million over five years) a broad-based effort to create a framework for the NVO. Participants in the proposal included ground- and space-based data centers and key players in the university community. Both the astronomical and computer science communities played an active role in devising the implementation plan. The NVO is predicated on seamless access to ground- and space-based data, and key next steps for NASA are the following: (1) to coordinate closely with the NSF-funded effort to develop the framework for the NVO, (2) at each data node, to identify costs of making extant space-based data compliant, and (3) to develop and invest in a long-term

¹⁵C. Handy, 1992, Balancing corporate power: A new federalist paper, *Harvard Business Review*, November-December, pp. 59-72.

¹⁶National Research Council, 1998, *Toward an Earth Science Enterprise Federation: Results from a Workshop*, National Academy Press, Washington, D.C., 51 pp.

¹⁷National Research Council, 2001, *Astronomy and Astrophysics in the New Millennium*, National Academy Press, Washington, D.C., 276 pp.

strategic plan that builds on the NSF framework effort and the existing investment in space-based data centers to meet the scientific requirements of the space science community.

An important element in NVO planning is the emphasis on developing bottom-up frameworks and toolkits to provide integrated services on whatever scale is appropriate to user needs and the scientific questions being asked. The NVO is *not* an effort to integrate all astronomical services via top-down control. The intention is to build the NVO as a *science-driven*, community effort with most of the funding distributed through peer review.

The NVO plans to accomplish the following:

- Establish a common systems approach to data pipelining, archiving and retrieval that will ensure easy access by a large and diverse community of users and that will minimize costs and times to completion.
- Enable the distributed development of a suite of commonly usable new software tools for querying, correlation, visualization, and statistical comparisons.
- Utilize high speed networks that will provide the connectivity among active archives and terascale computing facilities.¹⁸

Each institution participating in the NVO will maintain control over its individual data sets but will conform to metadata standards and protocols that are extensible far into the future. With properly designed interfaces, it will be possible for anyone in the community to add analysis tools and data facilities. Interoperability of such a distributed system will require a core management group that maintains standards and tight communications while supporting distributed research and development. Although these goals are challenging, astronomers have an established track record of operating in this manner.

The NVO is intended to be evolutionary so that it can respond to changing opportunities and to developments in both hardware and software. Fortunately, processing, storage, and networking are continuing to improve at an exponential rate, so it is likely that the hardware will keep pace with the growing volumes of astronomical data. Bandwidth remains a limiting factor, and so for the foreseeable future it is likely that the computation capabilities need to be close to the data so that large data sets do not have to be moved.

The NVO will also take advantage of the development of grid technology, which is being widely embraced by many fields, including medical technology, earth sciences, high-energy physics, and astronomy. In fact, the inclusion of current grid technology in astronomy in the United States is being accomplished in large measure through the NVO. Grid technology allows not only access to remote data facilities but also “single sign on” remote access to computing and analysis facilities.¹⁹

The NVO has established intimate links to the high-energy physics grid program (GriPhyN) and to the information-technology community that is responsible for the development of GriPhyN. One of the principal architects of the high-energy physics grid is also leading the development of the grid architecture for the NVO. The close cooperation between the GriPhyN and NVO projects will ensure that the astronomy community, through the NVO, will have access

¹⁸R.J. Brunner, S.G. Djorgovski, and A.S. Szalay, eds., 2001, *Virtual Observatories of the Future*, Astronomical Society of the Pacific Conference Series, San Francisco, California, p. 357.

¹⁹For more information on grid technologies and collaborations, see the Particle Physics Data Grid at <<http://www.ppdg.net/>> or the Global Grid Forum at <<http://www.globalgridforum.org/>>.

to current grid-based facilities that are also compatible with the grid networks being established in other fields.

The astrophysics senior review held in June 2000 stressed the importance of providing a coherent data system and recommended that NASA continue to examine what services such a system might realistically be expected to provide, how it might be maintained at the cutting edge of available computational and communications technologies, and what the appropriate trade-offs are between the costs of providing these services and investments in new missions.

While the astrophysics community is providing pioneering leadership in this field, other NASA-supported disciplines are beginning to explore ways of providing similar capabilities. For example, plans have also been developed for a prototyping study for a virtual solar observatory, modeled after the virtual observatory for astrophysics, and a recent senior review of the Sun-Earth Connection program recommended funding for the initiation of the virtual solar observatory.²⁰ These plans are consistent with an earlier recommendation made by an NRC task group on ground-based solar research.²¹

NSF and NASA should collaborate on the development of a distributed data system with access through the World Wide Web. Such access requires easily searchable catalogs, user-friendly access software, and the capacity to handle large volumes of data. A number of organizations, both in the United States and abroad, have taken the initiative to preserve and provide data sets online. Acknowledging the importance of providing data to the community, the task group encourages the cooperation of observatories and institutions, especially NSF and NASA, in efforts to archive and ensure access to their data. In fact, the task group believes that provisions for data archiving and distribution should be an integral part of planning for future observing facilities. In developing these plans, the space science community should take into account the lessons learned from similar endeavors in the earth sciences, such as the ESIP Federation.

Planetary Data System

The Planetary Data System (PDS) has already taken initial steps to achieve the same goals as the NVO by combining geographically distributed active archives, which store data under the supervision of scientific experts, with centralized project management and system engineering. A PDS management council, which includes the nodes managers as well as the overall project manager and system engineer, makes the major decisions. Nomenclature and data formats have been standardized for all data sets, and all archived data are peer reviewed. While PDS provides its users with high-level catalogs for searching for data, this capability is not yet integrated seamlessly across all nodes. In the future, the PDS plans to develop this capability, store more data online, and increase automation of the validation and ingest processes so that mission data can be archived more quickly and made available sooner. A system is currently being implemented for online distribution of all Mars data, beginning in October 2002 with data from the 2001 Mars Odyssey mission.²²

²⁰National Aeronautics and Space Administration, 2001, Final Report of the Senior Review of the Sun-Earth Connection Mission Operations and Data Analysis Programs, 27 pp.

²¹National Research Council, 1999, *Ground-Based Solar Research: An Assessment and Strategy for the Future*, National Academy Press, Washington, D.C., 47 pp. + 11 appendixes.

²²Elaine Dobinson, PDS manager, Jet Propulsion Laboratory, personal communication, March 2002.

Federation of Earth Science Information Partners

The ESIP Federation was created in response to difficulties of the EOSDIS system in responding to rapidly evolving technologies that, among other things, could improve both access and usefulness of ESE data, particularly for non-EOS communities.²³ There are four types of ESIPs, each serving a distinct user community. Type 1 ESIPs (the DAACs and NOAA's National Climatic Data Center) are responsible for standard data and information products. Type 2 ESIPs produce innovative science information products and services, primarily for the global change and earth science communities. Type 3 ESIPs (applications data centers) provide data products specialized for practical applications by nontraditional user communities, including teachers, students, policy analysts, and for-profit businesses. The type 2 and 3 ESIPs were chosen through a competitive proposal process. Finally, type 4 ESIPs are sponsoring agencies (currently only NASA) of the federation.

The ESIPs are an experiment in creating and governing a federated system of heterogeneous units, driven by competition, with each unit relatively small, manageable, and able to respond to changing scientific and technical opportunities. The ESIPs developed their own governance structure in 2000, and the federation became a not-for-profit organization in 2001.²⁴ Ten new partners have joined since the federation was created in 1998.

The objectives of the federation are (1) to increase the diversity and breadth of users and uses of earth science data, information, products and services; and (2) to explore new ways to provide data and information operability.²⁵ The first objective is being met by providing services in a wide range of application areas, including land management, commercial fishing, precision farming, K-12 instruction, weather, ranching, urban planning, and energy (see Chapter 3). The second objective is being met by providing catalog-level searching, distributed data exchange, and data discovery and access prototypes. More than 65 new information services are being provided, either by individual ESIPs or by self-organized clusters of ESIPs.²⁶ However, interoperability at the data level has not yet been achieved. Clearly, the federation has achieved many positive things, but a formal evaluation of its success has yet to be done.

Federation concepts are also being incorporated into plans for SEEDS. One of the objectives of the SEEDS program is to "establish a unifying framework of standards, core interfaces and levels of service to facilitate access to data and information as provided by a distributed, heterogeneous network of data systems and service providers."²⁷ An ESIP cluster is assisting with this issue, as well as providing metrics about design and performance, and is looking for ways to leverage current capabilities and expertise in existing data systems.²⁸ SEEDS is still being formulated, so the task group cannot comment on the adequacy of the planning to meet this objective. However, it can comment on the importance of the objective itself. The task group believes that creation of a federated, distributed system of active archives should indeed be a key

²³National Research Council, 1998, *Toward an Earth Science Enterprise Federation: Results from a Workshop*, National Academy Press, Washington, D.C., 51 pp.

²⁴Bruce Caron, president of the ESIP Federation, personal communication, February 2002.

²⁵Briefing to the task group by John Townshend, past president of the ESIP Federation, University of Maryland, July 30, 2001.

²⁶There are currently 11 clusters, each of which includes 4 to 13 ESIPs, working on cross-cutting issues. See <<http://www.esipfed.org/business/clusters/index.html>>.

²⁷Briefing to the task group by Steven Wharton, NewDISS program formulation manager, July 30, 2001.

²⁸See <<http://www.esipfed.org/business/clusters/newdiss/index.html>>.

component of the SEEDS program and that much can be learned from the approaches being prototyped and evaluated by the space science community and by the ESIP Federation.

Recommendation. NASA should encourage efforts by the scientific community to develop plans for federations of data centers and services that would enable complex querying, mining, and merging of data from different instruments and missions in order to answer complex, large-scale scientific questions.

- **The National Virtual Observatory, an astrophysics project funded recently by the National Science Foundation (NSF), will develop the architecture, standards, and so forth for creating a distributed system of data centers that can be cross-accessed and queried in a transparent manner by users. NASA should coordinate with the NSF-funded work on the NVO, which is predicated on seamless joint access to ground- and space-based data, to ensure that space data are compliant with NVO standards.**

- **NASA should encourage close communications among the groups operating or developing federated systems in order to transfer best practices among its various scientific programs.**

- **The successful implementation of methods for making complex queries of multiple databases is likely to be technically challenging and costly. The level of appropriate investment by NASA in federated data systems should be evaluated at regular intervals and should be based on (1) the importance of the scientific questions that can be addressed through the simultaneous mining of multiple databases, (2) demonstrated scientific return from past investments, and (3) the readiness of computational and communications technology to support data mining.**

ELEMENTS OF EFFECTIVE DATA MANAGEMENT

In examining the various approaches to archiving and dissemination, the task group has identified a number of elements of the overall data management system that have proven to be important in meeting the expectations of the scientific community. These elements are listed below and should be included in planning for future missions and facilities:

- Archives and data centers should have (1) scientists on staff with a strong background in the scientific discipline being supported, and (2) scientific working groups to help set priorities for acquiring, managing, and discarding data.

- Prelaunch funding should be provided for software development to ensure the timely development of pipelines for processing newly acquired data.

- Multiyear funding should be provided for research, including research using archived data, on the basis of quality of the proposals received. A recent senior review of extended planetary missions, for example, noted the success of the archival research programs maintained in astrophysics and suggested that these programs might profitably be emulated by the Planetary Data System.

- Guest investigator programs should be established to allow the community to conduct research not planned by the initial project teams.

- Early and open access to data should be provided to permit follow-on proposals to take advantage of new discoveries.
- A mechanism should be established (such as the senior reviews in space science) for making trade-offs among operations of long-lived missions and operations of active archives and data centers in a way that reflects the scientific merit of the range of possible investments.

The importance of managing data and information from NASA's space missions will only continue to grow in coming years. Data volumes are increasing, both because of the accumulation of data from a steadily growing number of space missions and because improvements in technology have enhanced the data rates from individual missions. Maintaining the data in forms that are readily accessible and that meet the needs of very diverse user communities presents intellectual challenges that are at least the equal of the challenges of building and launching hardware into space. NASA can become a leader in developing the techniques and tools for querying and mining large nonproprietary data sets. Playing this leadership role will require a new emphasis on software management; rigorous review of the balance between investments in software and hardware to optimize the science return from both individual missions and suites of missions; and development of new techniques for exploring and intercomparing data contained in a distributed system of active archives, data centers, and data services located both in the United States and abroad.

Appendixes

Appendix A

Letter of Request

National Aeronautics and
Space Administration
Headquarters
Washington, DC 20546-0001



MAR 22 2000

Reply to Attn of:

SR

Professor Claude Canizares
Chair, Space Studies Board
National Academy of Sciences
2101 Constitution Avenue, NW
Washington, DC 20418

Dear Professor Canizares:

The House-Senate Conference Report on Appropriations for NASA includes the following paragraph:

"The conferees are concerned that the large amount of data being collected as part of NASA science missions is not being put to the best possible use. To allay these concerns, the conferees direct NASA to contract with the National Research Council (NRC) for the study of the availability and usefulness of data collected from all of NASA's science missions. The study should also address what investments are needed in data analysis commensurate with the promotion of new missions."

We are surprised by these concerns, since the NASA Science Enterprises have, during the past few years, made deliberate efforts to make data from our science missions accessible and user-friendly. On the other hand, we recognize that there is certainly room for improvement.

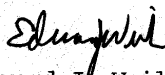
The purpose of this letter is to request that the Space Studies Board (SSB) conduct the study directed by Congress. We are addressing the SSB since it is the NRC entity most closely connected with the content and usage of science research in Earth and space science. We request that you address the following issues:

- (1) How available/accessible are data from science missions (after expiration of processing and proprietary analysis periods, if any) from the point of view of both scientists in the larger U.S. research community, as well as U.S. education/public outreach specialists? What, if anything, should be changed to improve accessibility?

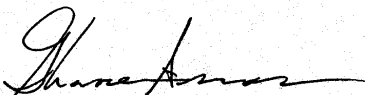
- (2) How useful are current data collections/archives from NASA's science missions as resources in support of high priority scientific studies in each Enterprise? How well are areas such as data preservation, documentation, validation, and quality control being addressed? Are there significant obstacles to appropriately broad scientific use of the data? Are there any changes in data handling and data dissemination that would improve usefulness?
- (3) Keeping in mind that NASA receives appropriated funds for both mission development as well as analysis of data from earlier or currently operating missions, is the balance between attention to mission planning and implementation versus data utilization appropriate in terms of achieving the objectives of the Enterprises? Should the fraction of a mission's life-cycle cost devoted to data processing, storage, and accessibility be changed?

We look forward to receiving your proposal, including study costs and estimated completion date. Please keep in mind that earlier drafts of the above-mentioned congressional mandate included a report due date of June 15, 2000, which we should still aim for. If you require further information on our current plans or if you would like to discuss study charter and implementation, please work with Dr. Jack Kaye, Director, Research Division in the Office of Earth Science and Dr. Guenter Riegler, Director, Research Division, Office of Space Science. For contractual matters, please contact Ms. Dolores Holland at (202)358-0834.

Sincerely,



Edward J. Weiler
Associate Administrator
for Space Science



Ghassem R. Asrar
Associate Administrator
for Earth Science

cc:
SR/Dr. G. Riegler
SR/Ms. D. Holland
YS/Dr. J. Kaye
NRC/Mr. J. Alexander

Appendix B

The Data Life Cycle

Data and information considerations play a critical role in all parts of the development of a space or earth science mission (see Figure B.1).

- *Mission planning* includes attention to data management as part of the overall project strategy. Science goals provide the basis for defining the requirements for data content, quality, and level of analysis, and these requirements must be factored in to the design of the project. During this planning phase, policies are established concerning the format of the data, data rights, and where and how the data will be processed, delivered, and archived.

Many missions are part of a larger program (e.g., Earth Observing System, Mars Exploration Program), and the availability of information from one mission in a series is often needed to support the design of subsequent missions. Indeed, the formulation of mission concepts is usually based on the results of previous missions. For example, the recently launched Microwave Anisotropy Probe mission was conceived in response to the results of the Cosmic Background Explorer mission. Thus, designers of new mission concepts must have access to well-formulated and complete information. In the case of some programs, such as the Mars Exploration Program, the time frame for conception of new missions is very short since launch opportunities exist at frequent intervals (every 26 months in the case of the Mars program). Other missions have a finite lifetime, and so early observations are important in guiding subsequent experiments. The Space Infrared Telescope Facility has planned a Legacy Program, which will make reduced observations from the earliest phases of operation available to the community so that it can optimize the use of the limited observing time available with this telescope. In such cases, adequate funding must be provided for timely research and analysis in order to generate the needed planning and management information.

- *Mission selection* involves many users of information, including managers at NASA Headquarters for program and project formulation and budgeting, and officials in the executive and legislative branches for program and project funding.

- *Mission and project design and building* depend on information such as spacecraft and instrument performance, whereas *mission testing* depends on information about the operational environment that has been obtained from other missions.

- *Mission operations* must be considered in all aspects of mission design, because the investment in operational infrastructure (e.g., Tracking and Data Relay Satellite System and Deep Space Network resource requirements and time allocation), plus the associated data and information infrastructure to support operations require a considerable fraction of the resources

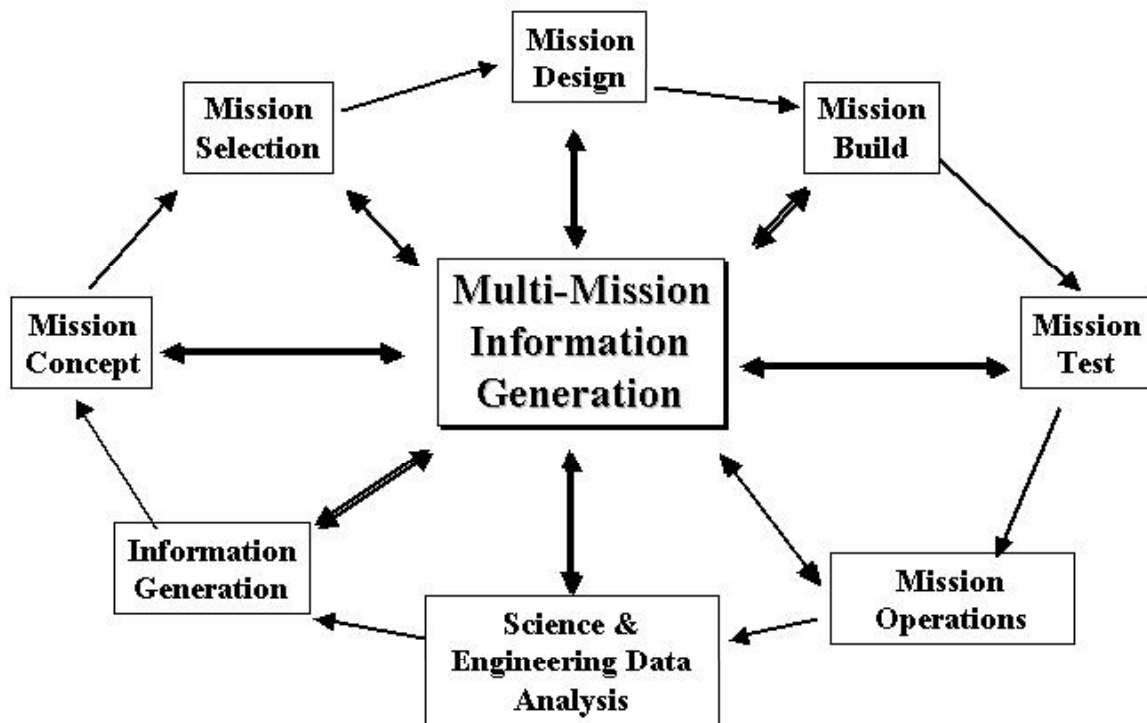


FIGURE B.1 Information supports every phase of mission development and operation, and mission development and operation in turn generate information that can be applied to future missions.

of any space mission. Trade-offs are needed, for instance, between preliminary processing of data onboard a spacecraft and processing on the ground.

- *Analysis of science data* from a mission must be supported by information on factors such as the state of the spacecraft at the time the data were gathered. Data from one instrument can shed light on the state of another instrument; for example, a weather sensor on a Mars lander can provide temperature information for the calibration of a camera.
- Finally, the *conversion of data into information* is the “value added” process that creates the products of the mission—knowledge about the Earth, the planets, and the cosmos.

Figure B.2 illustrates the elements of the information flow from a space mission or collection of missions that must be designed and supported. Data collected both from space and ground facilities must often be integrated. For example, remotely sensed data from earth observing satellites must be validated with “ground truth” data. Both space (e.g., the Hubble Space Telescope) and ground (e.g., the Keck telescope) astronomical observations can shed light on astrophysical processes.

Viking Mars data collected in the late 1970s are still being used to provide context and augment modern Mars observations from the Mars Global Surveyor. This example illustrates the necessity of secure, long-term data and information archiving with user-friendly access. Long-term maintenance of data, in turn, requires the ability to cope with rapidly changing technology. NASA and other agencies still have older data stored, precariously, on paper and computer tape, whereas new data are stored on CDs and on silicon media. Coping with these diverse forms and formats poses a challenge to users.

Data collection, whether in space or on Earth, requires ancillary data in addition to the data from the science instruments themselves. Such metadata includes information on where an instrument was pointed, the instrument’s temperature, its state of power, and so on. Transmission of the data and metadata, particularly from deep space missions, is often a bottleneck because of the increasing power and antenna aperture required as spacecraft fly farther from the Earth.

Data captured by a ground station must be calibrated and analyzed merely to account for the idiosyncrasies of the instrument in its environment. Sometimes this calibration can be very time consuming and expensive, particularly when an instrument is flying in an unknown environment. In the case of recent Mars missions, for example, the presence of atmospheric dust confused interpretation of thermal readings from the surface. Similarly, clouds on Earth present problems for satelliteborne instruments attempting to acquire surface measurements in the visible and infrared spectrums. Some of the complexities in developing data pipelines result from the need to use high-level data products from one instrument (e.g., high-resolution cloud masks) as input to processing of data from other instruments so that the observations can be interpreted properly.

In order to draw scientific conclusions about the state of an observed planet, astronomical object, or region on Earth, data from a number of instruments must often be synthesized. For example, imaging and laser altimetry data from the Mars Global Surveyor have been combined to draw conclusions about the possibility of liquid water on the surface of the planet.¹ The process of synthesizing data generates information, which is then interpreted to produce knowledge. For example, the determination of whether humans are contributing to global warming of the Earth requires considerable synthesis of data collected from a wide variety of

¹M.C. Malin and K.S. Edget, 2000, Evidence for recent groundwater seepage and surface runoff on Mars, *Science*, v. 288, pp. 2330-2335.

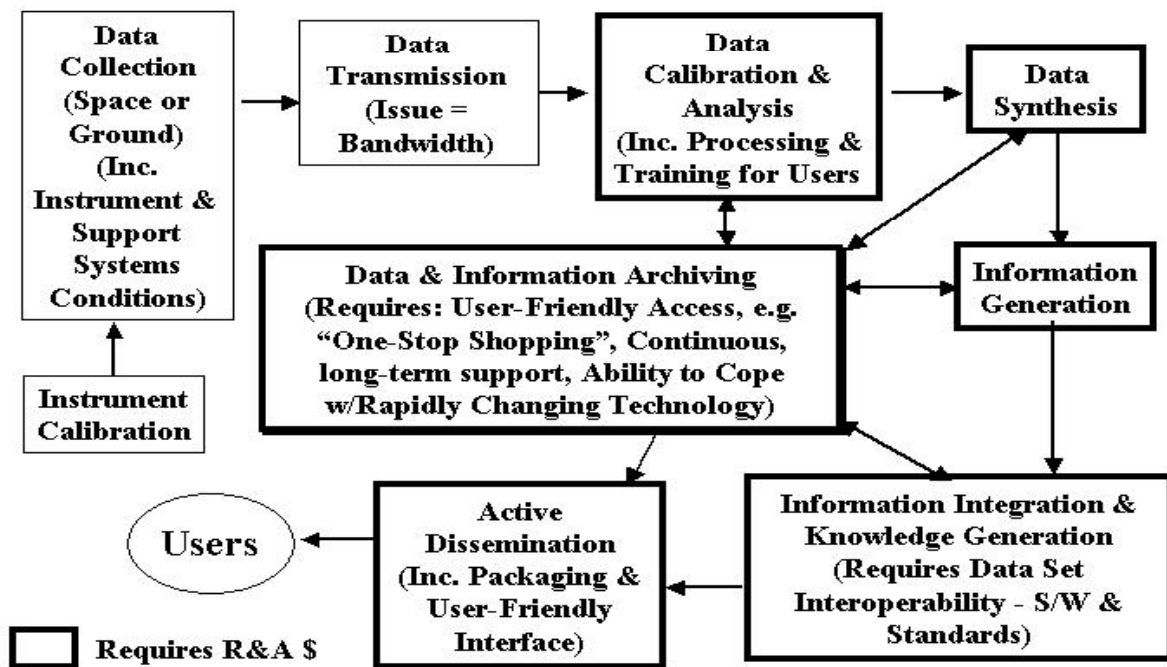


FIGURE B.2 The flow of data from initial acquisition to end users. This diagram illustrates the various steps in processing and disseminating data and some of the areas where significant investment is required in order to realize the full scientific potential of missions (e.g., for calibration, bandwidth, interoperability of various active archives and mission sites, user-friendly interfaces, and so on).

platforms (satellite, aircraft, ship, ground) over many years.² For the synthesis to occur data sets must be in a form in which they can be integrated—at a minimum they must use the same quantities! Standard software and community-accepted standards for data analysis are essential. Data that are archived must also be saved in standard forms so that they can be stored, retrieved, and used efficiently.

Finally, information must be widely disseminated to users to ensure a proper return on the investment of the time and cost of collection. Users include not only the scientists who generate knowledge from the data, but a very large nonscientist community, including engineers who design and implement future missions; managers who make decisions about mission design, selection, and funding; decision makers such as Congress, and the general public. The public includes students, educators, the news media, commercial enterprises, and interested people worldwide.

Table B.1 summarizes the categories of users who need information at different stages of the space mission process shown in Figure B.1. Engineers and scientists are actively involved in all stages, whereas managers are usually making decisions only in the “project” portions of a mission. Congress and the budgetary and policy elements of an administration are involved primarily in the selection process. The public is the ultimate consumer of space-derived knowledge, but it is also a participant in information generation (e.g., the news media and educators), in influencing mission conceptualization and selection, and even, in rare instances, in mission operations (e.g., student selection of targets for the Mars Orbiting Camera).

TABLE B.1 Information Users in the Stages of a Space Mission

| Stage | Engineers | Managers | Congress | Scientists | Public |
|---------------------------------------|-----------|----------|----------|------------|--------|
| Mission concept | x | x | | x | x |
| Mission selection | x | x | x | x | x |
| Mission design | x | x | | x | |
| Mission building | x | x | | x | |
| Mission testing | x | x | | x | |
| Mission operations | x | x | | x | |
| Science and engineering data analysis | x | | | x | |
| Information generation | x | | x | x | x |
| Multi-mission information generation | x | | x | x | x |

²See, for example, Intergovernmental Panel on Climate Change, 2001, *Climate Change 2001: The Scientific Basis*, Cambridge University Press, Cambridge, U.K.

Appendix C

Questionnaire to the Active Archives, Data Centers, and Data Services

Information Requested from the Space Science and Earth Science Facilities

- Number of unique users in FY95 and FY00.
- Actual FY00 budget and projections to FY05.
- Total number of staff (including active archive and Civil Servants) in FY00.
- Volume of holdings in FY00 and projected holdings to 2005.
- List of instruments providing data between now and FY05.
- List of data holdings as of January 2001.
- Current membership and the last 2 reports of your User Working Group.
- Metrics and/or statistics on users, holdings, publications that cite NASA data, etc.
- Results of user surveys (if any).
- Who is processing/will process the data from each instrument (active archive, science computing facilities, other)?
- Who are the current users and what user groups should be served in the future?
- How is the satisfaction of current users assessed?
- How are new user groups identified and entrained?
- What are you doing to make data: (1) accessible and (2) useable to nonscientists?
- How do you determine how useful your data has been to: (1) scientists, (2) private sector companies, (3) policy makers, (4) educators, and (5) the general public?
- How do you or will you decide how to retire any data sets?

Additional Information Requested of the Space Science Active Facilities

- Brief history of the center (e.g., 2-3 sentences on when it was formed, how it evolved)
- Host institution
- Disciplines served by the center
- Mission of the center

Additional Information Requested of the Distributed Active Archive Centers

- For what instruments are you using the ECS? Is the ECS working to specification?

Appendix D

Biographies of Task Group Members

Sidney C. Wolff, *Chair*, is immediate past director of the National Optical Astronomy Observatories. Her science interests include stellar atmospheres, stellar evolution, galactic/open clusters, star formation, and astronomical instruments and techniques. Dr. Wolff is the first woman to head a major observatory in the United States and has earned international recognition for her research on stellar atmospheres and the evolution, formation, and composition of stars, with emphasis on A-type stars. She served as president of Astronomical Society of the Pacific in 1985 and 1986 and was elected president of the American Astronomical Society in 1992. She received the Meritorious Public Service Award from the National Science Foundation for her outstanding support of the Gemini Project and her success in revitalizing it. Dr. Wolff is a former member of the National Research Council's Committee on Space Astronomy and Astrophysics, 1981-1984; Astronomy and Astrophysics Survey Committee, 1989-1992; Board on Physics and Astronomy, 1992-1995; and Task Group on Alternative Organizations of the SSB (Space Studies Board) Committee on the Future of Space Science, 1994-1995.

Thomas A. Herring, *Vice-Chair*, is professor of geophysics in the Department of Earth, Atmospheric and Planetary Sciences at the Massachusetts Institute of Technology. His areas of expertise and research interests involve the applications of high-precision geodetic measurement systems, primarily the Global Positioning System and Very Long Baseline Interferometry (VLBI). Dr. Herring's professional and advisory activities include the following: he is a fellow of the American Geophysical Union; member of the International Association of Geodesy's Special Study Group on Atmospheric Refraction, 1986-1987; chairman of the NASA Advisory Panel on the Applications of Water-Vapor Radiometry, 1988-1990; member of the NASA Advisory Panel on the Role of VLBI in the 1990s, 1988-1989; cochairman of the NASA Measurement Technique and Technology Panel, 1989; and member of the NASA Solid Earth Working Group. He is a member of NASA's Geoscience Laser Altimeter System science team. Dr. Herring's National Research Council service includes membership on the Committee on Geodesy, 1990-1994 (chair 1994-96); U.S. Geodynamics Committee, 1995-1996; Committee on Optimizing the Differential Global Positioning System Infrastructure for Scientific Applications, 1995-1996; and Committee on Gravity Measurements from Space, 1996-1997.

Joel Bregman is a professor in the Department of Astronomy at the University of Michigan. His research interests are in theoretical and observational studies of interstellar and intergalactic gas, and multiwavelength space astrophysics (X-ray, ultraviolet, infrared). He is an investigator on the Advanced Satellite for Cosmology and Astrophysics, Roentgen Satellite, Infrared Space Observatory, Hubble Space Telescope, Far Ultraviolet Spectroscopic Explorer, Chandra, and

X-ray Multi-Mirror Mission, and an observer at the radio facilities Very Large Array, Dominion Radio Astrophysical Observatory, Institut de RadioAstronomie Millimétrique, and the optical facility Michigan-Dartmouth-MIT Observatory. Dr. Bregman serves on the NASA Astrophysics Working Group, High Energy Archive Working Group, and Infrared Processing and Analysis Center Users Committee. He is a member of the American Astronomical Society and the International Astronomical Union.

Michael J. Folk is a technical program manager at the National Center for Supercomputing Applications (NCSA), University of Illinois. His professional interests are primarily in the area of scientific data management. He has led the HDF (Hierarchical Data Format) Project at NCSA since 1988. Through his work with HDF, Dr. Folk is involved with data management issues in NASA and the earth science community, particularly the Earth Observing System Data and Information System. He has also led the effort to provide a standard format to address data management needs of the U.S. Department of Energy's ASCI project, which involves data input/output, storage, and sharing among terascale computing platforms. Before coming to NCSA, Dr. Folk taught computer science at the university level for 18 years. Among Dr. Folk's publications is the book *File Structures: A Conceptual Toolkit*.

Richard G. Kron is a professor in the Department of Astronomy and Astrophysics at the University of Chicago. He is also a scientist at the Experimental Astrophysics Group, Fermi National Accelerator Laboratory (Fermilab). His research interests include studies of galaxies with the Hubble Space Telescope and ground-based telescopes. One of Dr. Kron's responsibilities within the Experimental Astrophysics Group is monitoring the efficiency of data acquisition for the Sloan Digital Sky Survey. The Fermilab group is responsible for processing the imaging and spectroscopic data of the Sloan Digital Sky Survey. He has served in the position of scientific spokesperson for the Sloan Digital Sky Survey since July 2001. Dr. Kron's prior National Research Council service includes membership on the Steering Committee for the Task Group on Space Astronomy and Astrophysics, Panel on Cosmology, and the Space Studies Board.

James F.W. Purdom is a senior research scientist at the Cooperative Institute for Research in the Atmosphere (CIRA) at Colorado State University. Before joining CIRA in 2001, he spent four years as director of the Office of Research and Applications in the National Oceanic and Atmospheric Administration's National Environmental Satellite, Data, and Information Service. Dr. Purdom's research focuses on remote sensing of the earth and its environment from space, as well as the development and evolution of atmospheric convection, with an emphasis on the study of mesoscale processes using satellite data. He received the U.S. Department of Commerce Silver Medal in 1994, the National Weather Association Special Award in 1996, and the American Meteorological Society Special Award in 1997. He currently chairs the World Meteorological Organization's Commission on Basic Systems Open Program Area Group on Global Observing Systems.

Donna L. Shirley is assistant dean of engineering for advanced program development at the University of Oklahoma, where she is responsible for coordinating engineering education activities. She is also president of Managing Creativity, a speaking and consulting firm, and is a well-known speaker, consultant, and trainer on the management of creative teams. She is the

author of *Managing Martians*, published in June 1998, and *Managing Creativity*, available at her Web site. Ms. Shirley has an M.S. in aerospace engineering, plus more than 38 years experience in engineering of aerospace and civil systems, including 30 years in management. She retired in August 1998 as manager of the Mars Exploration Program at the Jet Propulsion Laboratory. Prior to this position, Ms. Shirley managed the team that built Sojourner, the Microrover, which was landed by the highly successful Mars Pathfinder project on the surface of Mars on July 4, 1997. Ms. Shirley has numerous awards, including three honorary doctorates of humane letters.

Walter H.F. Smith has been a geophysicist at the National Oceanic and Atmospheric Administration since 1992. His scientific research areas include the interpretation of topography and gravity fields; the geophysics of the ocean basins; the use of satellite altimetry for mapping the ocean floors, monitoring climate, and forecasting hurricane intensification; and the use of satellite gravity data for studying mass flux in global climate change. Dr. Smith serves or has served on numerous committees, including the International Council for Science's Scientific Committee on Oceanic Research Working Group 107 on Improved Global Bathymetry and the General Bathymetric Charts of the Oceans. He is recognized for his work in developing a method for reconnaissance mapping of the ocean floors using satellite altimetry. Dr. Smith served on the National Research Council's Committee on Earth Gravity from Space, 1996-1997.

Nick Van Driel manages the Land Cover Characterization Program at the U.S. Geological Survey's (USGS's) EROS Data Center. He began his career with the USGS in Reston, Virginia, as a research geologist specializing in computer applications. Dr. Van Driel's subsequent assignments at USGS headquarters include that of information systems coordinator for the Geologic Division, deputy chief of the Office of Scientific Publications, and director of the Geographic Information Systems Laboratory in Reston. In 1994, he transferred to the EROS Data Center to manage its research program. He helped create the Land Cover Characterization Program in 1996, which has developed a Global Land Cover Database and the recently completed National Land Cover Dataset. His publications include articles on the application of satellite data to land cover mapping.

Donald J. Williams is retired as chief scientist in the Research and Technology Development Center at Johns Hopkins University's Applied Physics Laboratory (APL). He first joined APL's Space Department in 1961 and participated in the laboratory's early space activities. In 1965, he joined NASA's Goddard Space Flight Center, where he headed the Particle Physics Branch. In 1970, he was appointed director of the National Oceanic and Atmospheric Administration's (NOAA's) Space Environment Laboratory in Boulder, Colorado. In 1982, he rejoined APL's Space Department. During 1982-1989, he was supervisor of the Space Sciences Branch and from 1990 to October 1996 was director of the Milton S. Eisenhower Research Center. He has been the principal investigator of a variety of NASA, NOAA, Department of Defense, and European satellite programs. His research activities are in the field of space plasma physics, with an emphasis on planetary atmospheres. Dr. Williams has served on several national and international planning and advisory committees on space plasma physics. He is a past chair of the National Research Council's Committee on Solar-Terrestrial Research and a member of the Committee on Solar and Space Physics. Dr. Williams is past president of the International Association of Geomagnetism and Aeronomy. He is principal investigator for the energetic

particles detector on NASA's Galileo spacecraft, and the energetic particle detector and ion composition experiment on the Japanese/NASA Geotail satellite.

Roger V. Yelle is associate professor in the Department of Physics and Astronomy at Northern Arizona University. Dr. Yelle studies atmospheres and icy surfaces in this solar system and the atmospheres of extrasolar planets. He analyzes telescopic and spacecraft data and constructs theories and models to determine the composition and structure of atmospheres and their interaction with surfaces. His current projects include the following: the structure of extrasolar, Jupiterlike planets; thermal modeling of the energetics of the Jovian stratosphere and upper atmosphere in order to determine the importance of radiative and dynamical processes, and the relationship between composition and thermal structure; and analysis of the ultraviolet spectra of Jupiter in order to constrain the abundance of aerosols and hydrocarbons and to understand the role of Raman scattering. Dr. Yelle also works on NASA planetary missions, and he is a team member on the Ion Neutral Mass Spectrometer experiment on the Cassini mission to the Saturn system. He is also a member of the Miniature Integrated Camera Spectrometer team on NASA's Deep Space 1 mission to an asteroid and a comet. Dr. Yelle is a member of the American Geophysical Union, American Astronomical Society, and European Geophysical Society. He is a former member of the National Research Council's Committee on Planetary and Lunar Exploration.

James R. Zimbelman is a geologist for the Center for Earth and Planetary Studies, National Air and Space Museum (NASM), Smithsonian Institution. His main area of expertise is planetary geology, with emphasis on the analysis of high-resolution remote sensing and imaging data of Mars, geologic mapping of Mars and Venus, computer simulations of lava flows, and field studies of volcanic and aeolian features. He has been curator for "Exploring the Planets" (NASM Gallery 207) since March 1998. Prior to his position at the Smithsonian, Dr. Zimbelman was staff scientist at the Lunar and Planetary Institute. His many honors and activities include the following: chairman, Planetary Geology and Geophysics Review Panel, NASA, 1997-1999; NASA Venus Data Analysis Program Review Panel, 1992; and chairman, Regional Planetary Image Facility Directors and Data Managers Group, NASA, 1994-1997. Dr. Zimbelman is a fellow of the Geological Society of America, and member of the American Geophysical Union and the Geological Society of America.

Appendix E

Meeting Agendas

MEETING 1

Wednesday, January 31, 2001

Closed Session

7:30 a.m. Continental Breakfast

8:00 Welcome and Introductions
Review of Agenda

Sidney Wolff
Thomas Herring

8:15 Bias and Conflict Discussion

Joseph Alexander

10:15 Break

10:30 Discussion of Statement of Task
Discussion of Schedule
Discussion of Open Session Speakers

12:00 Lunch

Open Session

1:00 p.m. Space Science Perspective
Space Science Enterprise Science Questions
Budget Discussion
Space Science Data Policies/Philosophies

Guenter Riegler
Guenter Riegler
Joseph Bredekamp

4:00 Space Science Education and Outreach
Broker Facilitators
Forums

Cassandra Runyon
Terry Teays

5:00 Adjourn

5:00-6:00 Reception

Thursday, February 1, 2001

Open Session

7:30 a.m. Continental Breakfast

8:00 Earth Science Perspective
Earth Science Enterprise Science Questions
Budget Discussion
Earth Science Data Policies/Philosophies
Earth Science Education and Outreach

Jack Kaye

Jack Kaye
Blanche Meeson

12:00 Lunch

Closed Session

1:00 p.m. Preparation of Questions for Standing Committees
Other Information/Input Needed
Relevant NRC Reports
Discussion of Speakers, Location of Next Meeting

Discussion of Draft Outline
Initial Writing Assignments

4:00 Adjourn

MEETING 2

Monday, April 30, 2001

Closed Session

7:30 a.m. Continental Breakfast

8:00 Welcome and Overview of Meeting Goals
Introduction of New NRC Staff
Change in Status of Committee Chairs
Progress Since the Last Meeting
Proposal to Write a Short, High-Level Report
Review of Committee Charge
Bias/Conflict of Interest Discussion

Sidney Wolff

Joseph Alexander

8:20 What are the 3 key issues that should be addressed in the report?
(10 minutes each)
Earth Science Issues
• Thomas Herring
• Walter Smith
• Nick Van Driel

Space Science Issues

- Joel Bregman
- Richard Kron
- Donna Shirley
- Roger Yelle
- James Zimbelman
- Sidney Wolff

Computer Science Issues

- David DeWitt
- Michael Folk

10:30 Break

10:45 NASA's Science Priorities *Thomas Herring, Richard Kron*

11:00 Main Themes of the Report *Committee*

12:00 Lunch

Open Session

1:00 Space Science Enterprise
• Summary of information gathered so far *Edmund Reeves*
• Discussion and analysis *Committee*
• What additional information is needed?

3:00 Break

3:15 Earth Science Enterprise
• Summary of information gathered so far *Edmund Reeves*
• Discussion and analysis *Committee*
• What additional information is needed?

5:00 Adjourn

6:30 Committee Dinner

Tuesday, May 1, 2001

Closed Session

7:30 a.m. Continental Breakfast

8:00 Overview of Plans for the Day *Sidney Wolff*

8:15 Review of Relevant Advisory Committee Reports
• Recommendations from previous NRC reports *Anne Linn*
• Help from NRC standing committees on this study
• Recommendations from NASA advisory committees *Edmund Reeves*

Open Session

- 9:30 Conference Calls with Advisory Committee Chairs
- [Space Science] Mission Operations and Data Analysis Task Force
 - Earth Science Data and Information Systems and Services
- David Black*
Sara Graves

10:15 Break

Closed Session

10:30 Main Themes of the Report *Committee*

Open Session

- 11:30 Conference Calls with Advisory Committee Chairs (continued)
- [Space Science] Task Group on Science Data Management
- Jeffrey Linsky*

Closed Session

12:00 Lunch

- 1:00 p.m. Plan of Action *Committee*
- Outline for the report
 - Writing assignments
 - Additional information gathering
 - Timeline for completing the report
 - Schedule of meetings
 - Focus of next meeting

3:00 Break

3:15 Outstanding Issues *Committee*

4:00 Adjourn

MEETING 3

Monday July 30, 2001

Closed Session

7:30 a.m. Continental Breakfast

8:00 Introductions and Goals of the Meeting *Sidney Wolff*
 New Committee and Staff

8:30 Overview of Meeting *Sidney Wolff*
 Questions for Speakers
 Progress Since the Last Meeting *Monica Lipscomb*

- Supplemental information on data centers
- New information from NRC and NASA advisory committees

| | | |
|-------|---|--|
| 8:45 | Discussion of Report Summary of Recommendations from the Last Meeting New Outline for the Report Reports from Lead Authors <ul style="list-style-type: none"> • Nick Van Driel • David DeWitt • James Zimbelman | <i>Sidney Wolff</i> <i>Sidney Wolff</i> |
| 10:30 | Break | |
| 10:45 | Reports from Lead Authors (continued) <ul style="list-style-type: none"> • Michael Folk • Donna Shirley • Joel Bregman • Thomas Herring • Walter Smith • Sidney Wolff | |

Open Session

| | | |
|-----------|------------------------------|-----------------------|
| 11:30 | National Virtual Observatory | <i>Ethan Schreier</i> |
| 12:00 | Lunch | |
| 1:00 p.m. | EOSDIS | <i>Dolly Perkins</i> |
| 1:30 | Federation | <i>John Townshend</i> |
| 2:00 | NewDISS | <i>Steven Wharton</i> |
| 2:30 | Discussion | <i>All</i> |
| 3:00 | Break | |
| 3:15 | NASA CIO | <i>Lee Holcomb</i> |
| 3:45 | Discussion | <i>All</i> |
| 5:00 | Adjourn | |
| 6:30 | Committee Dinner | |

Tuesday, July 31, 2001

Closed Session

| | |
|-----------|-----------------------|
| 7:30 a.m. | Continental Breakfast |
|-----------|-----------------------|

- 8:00 Review of Information Gathered from the Previous Afternoon
- 9:00 Report
- Review of recommendations
 - Themes and organization of the report
- 10:30 Break
- 10:45 Review of Recommendations and Themes (continued)
- 12:00 Lunch
- 1:00 p.m. Plans for Completing the Report
- Writing assignments
 - Date of the next meeting
 - Additional information to be gathered
- 1:30 Review of Recommendations and Themes (continued)
- 3:00 Adjourn

Appendix F

Acronyms

| | |
|---------|---|
| ADC | Astronomical Data Center |
| ASF | Alaska SAR Facility |
| CFC | chlorofluorocarbon |
| CIO | chief information officer |
| COTS | commercial off-the-shelf |
| CSIO | chief science information officer |
| DAAC | Distributed Active Archive Center |
| ECS | EOSDIS Core System |
| EDC | EROS Data Center |
| EOS | Earth Observing System |
| EOSDIS | Earth Observing System Data and Information System |
| ESE | Earth Science Enterprise |
| ESIP | Earth Science Information Partner |
| FFARS | Food and Fiber Applications of Remote Sensing |
| FGDC | Federal Geographic Data Committee |
| GALEX | Galaxy Evolution Explorer |
| GLOBE | Global Learning and Observations to Benefit the Environment |
| GSFC | Goddard Space Flight Center |
| HEASARC | High Energy Astrophysics Science Archive Research Center |
| HST | Hubble Space Telescope |
| IRSA | Infrared Science Archive |
| LaRC | Langley Research Center |
| MAST | Multi-mission Archive at Space Telescope |
| MGS | Mars Global Surveyor |
| NARA | National Archives and Records Administration |
| NASA | National Aeronautics and Space Administration |
| NED | NASA/Infrared Processing and Analysis Center Extragalactic Database |
| NOAA | National Oceanic and Atmospheric Administration |
| NRC | National Research Council |
| NSF | National Science Foundation |
| NSIDC | National Snow and Ice Data Center |
| NSSDC | National Space Science Data Center |
| NVO | National Virtual Observatory |
| OAIS | Open Archival Information System |
| ORNL | Oak Ridge National Laboratory |

| | |
|---------|---|
| OSS | Office of Space Science |
| PDS | Planetary Data System |
| PI | principal investigator |
| PO.DAAC | physical oceanography DAAC |
| R&DA | research and data analysis |
| RESAC | Regional Earth Science Applications Centers |
| RPIF | Regional Planetary Image Facility |
| SCF | science computing facility |
| SDAC | Solar Data Analysis Center |
| SEDAC | Socioeconomic Data and Applications Center |
| SEEDS | Strategic Evolution of ESE Data Systems |
| SIRTF | Space Infrared Telescope Facility |
| TOMS | Total Ozone Mapping Spectrometer |
| TRMM | Tropical Rainfall Measurement Mission |
| USGS | U.S. Geological Survey |