PB84-195361

Solar-Terrestrial Data Access Distribution, and Archiving

National Research Council, Washington, DC

Prepared for

National Oceanic and Atmospheric Administration
Washington, DC

Feb 84

50272-101

REPORT DOCUMENTATION PAGE	1. REPORT NO.	2.	3. Recipient's Accession No. PB84 195361
4. Title and Subtitle Solar-Terrestrial Data Access, Distribution, and Archiving			5. Report Date February 1984
			6.
7. Author(s) Joint Data Panel, Com. on Solar-Terr. Res.; Com. on Solar & Space Phys.			8. Performing Organization Rept. No.
9. Performing Organization Name and Address National Research Council Commission on Physical Sciences, Mathematics & Resources 2101 Constitution Avenue, NW Washington, DC 20418			10. Project/Task/Work Unit No.
			11. Contract(C) or Grant(G) No. (C) (G)
12. Sponsoring Organization Name and Address NASA, NSF, NOAA, Air Force			13. Type of Report & Period Covered Final
			14.

15. Supplementary Notes

16. Abstract (Limit: 200 words)

The report recommends that a central data catalog and data access network (CDC/DAN) for solar-terrestrial research be established, initially as a NASA pilot program. The system is envisioned to be flexible and to evolve as funds permit, starting from a catalog to an access network for high-resolution data. The report describes the various functional requirements for the CDC/DAN, but does not specify the hardware and software architectures as these are constantly evolving. The report also stresses the importance of a steering committee, working with the CDC/DAN organization, to provide scientific guidelines for the data catalog and for data storage, access, and distribution.

17. Document Analysis a. Descriptors

Data, network, catalog, distribution, archiving, solar-terrestrial, solar physics space physics, computer networks

b. Identifiers/Open-Ended Terms

c. COSATI Field/Group

18. Availability Statement	19. Security Class (This Report)	21. No. of Pages
Release unlimited	UNCLASSIFIED	43
	20. Security Class (This Page) UNCLASSIFIED	22. Price ——

See ANSI-Z39.18) See Instructions on Reverse OPTIONAL FORM 272 (4-77) (Formerly NTIS-35) Department of Commerce

NRC NAS NAE IOM

PB84-195361

Solar-Terrestrial Data Access, Distribution, and Archiving

Solar-Terrestrial Data Access, Distribution, and Archiving

Joint Data Panel
of the
Committee on Solar and Space Physics
Space Science Board
and
Committee on Solar-Terrestrial Research
Board on Atmospheric Science and Climate

Commission on Physical Sciences, Mathematics, and Resources

National Research Council

NATIONAL ACADEMY PRESS
Washington, D.C. 1984

i.a

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

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

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

The Committee on Solar and Space Physics is pleased to acknowledge the support of the National Aeronautics and Space Administration. The Committee on Solar-Terrestrial Research is pleased to acknowledge the support of the National Science Foundation, the Air Force, and the National Oceanic and Atmospheric Administration.

Committee on Solar and Space Physics
or
Committee on Solar-Terrestrial Research
National Research Council
Washington, DC 20418

Joint Data Panel

MARGARET ANN SHEA, Air Force Geophysics Laboratory, *Co-chairman*
DONALD J. WILLIAMS, Applied Physics Laboratory, Johns Hopkins University, *Co-chairman*
KENNETH L. ACKERSON, University of Iowa
JOE H. ALLEN, National Geophysical Data Center
ROBERT GOLD, Applied Physics Laboratory, Johns Hopkins University
ROY L. JENNE, National Center for Atmospheric Research
MICHAEL L. KAISER, NASA/Goddard Space Flight Center
ROBERT L. McPHERRON, University of California, Los Angeles

Liaison

MICHAEL J. WISKERCHEN, National Aeronautics and Space Administration
ANTHONY J. VILLASENOR, National Aeronautics and Space Administration

NRC Staff

RICHARD C. HART, Committee on Solar and Space Physics
THOMAS M. USSELMAN, Committee on Solar-Terrestrial Research

Committee on Solar and Space Physics

STAMATIOS M. KRIMIGIS, Applied Physics Laboratory, Johns Hopkins University, *Chairman*
LOUIS J. LANZEROTTI, Bell Laboratories, *Chairman* (until September 30, 1983)
R. GRANT ATHAY, High Altitude Observatory
LENNARD A. FISK, University of New Hampshire
ROBERT W. FREDERICKS, TRW
JOHN W. HARVEY, Kitt Peak National Observatory
JACK R. JOKIPII, University of Arizona
MARGARET KIVELSON, University of California, Los Angeles
MICHAEL MENDILLO, Boston University
ANDREW F. NAGY, University of Michigan
MARCIA NEUGEBAUER, Jet Propulsion Laboratory
KONSTANTINOS PAPADOPOULOS, University of Maryland
SUSAN SOLOMON, Aeronomy Laboratory/NOAA
DARRELL F. STROBEL, Naval Research Laboratory
GEORGE L. WITHBROE, Center for Astrophysics

Liaison Representatives

MELVYN L. GOLDSTEIN, Goddard Space Flight Center
JAMES RUSSELL, III, Langley Research Center

Ex-Officio

THOMAS M. DONAHUE, University of Michigan
DEVRIE S. INTRILIGATOR, Carmel Research Center
ROBERT M. McQUEEN, High Altitude Observatory

NRC Staff

RICHARD C. HART

Space Science Board

THOMAS M. DONAHUE, University of Michigan, *Chairman*
DON L. ANDERSON, California Institute of Technology
RAYMOND E. ARVIDSON, Washington University
JACQUES M. BECKERS, University of Arizona
DANIEL B. BOTKIN, University of California, Santa Barbara
ANDREA K. DUPREE, Center for Astrophysics
FREEMAN J. DYSON, The Institute for Advanced Study
RICCARDO GIACCONI, Space Telescope Science Institute
JAY M. GOLDBERG, University of Chicago
DONALD M. HUNTEN, University of Arizona
STAMATIOS M. KRIMIGIS, Applied Physics Laboratory, Johns Hopkins University
ROBERT M. MacQUEEN, National Center for Atmospheric Research
HAROLD MASURSKY, Center for Astrogeology
CARL E. McILWAIN, University of California, Sa Diego
BERNARD M. OLIVER, Hewlett-Packard Company
RONALD G. PRINN, Massachusetts Institute of Technology
J. WILLIAM SCHOPF, University of California, Los Angeles
EDWARD C. STONE, JR., California Institute of Technology
ANTHONY L. TURKEVICH, University of Chicago
RAINER WEISS, Massachusetts Institute of Technology
GEORGE WETHERILL, Carnegie Institution of Washington

DEAN P. KASTEL, *Staff Director*

Committee on Solar-Terrestrial Research

DEVRIE S. INTRILIGATOR, Carmel Research Center, *Chairman*
JOHN V. EVANS, Communications Satellite Corporation *Chairman* (until June 30, 1983)
HANNES ALFVEN, University of California, San Diego
ODILE DE LA BEAUJARDIERE, Stanford Research Institute
CONWAY B. LEOVY, University of Washington
JANET G. LUHMANN, University of California, Los Angeles
MARTIN POMERANTZ, Bartol Research Foundation
THOMAS A. POTEMRA, Applied Physics Laboratory/Johns Hopkins University
ARTHUR D. RICHMOND, National Center for Atmospheric Research
NEIL SHEELEY, Naval Research Laboratory
RICHARD S. STOLARSKI, NASA/Goddard Space Flight Center
PETER STURROCK, Stanford University
THOMAS E. VanZANDT, Aeronomy Laboratory/NOAA

Ex-Officio

STAMATIOS M. KRIMIGIS, Applied Physics Laboratory/Johns Hopkins University
JERRY D. MAHLMAN, Geophysical Fluid Dynamics Laboratory/NOAA
PATRICIA H. REIFF, Rice University
MARGARET ANN SHEA, Air Force Geophysics Laboratory

NRC Staff

THOMAS M. USSELMAN

Board on Atmospheric Sciences and Climate

THOMAS F. MALONE, West Hartford, Connecticut, *Chairman*
FERDINAND BAER, University of Maryland
LOUIS J. BATTAN, University of Arizona
WERNER A. BAUM, Florida State University
ROBERT A. DUCE, University of Rhode Island
JOHN A. EDDY, National Center for Atmospheric Research
PETER V. HOBBS, University of Washington
FRANCIS S. JOHNSON, University of Texas-Dallas
ROBERT W. KATES, Clark University
MICHAEL B. McELROY, Harvard University
JAMES C. McWILLIAMS, National Center for Atmospheric Research
VOLKER A. MOHNEN, State University of New York-Albany
ANDREW F. NAGY, University of Michigan
WILLIAM A. NIERENBERG, Scripps Institution of Oceanography
ROGER R. REVELLE, University of California-San Diego
JUAN G. ROEDERER, University of Alaska
NORMAN J. ROSENBERG, University of Nebraska
STEPHEN H. SCHNEIDER, National Center for Atmospheric Research
JOSEPH SMAGORINSKY, Princeton University
JOHN W. TOWNSEND, JR., Fairchild Space & Electronics Company
THOMAS H. VONDER HAAR, Colorado State University

Ex-Officio

DEVRIE S. INTRILIGATOR, Carmel Research Center
VERNER E. SUOMI, University of Wisconsin-Madison

JOHN S. PERRY, *Staff Director*

Commission on Physical Sciences, Mathematics, and Resources

PREFACE

In January 1982 the Committee on Solar and Space Physics (CSSP) and the Committee on Solar-Terrestrial Research (CSTR) jointly established a Data Panel and asked it to develop specific recommendations for improving data management in the field of solar-terrestrial physics. It seemed timely to try to sort out the many ideas and studies on this subject that were then available. Chief among these studies was the report issued by the Committee on Data Management and Computation (CODMAC) *Data Management and Computation, Volume 1: Issues and Recommendations* (NRC, 1982). The CODMAC report presented a detailed description of the problems related to space science data acquisition, analysis, and distribution. From a consideration of these problems, CODMAC developed a series of general and policy recommendations aimed at alleviating the basic causes of many of the existing problems.

From this general perspective the Data Panel has considered specific approaches toward improved data accessibility and storage. Due to the successful operation of existing large mass-storage systems and the impending availability of video disks and digital optical disks, data storage is no longer the problem it has been in the past. Consequently, the Data Panel focused on the problem of data accessibility as the present highest priority problem in solar-terrestrial data management. In developing a solution to this problem, both today's technological capabilities and the economic realities of funding agencies have been considered. Consequently, a phased approach is presented in which the first step requires only modest funding to provide substantial improvements; significant additional improvements would be possible with increased funding. The evolution of practical and usable data access will thus occur as resources become available.

The essential features of the approach that we recommend (see Chapter 1) are as follows: An on-line data catalog that can be accessed at a central point, a computer network to maintain the catalog and provide access to the data, and a means of involving active research scientists in the definition and maintenance of the system. These key elements are described in Chapter 2. In Chapter 3, we present considerations on data archiving procedures. Finally, in the Appendix, we briefly discuss some thoughts and concerns about the computer hardware and software compatibility problems that will be encountered as this approach is implemented.

In addressing the problem of data accessibility, it was tempting to become immersed in the latest data handling technology and software architectures in order to choose a specific solution. However, any specific solution agreed upon would, in all probability, be obsolete by the time this report was published. Consequently, we decided that a more appropriate approach to data accessibility problem was to describe the *functional requirements* that must be satisfied by any specific hardware and software solution. This report strictly follows this approach and recommends actions to be taken that, in time, will solve the data accessibility problem.

Consistent with the functional requirements approach, we point out general management attributes that must be incorporated into any proposed solution. Details of the management structure will be defined in a recommended pilot program.

The Data Panel has benefited greatly from discussions with CODMAC and presentations of Skylab data handling (G. Withbroe, Center for Astrophysics), the Climate Data Pilot Program (P. Smith, NASA/Goddard), SMM data handling (W. Wagner, High Altitude Observatory), the UARS data system (C.A. Reber and T. Taylor, NASA/Goddard), the OPEN data system (M. Kaiser, NASA/Goddard), and the Data System Users Working Group (J. Green, NASA/Marshall).

Contents

1

SUMMARY AND RECOMMENDATIONS

THE PROBLEM: DATA ACCESSIBILITY

From the beginning of the space age in 1957, the development of general and effective data management practices in solar-terrestrial research has remained a neglected area of concern. Modest funding has been made available for data archiving purposes, but fundamental questions of data accessibility and general data availability have been either neglected or treated on a piecemeal basis by individual experimenters. While some funding now is becoming available for these data management concerns, it is apparent that any permanent improvement in solar-terrestrial data management probably will have to use existing and expected programmatic resources from various agencies and projects. These improvements must be applicable not only to future data sets, but also to existing and past data sets that continue to be of great value not only for analysis for present problems but also for studies of long-term variations, such as solar cycle effects or climatic variations.

Two major problems of the data management issue have been data accessibility and data storage. These problems include questions of data location, description, formats, availability, and short-term (on-line) storage, intermediate (off-line) storage, and long-term (archival) storage. Recent technological advances in data storage have made great strides toward the solution of the data storage problem. Thus the problem of data accessibility has become the highest priority problem to be addressed in order to enhance the scientific use of solar-terrestrial data.

Data accessibility includes actual access to the data itself as well as the ability to readily obtain information about data location, availability, and status information (e.g., time period collected, level of processing, time resolution, quality, formats, and cost). Actual data access (electronic, tape, disk, microfilm, or paper copy) will vary according to individual circumstance.

To solve the problem of data accessibility requires that we develop ways to use existing resources to significantly improve solar-terrestrial data accessibility and thereby extract more science per dollar invested in the data collection process. Whatever the solution, it should easily allow the incorporation of additional future resources and/or projects into the data management system.

AN ERA OF CHANGE

The past 25 years have been a period of significant change in data handling concepts throughout solar-terrestrial research. Since the International Geophysical Year (1957-1958), the data collections of the World Data Center system have evolved from being primarily ground-based in origin to including large satellite data sets and from being mostly analog and tabular to emphasizing computer accessible digital data. Along with this evolution, major changes have taken place in the solar-terrestrial community's approach to solving scientific problems.

Since the discovery of the radiation belts, much successful space research has been accomplished by independently analyzing data from individual satellite-borne instruments. Great advances in our knowledge of the space environment were made using this approach, particularly during the discovery and exploration phases. Based on these advances, it became clear that the solutions to current scientific problems in solar-terrestrial research require data from a variety of instruments--a multiparameter data base--and individual researchers began to share data to pursue these problems jointly. The concept of obtaining a multiparameter data set suitable for particular space physics problems was fundamental in shaping the instrument complements and/or data handling systems in the Explorer 45 (S^3-A), AE, SMM, SME, and DE programs.

Today, few problems in solar and space plasma research can be solved by analyzing data from a single instrument (see *Solar System Space Physics in the 1980's: A Research Strategy*, NRC, 1980; and *Solar-Terrestrial Research for the 1980's*, NRC, 1981). The complexity of the problem often determines the extent of the required data base, which may span data ranging from severa. instruments on one satellite to data from several satellites plus a variety of ground-based observations.

This has brought about a significant change in the conduct of space research. In the past individual investigators (and their teams) exclusively analyzed data from a single instrument, whereas now it has become common to have groups of investigators jointly analyzing their combined data. The need for cooperative data analysis has led to the development of several pilot data handling programs within NASA. The Space Plasma Computer Analysis Network (SCAN), the Planetary Pilot Data System (PPDS), and the Climate Data Pilot Program are addressing a significant subset of the problems of coordinated scientific analysis. The concept of multiparameter data sets has been further encouraged in several satellite projects through the establishment of guest investigator and/or interdisciplinary scientist programs (e.g., ISEE, AE, Pioneer-Venus, and Galileo). In these programs, research proposals are solicited from scientists outside the project for studies specifically requiring data from more than one instrument.

Present solar-terrestrial research often requires multiparameter data sets. A number of data management concepts have been designed to meet that need: central and distributed data bases, computer networking, and data base management techniques are examples of such concepts. Because of the variable complexity of solar-terrestrial research problems, no single data management concept will dominate, and features of many approaches will be required in varying proportions for specific situations. Flexibility must be an essential characteristic of any data handling solution. The key to the success of fulfilling the needs of present data analysis, however, is how well the problem of data accessibility is solved.

RECOMMENDATIONS

The recommendations below address data accessibility as the key functional objective of solar-terrestrial data management. They also functionally describe a system that maintains flexibility and can easily evolve as resources become available. It is expected that, in general, resources from a number of sources can be used to establish and operate a solar-terrestrial data access network.

1. *We recommend that a pilot program be started by NASA that would lead to the establishment of a solar-terrestrial Central Data Catalog and Data Access Network (CDC/DAN).* The Central Data Catalog (CDC) should be established as a relational data base, should be supported by a query language, and should be accessible from remote terminals. This would allow data to be identified by relationships with other data elements. As a minimum the CDC should contain information as to data location, type, level of processing, time periods covered, quality, formats, cost, and availability. The CDC and the sources and users of solar-terrestrial data should be connected via computer networking to create the Data Access Network (DAN). The CDC will be the primary node for information about the data. The incorporation of user and data base nodes into the CDC/DAN permits, as resources allow, the growth of the CDC/DAN from a query catalog to an electronic mail and request service, to a browse capability of survey data sets including graphics, to the availability of on-line data sets throughout the network, to, finally, a browse capability of remote high-resolution data sets. In addition to the catalog and network, the CDC/DAN must also have a staff to manage the creation and maintenance of the catalog. Thus the CDC/DAN concept defines a data management organization. A possible location for the CDC node is the National Space Science Data Center (NSSDC), and the pilot program could begin by using subsets of existing NSSDC and National Geophysical Data Center (NGDC) data sets.

2. The CDC/DAN is expected to be a flexible and evolving structure. The data catalog structure, data base architectures, communication techniques, and methods of data access will evolve in time as more data become available to the network. In addition, different data require different data handling concepts, and so a variety of techniques will have to operate concurrently within the system. *We recommend the establishment of a Scientific Steering Committee (SSC) to work closely with the CDC/DAN organization to provide scientific guidelines for data storage, access, and distribution.*

3. *We recommend that agencies (e.g., NASA, NSF, DOD, NOAA, and DOE) sponsoring solar-terrestrial research projects configure their data handling systems to be fully compatible with the CDC/DAN.* Data catalogs should be made available as soon as possible, and operation within the

DAN as an on-line node with access to processed data should occur in a reasonable time interval after initial data reception (e.g., 1 year). Examples of such programs are AMPTE, UARS, OPEN, HILAT, CRRES, and RGON.

4. Many of the problems of software compatibility and transportability in solar-terrestrial data analysis have been repeatedly and independently solved at different institutions. This has resulted in duplication of effort and in software that is needlessly unique to individual sites. *We recommend that the CDC/DAN serve as a software clearinghouse and that the SSC suggest standardized practices for user software and documentation that would enhance their transportability.*

5. Experimenters are not alone in developing individual data bases; researchers using data from several instruments often construct unique data sets that could be very useful to the solar-terrestrial research community. *We recommend that funding agencies provide incentives to encourage researchers to provide useful and appropriate data sets to the CDC/DAN in a timely manner.* Such a policy would widen the use of these data sets and could ensure that highly processed data sets be archived if appropriate.

6. *We recommend that agencies having or sponsoring operational or research programs that acquire solar-terrestrial data develop plans at project initiation to provide resources for appropriate cataloging and processing into archival format.*

7. The present archives of solar-terrestrial data contain a variety of data sets and storage media; consequently, problems exist in merging these older data with the new. In addition, it is recognized that some data are of marginal value and should be either purged or archived in a special manner. *We recommend the establishment, as needed, of data archival advisory groups composed of discipline specialists to review archival activities and to provide guidance concerning priorities for continuation of existing programs and searching for unique but nonarchived data collections.* A liaison between the SSC and discipline-oriented data archival groups should be maintained to guide the application of new technology, to provide access to archives of past data, and to provide coordination with the CDC/DAN.

2

CENTRAL DATA CATALOG AND DATA ACCESS NETWORK

The main functional requirement of a solar-terrestrial data management system is to provide ready access to the variety of data required for problem solving. The effective use of data requires more than access to the actual data itself; it also requires access to knowledge of data location, type, description, formats, availability, level of processing, time period, time resolution, quality, ancillary data available, instrument descriptions, and cost. A central data catalog containing data descriptors and a data access network are key elements of a solar-terrestrial data management scheme. The data catalog should be constructed as a relational data base, should be supported by a query language, and should be the main information node on the data access network. The Central Data Catalog (CDC) should be a simple, user friendly, information catalog that allows users to determine where and how to obtain specific data. Together with the Data Access Network (DAN), it should easily evolve to a library mode offering browse capabilities and, ultimately, to direct data transfer.

An additional necessary ingredient of the CDC/DAN system is the involvement of the scientific community. Only if active research scientists are involved in the creation and maintenance of this system will it become the valuable tool we envision.

All three of these elements--CDC, DAN, and scientific involvement--are necessary for the system to be successful.

DESCRIPTION

Crucial to the success of any central solar-terrestrial data handling scheme is the ease with which data can be found. This requires that a general electronic catalog or directory of available solar-terrestrial data be constructed. This catalog should be in the form of a relational data base so that inquiries can be made that do not assume a priori that the user knows of the existence and properties of a specific set of data. For example, if a user is interested in ultimately obtaining data on the density and speed of the solar wind during a particular interval, he should not have to know what spacecraft were operating at that time. The catalog should inform the user as to the availability of data for the interval in question. In addition, the user should be able to obtain from the system a useful description of the data (e.g., source, data type, format, level of processing, ancillary data available, cost, and mode of access). Preparation of the data descriptions obviously will involve considerable care and effort. In addition to the staff of the CDC/DAN organization who would be responsible for managing and maintaining the catalog, we propose a scientific steering committee (see next section) that would be responsible for the definition of the level of detail and quality of the data descriptions in the data catalog.

The data catalog is but the first of an envisioned five levels of interaction between the user and the desired data. The catalog serves as a guide to the data, giving its description and location, and thus is the first step in attaining the goal of a solar-terrestrial data service. The following levels, 2 through 5, can be incorporated as resources allow.

Level 2 would be the implementation of an electronic mail request service for data. Once a particular data set is identified from the catalog, the user should, at the very least, be able to issue a request for that data from its source, e.g., a principal investigator, a project data base, an archive. Fulfillment of that request will depend on many factors such as cost and availability. Many data sets are and always will be available only in hardcopy form; thus the postal service will be the medium of transfer. The next three levels in the user-data interaction all presume that the data can be made available electronically.

Level 3 provides an ability for the user to perform a remote low-resolution browse of a desired data set. This will involve only survey or overview data sets and will require some rudimentary graphics capability so the user can display the data in a useful form. This level of interaction should be possible with present-day technology.

Level 4 involves actual on-line transfer of data sets. Level 5 extends the browse capability to the full desired data set itself. Both of these final two levels will place great demands on the communications hardware between the user and the repository of data, and may be realistic for only a fraction of the solar-terrestrial data available.

Obviously, this idealized picture of interactions between a scientist and the data will be difficult to implement, with funding being the foremost stumbling block. We believe, however, that a step-by-step approach as suggested here is realistic. Many of the data sets currently residing in the NSSDC and the NGDC could be entered in the CDC. Description of new data sets could be added to a core of existing catalog entries, and the usefulness of the catalog would grow with time. The next levels, 2 through 5, could follow as time, funding, and, most importantly, interest allow.

Figure 2.1 shows a block diagram of the CDC/DAN with the CDC as the information node. A variety of additional nodes are shown that may be either simply users of the CDC or participating data systems, supplying catalogs and/or data. Several data sets and users are shown to indicate that individual researchers can easily participate in the CDC/DAN. While much of the information transfer in the CDC/DAN eventually will occur electronically, data transfer also must include the mails. Many data sets are only available in hardcopy form, and it may be best for some to remain that way. Other data sets (e.g., imaging) may be best accessed by most users in hardcopy form. Finally, there will be users whose only mode of data access will be hardcopy. This illustrates the need for recognizing a balance between electronic and hardcopy modes of data access--a balance determined by the types of data and user's facilities being accessed.

Flexibility is maintained in that users can employ all attributes of the network and/or contact data sources directly. Because data sets vary greatly in size and complexity, data storage problems are best addressed on an individual basis. Some form of storage should be considered whereby data, depending on usage, move from fast on-line storage to moderate speed mass-storage, and finally, to off-line archival storage.

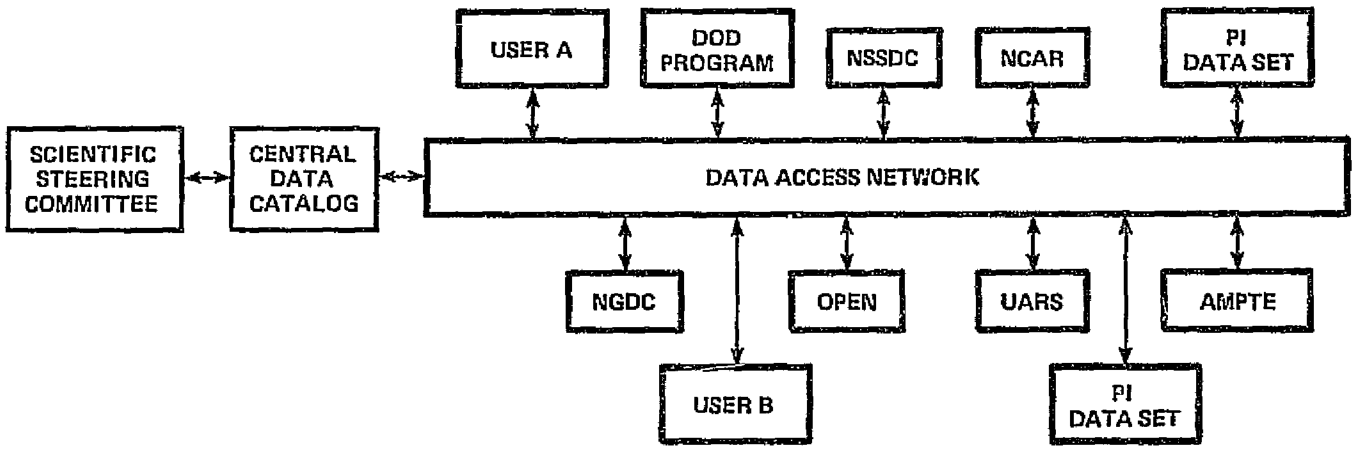

FIGURE 2.1 The Central Data Catalog and Data Access Network (CDC/DAN). The CDC is the main information node in this multiuser network and contains a query-language-supported, relational data base catalog of data descriptors, allowing access to data throughout the network. The DAN will accommodate a variety of participants and, as resources allow, can provide various modes of data transfer including fully electronic transfer. The scientific involvement is provided by the Scientific Steering Committee (SSC) that oversees the activities of the CDC.

SCIENTIFIC INVOLVEMENT

The CDC/DAN must maintain flexibility in order to incorporate a variety of data sets, each with individual characteristics, and to evolve as new data handling capabilities become available. Choices will have to be made on whether or not to apply contemporary data handling techniques to a particular data set. Decisions will be required concerning the scientific value of specific data sets relative to the cost of their incorporation into the CDC/DAN. Further decisions will be required concerning, for example, the CDC structure, data descriptions, ancillary data required, communication techniques, and modes of data access.

As recognized in the CODMAC report and in *Geophysical Data Interchange Assessment* (NRC, 1979), scientific involvement in data system planning, in data processing, and in data distribution is essential. Most of the decisions that must be made in creating, operating, and guiding the evolution of the CDC/DAN require a scientific understanding of the data sets and their interrelationships. It is therefore important that a Scientific Steering Committee (SSC) be established to work closely with the CDC/DAN in all phases of its development and operation. Solar-terrestrial research involves a broad range of disciplines, the data are taken by many different techniques using a wide variety of instruments and observation platforms from the Earth's surface to satellites in deep space, and correlative uses of the data often are of importance outside the particular program or area for which they were obtained. For these reasons, the SSC should be interdisciplinary and made up of researchers and scientific data management specialists from all aspects of solar-terrestrial physics.

Of equal importance is the staff of the central data catalog node of the CDC/DAN. They should have scientific backgrounds and devote some part of their professional time (20 to 50 percent) to research activities using the CDC/DAN. Similarly, staff with scientific backgrounds should play essential roles in data preparation, maintenance, and distribution at each of the nodes of the DAN. Only by interacting with the central data catalog as a user can the staff responsible for its maintenance and operation obtain an understanding of its utility, of problems that need correcting, and of opportunities for system improvements.

IMAGING DATA

Imaging data presents special problems in terms of data quantity, computational load for enhancements, and transformation codes for remote display devices. The uses of image displays fall into two broad categories that require dramatically different levels of access to the data: data surveys and detailed analyses.

Survey Requirements

For a large number of the activities that involve access to image data, the principal requirement is access to a visible representation of the image itself. For such tasks (e.g., creation of data catalogs, event identification, and instrument health verification), there is no requirement for access to the digital data that make up the image, nor for the intensive activities involved in contrast stretching and enhancement. Hardcopy images of reasonable quality or video representations would suffice.

One approach that meets the requirements of image surveys would be to produce standard "analog" video disks containing the images. The capacity of each disk is in the range of 50,000 images per side with access times of approximately 5 seconds, being consistent with interactive use. By taking advantage of the existing hardware, we believe that significant economies could be realized in satisfying these survey requirements.

Detailed Analysis Requirements

In contrast to the requirements for survey-type activities, the careful, quantitative analysis of measurements obtained by imaging instruments requires access to the digital image itself. These detailed analyses include determination of absolute intensities, correction of instrumental effects, assignment of geographical positions, and a host of contrast enhancement and transformation techniques for the study of spatial and temporal variations.

The best way to maintain these data is in a form as close as possible to the original sensor output with such reformatting as is appropriate to the instrument. As long as the algorithms for reducing, normalizing, correcting, and calibrating the data are updated and accessible, the number of times the entire set of images must be reprocessed and updated is minimized. If such an approach is adopted, a read-only medium such as an optical disk is most appropriate. The data density on an optical disk is extremely high, and the medium shows promise of being of archival quality. The necessary algorithms could be included together with the sensor data.

ANCILLARY DATA

Solving solar-terrestrial research problems requires certain ancillary data as well as the multiparameter data sets discussed earlier. The required ancillary data generally will vary with the problem being studied and may include universal time, local time, spatial location, coordinate transforms, instrument and platform attitude, and a variety of solar-terrestrial parameters and indices such as Carrington longitude, sunspot number, flare characteristics, K_p, auroral electrojet indices, and D_{ST}. Such ancillary data are an integral component of the data base and should be available for all data sets incorporated into the CDC/DAN.

INTERAGENCY COORDINATION

The data used for solar-terrestrial research is collected from experiments and monitoring programs supported by a variety of federal agencies (e.g., NASA, NOAA, NSF, DOD, USGS, and DOE). Frequently, these agencies have different policies as to what data to collect, whether it should be processed into archival formats, where it is to be retained, how long it should be available, who should have access to it, and whether there should be a charge for access to the data or cost recovery for copying. For example, NASA retains copies of data taken by spacecraft experiments and usually provides access without charge to researchers, whereas NOAA attempts substantial cost recovery. Thus some data in the system might be accessible without charge and some might be accessible only at a price. In principle, these differences can be solved; however, if problems arise, the SSC should contact the relevant agencies and request them to develop a coordinated policy for access to data for scientific research. It may be necessary to establish an interagency panel to develop such a policy.

INTERNATIONAL COOPERATION

Solar-terrestrial research is a global science, and no nation is able to independently collect or retain all data needed by this scientific field. This has been recognized for many years in the gathering of ground-based data and has stimulated the establishment of international networks of observatories for cooperative study of the sun, monitoring the ionosphere, and measuring geomagnetic variations. In space science, multinational programs that share satellite resources are becoming more frequent. About two-thirds of the organizations sending solar-terrestrial data to NOAA's National Geophysical Data Center/World Data Center-A for Solar-Terrestrial Physics are outside the United States. Through the World Data Center system, channels for routine and special data exchange are established and have operated successfully for 25 years. Periodic revisions of the *Guide to International Data Exchange* (ICSU Panel on World Data Centers, 1979) will be required in order to obtain data in formats most suitable for incorporation into CDC/DAN.

3

DATA ARCHIVES

Data archiving is the long-term retention of data so that it can be accessed by users who were not directly associated with the data collection and processing. Archiving data products from both operational and research projects is important in order to preserve the data for future use for synoptic studies and for reanalysis should new information (scientific perspectives and/or instrument reevaluation) become available. Because total program resources are limited, it is necessary that careful attention be given to the creation of useful, accessible archives during the earliest stages of planning future data collection programs. Solar-terrestrial research projects currently producing data must be reviewed to establish priorities for the use of existing resources to archive the data. Unique data collections from past projects that have not been archived should be identified and evaluated for inclusion in the archive system. Existing archives must be reviewed for compatibility with the CDC/DAN.

PLANNING ACTIVITIES

For too long, the archiving of solar-terrestrial data and analytical results from data processing have suffered from insufficient planning. Decisions concerning which data or products should be archived, their formats, and their availability to users often were made on an ad hoc basis without sufficient consideration of whether adequate staff and financial resources were available, whether the archive would be usefully accessible, and whether this was the most appropriate use of resources. As a result, there exist archives that are inadequately cataloged, collections of data that are of dubious value, and badly stored data of unique and possibly inestimable value that are inaccessible to users.

Generally, the funding provided for solar-terrestrial research projects is expected to support instrument design, data collection during a mission, data analysis, and data archiving. The funds allocated to data analysis and archiving are often inadequate for the task. Moreover, where problems are encountered during early phases of a project, funds for analysis and archiving are often diverted to solve them. Consequently, experimenters often find themselves without the resources needed to provide accurate, well-documented data sets that could be employed by other users.

Although there are examples of useful archives and of data collection programs for which successful archiving was planned, we believe that all future programs, both research and operational, must include planning during their earliest stages for creation of accessible data archives. This planning must involve active research scientists as well as representatives of sponsoring agencies and data center staff.

The involvement of active research scientists is necessary not only in the creation of archives, but also in their maintenance. This includes keeping data current and error-free, revising data sets, and making decisions on purging of data.

FACTORS APPLICABLE TO DATA ARCHIVING

To assure the smooth transition of data from aquisition to archiving, the retention of derived products, and the merging of data from future programs with archival data from the past, certain factors should be considered:

- adoption of common standards,
- storage media,
- Data catalogs, and
- responsibility for archive preparation and services.

Common Standards

Common standards for parameters such as time, spatial location, and units of measure should be adopted whenever possible and documented with the data so that both are available to users. Some standards, however, will differ with discipline, observing platform, instruments, and so on, and corresponding definitions must make such distinctions clear. For example, documentation should specify common definitions for:

a. date,
b. day number in year,
c. hour within day,
d. time interval for data averaging,
e. underlying models for derived data products, and
f. units for physically meaningful quantities, e.g., particle flux.

Storage Media

Storage media for archives should be appropriate to the type of data retained and the need for access to it. The choice should be made so that users will have the maximum practical accessibility consistent with data quantity. For example, digital data could be kept on-line on disks, stored on magnetic tape, or retained indefinitely on optical disks. Data can also be retained in publications and on film as tabular or analog representations. Each of these storage media have limitations and uses that must be considered in overall planning.

Data Catalogs

With implementation of the CDC/DAN concept, data catalog entries will be prepared at the time of data acquisition and will be supplemented with appropriate ancillary data. These original catalogs will evolve to include information about processed data, data products, analysis of data for selected events or phenomena, availability of correlative data, and software. Such catalogs will be retained in the archive, as needed, to provide access to distinct data collections or will be merged with more general archive catalogs. Provision must be made for continual catalog updating as subject archives are modified.

Responsibility for Archive Preparation and Services

Within the context of the CDC/DAN, each data collecting program must clearly assign responsibility for each level of data capture, processing, and summarization with retention either at distributed sites or in a central facility. Unless specifically planned otherwise, sponsoring agencies must assume responsibility for assuring the appropriate processing of data into archivable formats, for creation of archivable catalogs of the complete data collection, for the preparation of derived data summaries or other products with accompanying catalogs, and for the preparation of browse files by which general users may efficiently scan larger data bases connected to the CDC/DAN. Since data archiving is a continuing task, federal agencies must plan for the transfer of responsibility for the data sets, catalogs, and documentation. Interagency coordination, as noted in Chapter 2, is essential.

ARCHIVING DATA IN THE FUTURE

As noted earlier, new data programs in solar-terrestrial research must include well-defined plans for data capture, processing, accessibility during the program, and eventual archiving. The recommended concept is to incorporate the data from existing and future programs into the CDC/DAN. The need for archiving both raw and processed data exists; it is recognized that each type of data has different access and archive requirements.

Archiving of Raw Data

Often, attempts to retain the large mass of raw data collected but never processed are a waste of resources because the effort and expense required for a later investigator to penetrate the collection and extract meaningful results make such ventures unlikely. It could be more efficient to undertake a new program to collect the same type of data using more modern techniques. As the costs of multidisciplinary and multiplatform programs have increased, their frequency has decreased and the need for maximum exploitation of collected data has grown. As the volume of data collected by modern programs increases with the use of new technologies, only selected parts of the original data may be processed into usable formats and the vast majority will remain in its raw form. It is essential therefore that the raw data be archived and that the archive be structured so that the contents can be easily accessed, either to enhance or correct earlier processing or to permit the analysis of unprocessed data. Historical studies have extracted new ideas from the sparse but carefully kept records of earlier centuries that were taken and saved for reasons unrelated to the then current research interest.

Technological and procedural innovations can make the archiving of raw data easier. Application of the optical disk promises to make possible the efficient storage of large quantities of raw data, with longer life expectancy and less need for expensive maintenance than present magnetic tapes.

Some future programs (e.g., OPEN) are likely to employ centralized data acquisition with distributed data processing. Project participants will deposit verified data conversion algorithms with the central facility so that others can access selected raw data and process it into usable form on their computer facilities. This system will lead to an improvement in the creation of raw data archives because the necessary processing algorithms and catalogs of data and software will be available. Whether these are at the same cr at a different physical location, they will be accessible through the CDC/DAN.

The concept of archiving raw data must be flexible. If adequate processed data are retained and if the production of summary indices or other representations of the original collection are judged sufficient to maintain the historical record, then the raw data need not be maintained.

We recognize that the data generating capabilities of future programs must be considered in decisions about maintaining and servicing archives. Moreover, the often transient nature of software and storage media as computer hardware and operation systems evolve will require careful planning of archival architecture to ensure longevity. Guidance in making the difficult choices about archiving raw data must be the responsibility of knowledgeable scientists who use the data.

Archiving of Processes Data

Processed data include data reduced to physical units, gridded data, indices, and other derived data products that summarize masses of reduced data. Included in these processed data archives are data collected by ground-based, balloon, rocket, ship, and space platforms. It is essential that these processed data archives be independently accessible in the CDC/DAN.

Some future programs (e.g., UARS) plan to have a central data collection and processing facility. Here, also, the eventual transfer of processed data to archival facilities is visualized. Other programs, which plan to concentrate primarily on "event" data (e.g., OPEN), will routinely process data in a monitoring mode to create a summary file that participants can browse to select events for intensive analysis.

ARCHIVING PAST DATA

Some types of scientific data, usually lists of observed events, have been acquired for centuries. Examples are the dates of solar eclipses, lists of earthquakes, and sunspot diagrams. Early global scientific programs such as the International Polar Years generated a few data collections that were archived but not systematically disseminated. The concept of systematic observation schedules resulting in large data bases maintained at regional data centers was planned for the International Geophysical Year (IGY) and resulted in the creation of the World Data Center (WDC) system. In practice, the World Data Centers were associated with centers having national or regional responsibility and expertise in particular disciplines, and the IGY data collections often were merged with archives of earlier data.

Because of the success of the WDCs in meeting the needs of scientists for access to IGY data, they were continued beyond the end of that program, and they now are the largest repositories of geophysical data outside of industrial facilities. Data collections from specialized programs have usually remained with the institution responsible for their acquisition.

Often, archives of data sets that span many years are kept in a variety of formats including bound publications, loose sheets of paper, charts and maps, photographic plates, microfilm, microfiche, and electronic storage media. While some of the tabular and analog data have been converted to digital formats, much remains as originally recorded. To ensure archival continuity between data of the past and those to be collected in future programs, the CDC should include provision for the cataloging of all types of data.

As technology improves to make practical the compact storage and efficient access to large quantities of data, the problem of which data to keep for the indefinite future is somewhat mitigated. Although data collecting programs may emphasize short-term or event analysis and research, the creation of suitable summary data is still important for historical archives. For example, contemporary solar research may be intensive into the mechanism of energy storage, transformation, and release in flares but not particularly interested in maintaining the continuity of the sunspot number, which provides a correlative index of solar activity extending back to the seventeenth century.

To establish priorities for merging the older data with the new, to guide the application of new technology, to provide improved access to archives of past data, to identify synoptic data to be maintained and archived, and to identify data sets of marginal archival value, guidance will be required from appropriate groups of scientists. Such groups should also guide the data centers in the acquisition of unique but previously nonarchived data collections.

Finally, there is often a strong "oral tradition" in reducing and analyzing complicated spacecraft data. In particular, imaging data can often require subtle interpretations that may be difficult to define in the CDC. Therefore the data file in the CDC should include the name of contact persons or previous data set users.

APPENDIX: SYSTEMS CONSIDERATIONS

In this appendix we describe a variety of problems to be expected in the development and operation of the CDC/DAN. Many of these problems involve hardware and software compatibility considerations that invoke various levels of standardization as solutions. While we present examples of standardization techniques that can ease the CDC/DAN problems, we emphasize that standardization has to be done sensibly and voluntarily in the solar-terrestrial research community and must be implemented to encourage, not stifle, individual initiative.

NETWORKING AND COMMUNICATIONS

The rapidly emerging field of networking will be important to the development of CDC/DAN. Networking of computer systems provides new possibilities for scientists to perform their analyses with greater access to the data and with ready availability of processing power and graphics peripherals. Perhaps the most important benefit of networking will be the ease of cooperative research among investigators in different locations or institutions.

While it appears that the simple interchange of formatted alphanumeric data should pose no serious problems in the near future, the solar-terrestrial research community should realize that the existing profusion of networking concepts will give rise to compatibility problems for some years to come.

The goal of a network should be to deliver records of data, which can include programs, data, print output, graphics, or interactive commands. The user should not have to be concerned with the mode of communications. Some of the network nodes will be the principal repositories for specific data sets and will be the central analysis facility for that data set. Such nodes should have most, if not all, of their principal data set on-line at all times. While the capability of putting any portion of the data set on the network is necessary, the network would easily be overloaded by unrestrained data requests. The SSC should negotiate policies among the nodes to determine when large data sets should be transferred by mail.

Whatever the mode of communication, the system should permit the transfer of binary data without great inefficiency or effort. The programs in each computer should be isolated from the communications devices in order to be independent of the communication methods used.

Sample Network

An example of a network that has demonstrated effectiveness in solar-terrestrial research is the Space Plasma Computer Analysis Network (SCAN). The SCAN system is a data base management and computer network system that has been partially operational since January 1982. SCAN has linked computers together with dedicated communication lines at Marshall Space Flight Center, Los Alamos National Laboratory, Stanford University, Utah State University, University of Iowa, University of Texas at Dallas, and Goddard Space Flight Center. The network provides a means of quick and easy transfer of data, computer programs, manuscripts, and messages to other scientists on the network, thus allowing participating space plasma scientists to conduct correlative research. The SCAN computer network is in a star configuration with the central node at Marshall Space Flight Center. Current plans call for increasing the number of nodes within the next few years.

The SCAN system has proven useful in correlative research; it is an effective means of data exchange (both spacecraft and ground-based) and allows scientists on the network to interact with data management systems not only at the central archive, but also at any of the remote nodes as well, where large data bases exist. Software exchange is an important aspect of the system and has reduced software development costs. Besides the trading of data, sharing of computational resources has proven to be very fruitful. The network continues to provide a productive environment for correlative data analysis workshops. Although "help-file" oriented, prior knowledge of the types of data and its location is necessary to use the SCAN system effectively. In addition, the system utilizes DECNET software and thus restricts users to those who have computer facilities in the DECNET.

The CDC/DAN concept is based on more than just computer networking--the creation and maintenance of the catalog and the involvement of active research scientists are essential. The experience gained with the SCAN system will certainly be useful in guiding the development of the CDC/DAN. The SCAN itself could be a node of the CDC/DAN.

GENERAL SOFTWARE COMPATIBILITY PROBLEMS

The area of software compatibility and portability is crucial to the effective use of data in the coordinated, multidisciplinary studies that are required to answer the most important questions in solar-terrestrial research. In this section we describe problems of software compatibility in three areas: system software, scientific analysis software, and graphics. By being aware of these problems, the solar-terrestrial research community, through its own efforts, can partially alleviate them.

System Software

System software consists of the commands and control entries to sign-on to a computer, edit and transfer files, compile and run programs, and read and write from the peripheral devices such as tapes and disks. System software is, unfortunately, one of the least standardized areas of all software--not only is every manufacturer's operating system unique, but also some manufacturers have several different command languages that may be run on the same machine.

From the user's viewpoint, it is primarily the system command structure that is important. If the principles of the command structures, their syntactical rules, and commonly used keywords were standardized, the task of learning to use a new machine would be simplified. Currently, there are no standards for command structures, and none are expected in the near future. Some manufacturers, however, are standardizing command structures within their product lines, and some progress is being made in "universal" operating systems. The applicability of such systems to the solar-terrestrial research community in general and the CDC/DAN in particular should be studied by the SSC.

Analysis Software

Analysis software is or prime interest to the scientist. These programs enable the examination of data and the creation of a framework in which to understand its implications. Most analysis software, however, will not run on other computers and often requires extensive revisions each time it is used on a new machine. Although analysis software is written in high-level languages whenever possible, there is still a compatibility problem. However, most solar-terrestrial data analysis is done on computers from only a few manufacturers and FORTRAN is a high-level language that is available from all of the principal manufacturers. Table A.1 lists the 12 manufacturers of the most widely used computing machinery and the higher level languages that they support. While the table indicates that FORTRAN is a universally supported language, many of the FORTRANs have extensions and machine-specific options. Only the ANSI standard subset of FORTRAN is

TABLE A.1 Languages Available on Different Computers

	Languages					
Manufacturer	FORTRAN	PL/1	BASIC	PASCAL	APL	Other
Control Data Corp.	Y		Y	S		
Cray Research Corp.	Y					
Digital Equipment Corp.	Y	S	Y	S	S	C
Data General	Y	S	Y	S		ALGOL
Gould (SEL)	Y		Y	S		C
Harris	Y		Y		S	
Hewlett Packard	Y		Y	S	S	
Honeywell	Y		Y			C
IBM	Y	Y	Y	S	Y	C
MODCOMP	Y			S		
Perkin Elmer	Y			Y	S	C
UNIVAC	Y		S	Y	S	

Y = Available for all machines
S = Available for some machines
C = Computer language known as "C"

compatible among all manufacturers, but this subset does not cover many of the common utility functions that are needed for data and peripheral device access, such as to read a block of data from a tape, to manipulate a string of bits in memory, to convert from one type of floating point number to another, and to unpack the magnetic tape blocking structures of other manufacturers. These support utility programs are usually written by each institution with machine-specific calls and features.

Graphics

Graphics software and hardware permit the visualization of the complex interrelationships that characterize modern data sets. Most graphics systems, however, are cumbersome and difficult to use. Not only are graphics software packages different on almost every system, but also most systems have some hardware items for which unique software must be written for each use. Computer-generated graphics generally separate into two classes. The first, and most common, are the vector-oriented devices in which all pictures are composed of a series of straight lines. The second type, which has been more extensively used with the advent of inexpensive memory, is raster and bit-mapped graphics. In these latter devices the picture is composed of a myriad of dots that are defined at fixed positions within the image. Bit-mapped raster graphics devices can generate true images. Unfortunately, there are no standard software systems to handle raster graphics. The ACM and ANSI graphics committees, however, have been actively working on standards for user-level vector graphics software and on hardware interfaces. These activities are part of a worldwide effort, and an ISO software standard is expected soon.

CDC/DAN SOFTWARE

In view of software and communications compatibility problems, the solar-terrestrial research community will have to take some definitive steps to ensure cooperative data analysis and data exchange in the future. While the users cannot accomplish a great deal in the area of system software, there are a number of possibilities for improving the transportability of applications programs and data. Simplest among these would be to restrict the body of data analysis programs to the ANSI standard subset of FORTRAN.

All of the nonstandard machine-dependent codes would be segregated into functionally well-defined subroutines to minimize the reprogramming effort for a new machine. While FORTRAN may not be the best language for all applications and the ANSI standard subset of FORTRAN can be limiting in a number of areas, FORTRAN is the only language that is almost universally supported at present.

Problems in transporting software programs usually occur in the input/output (I/O) and manipulation of data. Among the most important system-dependent routines that are needed are the following:

- A routine to read and write physical records. Programs should not call the I/O routines of a machine directly; they should call an intermediate subroutine. This routine then uses the system I/O functions of the particular machine to carry out I/O tasks on physical devices such as tapes or disks.
- Routines to use for bit manipulation. These are used to pack and unpack data in variable sized bytes that may cross word boundaries. For example, 1000 numbers sequentially packed into 14 bits each could be unpacked into separate memory cells with one FORTRAN call.

It would greatly enhance software compatibility if the SSC with the staff of CDC/DAN would standardize the calls to these utility routines. Then, even though the routines would be installation dependent, the FORTRAN source for analysis programs would be transportable among systems that implemented the standardized routines. Since most solar-terrestrial data processing is done on only a few types of machines, the CDC/DAN should act as a depository for working standardized routines for each machine type.

In addition to the above, there are a selection of other routines that should be readily available at each installation on the DAN:

- routines to block/unblock fixed length logical records,
- routines to block/unblock variable length logical records,
- routines to unblock systems records from various computers,
- routines to convert floating point words from various machines,
- sort/merge programs with flexible sort key manipulation,
- one-word sorts in memory,

- two-word sorts in memory,
- routines to convert back and forth between sequential day numbers and year-month-day, and
- simple access routines for commonly used data sets at the central facility.

Most data sets should have data access routines to handle much of the job of unblocking and unpacking the data. The methods used for the access routines should be as modular and conceptually as simple as possible so that learning time and complexity are kept low.

The basic data archive system should be a file management system that would allow user programs to access the files directly.

Some data archive procedures demand an unnecessary degree of conformity before a data file can be entered into the system and made available to users. The CDC/DAN system should be designed so that a new data base can be added and used within hours--improvements can be made later if they are necessary.

Documentation and inventory functions generally require an emphasis on convenience to the user, low learning time, ease of updating, and reasonable cost. These data sets of inventory information usually have rather small amounts of data, so there is not a concern with slow software if user convenience is being improved. Detailed inventories may have large volumes of data, and should typically be treated as ordinary data sets handled by file management techniques.

SYSTEMS COMPATIBILITY

The problems encountered by users of data prepared on different computers are partly caused by the systems software, and partly by the hardware. Many of these problems would be alleviated if points such as the following are considered in hardware selection:

1. *Read/write data*. Make certain that the routines to read/write a record from devices are easy to use. Operating systems should allow long records (up to at least 16-bit controller architectures) to be read, and longer records could be permitted if special procedures are used.

The reading routines should be able to return the length of the record that has been read, and its status (good, good after correction, parity, end of file, end tape). Typically, the call for I/O status should be separate from I/O initiation so that buffered I/O is permitted.

2. *Routines for bit manipulation*. The standard software package for each machine should include routines for bit manipulation. In these routines, the groups of bits to contain each number are of variable length; that is, they may be 3 bits, 27 bits, and may cross word boundaries. The maximum size of a bit group is the word size of the machine. The problem is that such routines are not available on the system library of most machines. This has caused a good deal of needless trouble in data exchange.

3. *Strings of bits or characters*. A user should be able to consider a string of bits or bytes of data as a sequential array of data in memory, in which the word boundaries of a given machine are relatively transparent to the unpacking process if routines such as described above are available. A problem occurs in that the byte order in various computers in the internal representations differ; then a subroutine is necessary to reorder the bytes before bit manipulation routines can be used.

4. *Facility to manipulate characters*. FORTRAN 77 does not include Encode/Decode statements and will cause a problem with certain tasks. Many computing companies will provide a Hollerith extension to FORTRAN 77. Without such software capability, individual users develop many different ways of handling the problems, thus complicating the tasks of software exchange. Statements for character manipulation such as Encode/Decode should be implemented on all machines.

The maximum size of data buffers permitted in Encode/Decode statements has often been limited to the approximate length of print lines, usually 150 characters. There is no need for such a restriction; it needlessly complicates the unpacking of an array of characters.

5. *Control over data transformations*. Users should be able to control data transformations. For example, systems have caused problems when they always translate characters in a certain data path. When the input data contains characters that cannot be translated, information is lost.

6. *Fast memory move*. Data manipulation often involves the movement of small blocks of data within computer memory. A subroutine, optimized for speed, should be available to move blocks of words or data within memory.

REFERENCES

ICSU Panel on World Data Centers (1979). *Guide to International Data Exchange*, International Council of Scientific Unions, 113 pp.

NRC (1979). *Geophysical Data Interchange Assessment*, Committee on Data Interchange and Data Centers, Geophysics Research Board, National Research Council, National Academy of Sciences, Washington, D.C., 70 pp.

NRC (1980). *Solar-System Space Physics in the 1980's*, Committee on Solar and Space Physics, Space Science Board, National Research Council, National Academy of Sciences, Washington, D.C., 82 pp.

NRC (1981). *Solar-Terrestrial Research for the 1980's*, Committee on Solar-Terrestrial Research, Geophysics Research Board, National Research Council, National Academy Press, Washington, D.C., 143 pp.

NRC (1982). *Data Management and Computation, Volume 1:* Issues and Recommendations, Committee on Data Management and Computation, Space Science Board, National Research Council, National Academy Press, Washington, D.C., 167 pp.

GLOSSARY

ACM	Association for Computing Machinery
AE	Atmospheric Explorer
AMPTE	Active Magnetospheric Particle Tracer Explorers
ANSI	American National Standards Institute
CDC	Central Data Catalog
CODMAC	Committee on Data Management and Computation
CRRES	Combined Release and Radiation Effects Satellite
DAN	Data Access Network
DE	Dynamics Explorer
DOD	Department of Defense
DOE	Department of Energy
HILAT	High Latitude Ionospheric Research Satellite
IGY	International Geophysical Year
I/O	Input/Output
ISEE	International Sun-Earth Explorer
ISO	International Standards Organization
NASA	National Aeronautics and Space Administration
NGDC	National Geophysical Data Center
NOAA	National Oceanic and Atmospheric Administration
NRC	National Research Council
NSF	National Science Foundation
NSSDC	National Space Science Data Center
OPEN	Origin of Plasmas in the Earth's Neighborhood
PPDS	Pilot Planetary Data System
RGON	Remote Geophysical Observation Network
SCAN	Space Plasma Computer Analysis Network
SME	Solar Mesosphere Explorer
SMM	Solar Maximum Mission
SSC	Scientific Steering Committee
S^3-A	Small Scientific Satellite-A
UARS	Upper Atmosphere Research Satellite
USGS	United States Geological Survey
WDC	World Data Center