X-601-69-415
PREPRINT

NASA TM X 63712

A FRAMEWORK FOR FUTURE DATA CENTERS*

JAMES A. FAVA

SEPTEMBER 1969

GSFC

GODDARD SPACE FLIGHT CENTER
GREENBELT, MARYLAND

*To be presented at the ASIS 32nd Annual Convention, San Francisco, California, October 1-4, 1969.

X-601-69-415

A FRAMEWORK FOR FUTURE DATA CENTERS

James A. Fava
National Space Science Data Center

September 1969

GODDARD SPACE FLIGHT CENTER
Greenbelt, Maryland

ABSTRACT

A discipline-oriented data center is an essential element for many segments of the scientific community. To be effective, the center must have: (1) professional staffs in the disciplines associated with the center and in the ADP fields, (2) sufficient equipment and software to reformat and change the form of data, and (3) an information retrieval system concerning its holdings. A variety of services must be performed by the center. Most data should be made available to all users, and the paper suggests that all data centers should charge for their output services.

CONTENTS

ILLUSTRATIONS

A FRAMEWORK FOR FUTURE DATA CENTERS

INTRODUCTION

The term "data" used by itself means different things to different people. However, when those involved with a data center discuss data, they generally use an adjective to modify the word and talk about a specific class of data. For example, space science data may be considered to be derived from quantitative measurements of phenomena taking place within the immediate vicinity of the earth and extending into interplanetary space, to include the planets.

Both space science and other environmental data may be collected for a number of reasons. The primary reason may be one of basic research in which an attempt is made to find out what is there, how it varies with time and space, as well as to understand its properties in terms of fundamental processes and principles. On the other hand, there may be an operational mission which must be supported. Regardless of the initial reason, much of the data, either in the fundamental or in a converted form, may be very useful to others—and for entirely different reasons. In many cases these data are extremely expensive to obtain (a large satellite program consisting of a number of satellites may cost hundreds of millions of dollars) and may take as much as 4 or 5 years of effort on the part of an individual research group. In addition, the actual volume of these data has become so large that it would be impractical to publish all the data. Although the initial use of these data may have already been exploited, the preservation of such data for secondary use is important. This secondary use often demands that the user be an expert in both data processing and the associated scientific discipline.

Data centers have been established to acquire and preserve the data originating in a number of scientific fields, e.g., oceanography, meteorology, space science, etc. Such centers differ from large documentation centers because they are involved with huge masses of quantitative measurements which may exist on microfilm, hard copies, computer printouts and plots, magnetic tapes, etc. If the documentation centers were to handle individual words in the documents, they would begin to approach the data volume problems of the data centers. One example is given to put this in proper perspective. A single ionospheric experiment operating for about 2 to 8 hours per day for 7 years has generated more than 1,360,000 composite ionograms.(1) Each ionogram represents 744 frequency scans recorded on analog tape during a time period of 12 seconds. These are presently being stored on 340,000 linear feet of microfilm.

A comparison between present data collection, reduction, and evaluation technology and that of 10 years ago illustrates how the problem of data centers have been magnified. Before the use of computers, it was customary to display raw data by manually recording individual measurements or by using chart recorders. The data generators and primary data users, such as scientists and engineers conducting research or an operational mission, would then work with those individual measurements using desk calculators, slide rules, and pencil and paper to reach their conclusions or present their data.

With the advent of computer technology, these users have been relieved of the problem of manipulating the individual measurements and are able to specify the repetitive operations to be performed by computing devices. Therefore, they are able to work effectively with data bases which are expanded by factors of thousands to millions. Let me illustrate this with an example. Blair and Ficklin (2) pointed out that the Stanford University/Stanford Research Institute experiment on OGO (Orbiting Geophysical Observatory) - 1 generated 2 x 10^9 bits of information per year. In order to process this quantity of data they devised a new data handling process. The output of the process is 16-mm cine films on which data are plotted along with the pertinent orbital and geophysical parameters. In this way, they were able to review in one day all the data generated by the experiment during a year. In addition to aiding in the analysis of the data, this technique greatly simplified the storage problem. The data from about 30 standard computer tapes were reduced to one 400-foot 16-mm cine film; the data generated in a month were reduced to about five reels of film.

During this same 10-year period, the number of active data generators and primary users has increased tremendously. These factors, plus the present tendency for research efforts to cut across disciplinary boundaries, result in the present increased need for data centers to have extensive capabilities and full-time professional staffs.

One such data center, the National Space Science Data Center (NSSDC), was established by NASA in 1965 to handle the data originating from space science experiments. The volume of data generated in these areas is probably larger than in most others, and the diversity of the user community is quite extensive and worldwide. For these reasons, it is felt that the experiences that have been encountered over the past few years would apply to data centers in other fields, e.g., medicine, social science, education, etc. Thus, this paper is an attempt to apply the experiences of the National Space Science Data Center to a generalized data center and to provide a framework around which data centers in the future can be developed. It contains a description of a generalized data center, a discussion of some of the broad functions which such a center should perform, and the relationship of this activity to the general flow of information throughout the professional and user communities associated with the particular discipline.

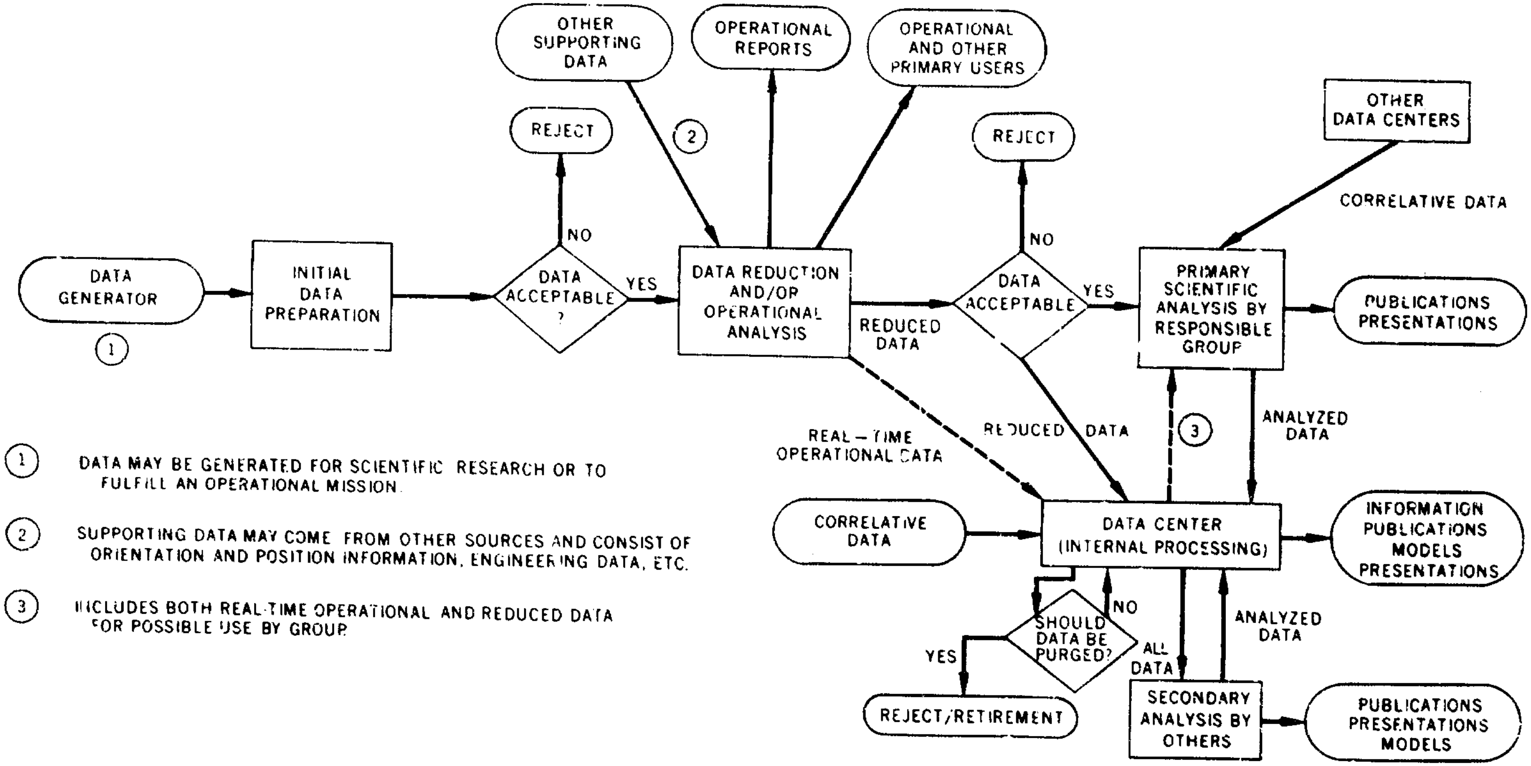
DATA FLOW

For the purposes of this paper, a single space science data measurement performed at a given location and time becomes a data point. A data point can be considered as a unit of fundamental information obtained from a sensor. Ludwig (3) has pointed out that a data point would generally correspond to 8-10 binary digits. In addition, he has indicated that the telemetering bit rate has increased from a few bits per second to as much as 64,000 bits per second for some of the more complicated NASA satellites. During this time period (1961-1967), the number of data points per day increased from 3 to 237 x 10^6. In order to use a data point, associated information such as time, spacecraft location, and attitude, certain housekeeping information, and appropriate characteristics of the basic measuring device often are required. Therefore, the data center must concern itself with both the basic data point measurement and the other associated information.

Once data are obtained, say from a satellite, some initial preparation may have to be accomplished to make the data useful. They may pass through an acquisition station and be relayed over a communication link to a processing facility. At this point, mechanical, electrical, computational, or other techniques, may be applied in order to change the data from one form to another, e.g., analog to digital. The data could then flow to an experimenter or, for use in a real-time mode, to an operational unit. In both instances the data may be processed and thus reduced into a useful, ordered, or simplified form for operational purposes or for scientific analysis. An idealized picture of data flow is shown in Figure 1.

The actual time involved in the flow from source to the center may range from weeks in the case of photographs to years in the case of some satellite data. In the latter case, individual scientists are responsible for the general conduct of the experiment and the subsequent primary analysis of the data. To be useful, good and valid data with the necessary documentation to adequately describe the experiment and the characteristics of the measuring sensors should reach the data center. It may not be necessary or feasible in some instances for a center to acquire all useful data. By maintaining a directory of specialized data bases, the data center may call upon, or refer the user to, these peripheral data collections.

The issue of quality cannot be emphasized enough. As shown in Figure 1, quality control of the data is a continuing effort throughout the data flow process. There is no valid reason for expending money and effort to preserve data of questionable quality.


DATA GENERATOR
1
INITIAL DATA PREPARATION
DATA ACCEPTABLE ?
NO
REJECT
YES
OTHER SUPPORTING DATA
2
OPERATIONAL REPORTS
OPERATIONAL AND OTHER PRIMARY USERS
DATA REDUCTION AND/OR OPERATIONAL ANALYSIS
REDUCED DATA
REAL – TIME OPERATIONAL DATA
DATA ACCEPTABLE
NO
REJECT
YES
REDUCED DATA
OTHER DATA CENTERS
CORRELATIVE DATA
PRIMARY SCIENTIFIC ANALYSIS BY RESPONSIBLE GROUP
PUBLICATIONS PRESENTATIONS
3
ANALYZED DATA
CORRELATIVE DATA
DATA CENTER (INTERNAL PROCESSING)
INFORMATION PUBLICATIONS MODELS PRESENTATIONS
SHOULD DATA BE PURGED?
NO
YES
REJECT/RETIREMENT
ALL DATA
ANALYZED DATA
SECONDARY ANALYSIS BY OTHERS
PUBLICATIONS PRESENTATIONS MODELS

(1) DATA MAY BE GENERATED FOR SCIENTIFIC RESEARCH OR TO FULFILL AN OPERATIONAL MISSION.

(2) SUPPORTING DATA MAY COME FROM OTHER SOURCES AND CONSIST OF ORIENTATION AND POSITION INFORMATION, ENGINEERING DATA, ETC.

(3) INCLUDES BOTH REAL-TIME OPERATIONAL AND REDUCED DATA FOR POSSIBLE USE BY GROUP.

Figure 1. Data Flow

The type and amount of data that will flow into a data center will depend upon the activities of the groups which it supports. In the case of the NSSDC, the amount of data that can be expected from a particular satellite will depend upon its mission and the number of experiments carried on-board. Figure 2 shows the number of successful experiments flown and the number in which some data has been acquired by NSSDC.

DISCIPLINES	SUCCESSFULLY FLOWN EXPERIMENTS *	SOME DATA AT NSSDC**
IONOSPHERES & RADIO PHYSICS	86	12
PLANETARY ATMOSPHERES	96	20
PARTICLES & FIELDS	396	54
SOLAR PHYSICS	28	11
ASTRONOMY	35	3
PLANETOLOGY (INC. SELENOLOGY)	127	45
TOTAL	768	145

* AS OF MAY 17, 1968

** AS OF SEPTEMBER 17, 1968

Figure 2. Data on Hand vs Successful Experiments

CHARACTERISTICS OF A DATA CENTER

Although different data centers may have unique characteristics, they also have many features which are common. A data center, although discipline-oriented (e.g., to meteorology, space, oceanography, medicine, etc.), is responsible for the archiving and subsequent use of the data obtained from a particular segment of the scientific community or a data generation activity. If the data cannot be handled by a diversified spectrum of users with a minimum of effort, they should remain with the original investigators and be noted as available. The data center should have at least the following capabilities:

- An information system about both the data in the data center as well as the availability of the specialized data collections that exist in other locations

- Microfilming, digitizing, and computing equipment with enough flexibility to be able to accept data in almost any form and be able to provide the data in a variety of ways so that it is readily usable by a diversified user community

- A specialized technical library and automated document retrieval system

- A professional staff in the scientific disciplines that carries on analysis and synthesis of the data

- A professional staff in the computer and information sciences that develops information systems, analysis routines, storage, and retrieval techniques based on latest capabilities in computers, data storage devices, communication links, and interactive input/output devices.

MAJOR FUNCTIONS OF A GENERALIZED DATA CENTER

Three of the more important functions are discussed in the following paragraphs. These are: (1) acquisition, (2) analysis, and (3) user services and products. A data center must perform a number of operations on the data that are similar to the operations performed by a documentation center, e.g., cataloging, indexing, storing, retrieving, duplicating, etc.; however, these will not be discussed.

Acquisition

To be successful, a data center must have a very active acquisition effort. Those responsible for acquisition must be professionals, technically competent in their disciplines. During the early planning phases of any large-scale, data-gathering program—whether for research, survey, or operational purposes—the acquisition specialists of the appropriate center(s) should be involved. They could suggest data processing techniques which would optimize the use of the data both for the goals of the program and for the input/output functions of the center. In addition, the collection of the necessary correlative data can be anticipated at this time. Individuals involved with smaller scale research efforts should be advised by the center as to the best means to preserve the data for use by others. A flexible input/output system of the data center is of great advantage in communicating these data to a wide variety of users.

Once a data-gathering program is approved, the acquisition staff must start working with the generator during the time that data reduction plans are being formulated. It is at this time that the function of the center and the problems associated with archiving the data must be clearly understood by the generators.

While working with the generators, the center representatives must maintain a flexible but persistent schedule. This schedule should allow for the rejection of data of questionable quality and of data with inadequate documentation and allow for slippages of program schedules. The data to be submitted to a center should be in a form which requires the least expenditures of resources—money, manpower, computer time, etc., considering both the data generator and the data center. Normally this form of data will be a natural product of data processing and only needs to be preserved at the proper point. Again in the case of NSSDC, two types of data are acquired. These are reduced data and analyzed data records. Karlow and Vette (4) have defined these as follows:

> "Reduced Data Records - Data records prepared from raw data records by a compacting, editing, correcting, and merging operation performed under the supervision of the principal investigator. Data in this form contain all the basic usable information obtained from the experiment and generally include the instrument responses measured as functions of time along with appropriate position, attitude, and equipment performance information necessary to analyze the data in an independent fashion. The engineering corrections such as temperature, voltage, dead time, gain changes, and other similar corrections to the instrument response will have been made. Unusable noisy data and periods of questionable instrument performance will have been removed as well as duplicate portions of information. Time averaging and the conversion of the instrument response to physical units will not have been accomplished in most cases. Visual data, such as photographs derived from data processing techniques, may also be considered as reduced records."

> "Analyzed Data Records - Data records prepared from reduced data by the principal investigator, his co-workers, and other space scientists which display the scientific results of the experiment. In general, the physical quantities derived from the sensor responses are displayed in various appropriate coordinate systems and correlated with other geophysical measurements. The results may be time averaged over meaningful intervals, displayed in the form of parameters of specific physical models or theories or as best-fit parameters of empirical descriptions. This form may include charts, graphs, photographs, and tables which are the results of data processing and analysis techniques employed by the analyzing scientist. Examples of these appear in his published works, but the total number are usually too large to be published in their entirety."

A data center may collect data which have been recorded on (a) microfilm, (b) digital magnetic tapes, (c) photographic positives and negatives, (d) graphs and roll charts, (e) microfiche, (f) computer generated plots, or (g) printed material.(5) Since it is necessary to have special-purpose equipment to handle analog tape data, a center should not normally be expected to accept such data. Figure 3 shows the holdings of the NSSDC in these various categories.

MEDIUM	AUGUST 1967	MARCH 1969
SHEETS AND BOUND VOLUMES, SHEETS	175,000	257,000
DIGITAL MAGNETIC TAPES, ½" × 2400'	291	2,864
MICROFILM, 100-FT ROLLS	7,800	11,001
PHOTOGRAPHIC FILMS		
9½" WIDTH, LINEAR FEET	14,000	18,000
70-MM WIDTH, LINEAR FEET	33,200	177,000
35-MM WIDTH, LINEAR FEET	0	759,000
4×5 INCH, EACH	2,100	2,445
8×10 INCH, EACH	0	400
16×20 INCH, EACH	93	93
20×24 INCH, EACH	2,200	7,600
PHOTOGRAPHIC PRINTS		
9½" WIDTH, LINEAR FEET	0	9,000
70-MM WIDTH, LINEAR FEET	0	14,000
8×10 INCH	600	3,100
11×14 INCH	200	500
16×20 INCH	93	93
20×24 INCH	2,200	5,000

Figure 3. Growth of the Data Base at NSSDC

Analysis

When appropriate, data centers should develop a strong capability for analysis to meet the user needs for various data products. The end products of such analysis should be new and useful products, compilations, or models which are desired by the user community. Only in this way will centers be able to attract professionals of sufficient competence in the various disciplines to guarantee the proper data inputs and internal data management. The creation and documentation of a particular model of some environmental parameters could be considered as a state-of-the-art survey in a scientific field as well as a useful new output. Such a model, in lieu of a well-developed theory, may serve to identify certain data as no longer useful. Thus, these data subsets could be retired from the active data base or purged completely. Any high-volume data center must establish a data retirement or purging system. It would not be practical to acquire and archive all data. However, decisions involving purging or retirement should normally be left to the judgment of professionals and not be made by an arbitrary policy or procedure established by some administrative group.

Once a data center is able to attract competent professionals and develops a strong capability for analysis, several information analysis centers will evolve within the center. It must be realized that both the analysis and information-type functions require a number of years to develop. The data center must reach a certain minimum size, both as to resources and the types and amounts of data, before it can really become effective. This minimum size will depend upon both the discipline(s) associated with the center and the segment of the scientific community to which the center is responsive.

User Services and Products

There is no valid reason for having a data center if the center cannot provide a wide variety of services and products to users. Everyone concerned with data centers must realize that a center will probably never have sufficient resources to satisfy all user demands for service.

Services and products of a data center should include, but are certainly not limited to, the following:

1. Disseminating catalogs and data center publications

2. Retrieving, reformatting, and furnishing data

3. Furnishing necessary space and use of facilities for visiting scientists

4. Preparing and publishing models

5. Evaluating and analyzing data to meet individual requests

6. Summarizing and preparing graphic displays

7. Providing data directories and referral services

8. Consulting, reducing, and processing data

In many instances the major secondary users do not require the data per se, but require products that are derived from extracting, compiling, evaluating, reformatting, and synthesizing the data. Such products may be charts, atlases, models, statistical studies of properties and phenomena, handbooks, etc. The users of these products may not be the scientists intimately involved in the particular discipline. More commonly, they would include such groups as (a) scientists in related disciplines, (b) engineers and designers, (c) planners, (d) management, (e) operational activities, (f) educational activities, (g) recreational activities, (h) commercial activities, and (i) general public.(6) The use of the data center will grow in relation to its useful products. For example, Figure 4 showe the rate of growth of the number of requests received by NSSDC.

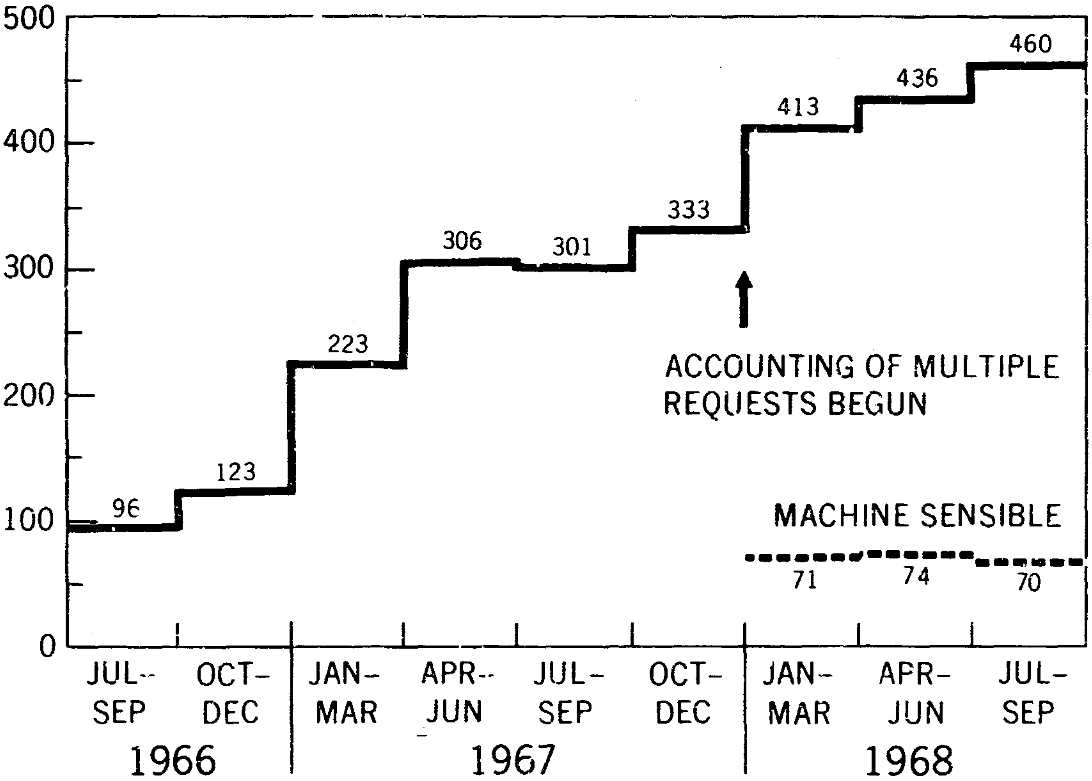


Figure 4. Growth of Requests. As of January 1968, requests requiring machine processing were identified apart from other requests, and single requests requiring different forms of data were treated as multiple requests.

THE DATA CENTER CONCEPT IN THE OVERALL SCIENTIFIC AND TECHNICAL INFORMATION SYSTEM

It should be ˬmphasized that a data center does not replace any element in an information system serving a particular scientific discipline. The center merely represents a new addition in the overall system. It is essential in those fields where vast amounts of data are generated at considerable expense which have a wide use outside the specialized scientific or operational activity which generated the data. The primary communication of information within the particular discipline and its peripheral areas should continue to be provided by the professional societies, meetings, publications in journals, and technical reports. The mission-oriented and cross-disciplinary information analysis centers are not replaced by the activities of the large data center. However, new information analysis centers in the disciplines covered by a data center will evolve.

Discipline-oriented data centers, e.g., environmental sciences, medicine, etc., could become subsets of a "National Data Center Subsystem" in the evolving U.S. National scientific and technical information system as described by Simpson. (7)

AVAILABILITY OF DATA TO USERS

For government-funded data centers, any U.S. citizen should be allowed to purchase the output, except for classified data. Over the past year or so there has been a gradual shift in the user-charge policies of both the documentation and information analysis centers. Many such centers are beginning to charge for their services. While they may be able to recover a part of their output cost, it is doubtful if such centers would ever be totally self-sufficient. Data Centers should also charge for their services, and a uniform user-charge policy should be adopted for all documentation, information analysis, and data centers.

The interchange of data on an international level should be encouraged. Of course. there will always be certain classes of data - classified, proprietary, etc. - that will not be exchanged. To facilitate the international exchange of data in the environmental sciences, World Data Centers were established in 1957 to support the International Geophysical Year. National data centers in the U.S. concerned with environmental data should support these World Data Centers.

RELATIONS BETWEEN DATA CENTERS

Because data centers will, I believe, tend to become discipline-oriented and will tend to serve the needs of a particular portion of the scientific community, there does not appear to be any requirement for a monolithic data center or for high-speed data links among all data centers. There is, however, a genuine need for close coordination and cooperation among such data centers both now and in the future. This would facilitate the identification of problem areas, the reduction of unnecessary overlap, and the development and spread of technological advances in storage, manipulation, and retrieval.

In addition, each data center should be aware of the holdings and services of the others so that requests may be funneled to the correct center for action. It is quite possible, with advances in high-density storage media, that high-speed links to data on-line will become a way of life. For example, at some time in the future, a user may have access to the data on-line using a console at his own location.

REFERENCES

1. Dubach, L. H., Private Communication.

2. Blair, W. E., and B. P. Ficklin, "Summary of Digital Data-Processing Systems for the OGO SU/SRI Very-Low-Frequency Experiments," Stanford Research Institute, Menlo Park, California, July 1967.

3. Ludwig, G. H., "Space Sciences Data Processing," NASA TN D-4508, May 1968.

4. Karlow, N., and J. I. Vette, "Flow and Use of Information at the National Space Science Data Center," NASA/NSSDC 69-02, January 1969.

5. Vette, J. I., "The Operation of the National Space Science Data Center," NASA/NSSDC 67-41, October 1967.

6. Pritchard, D. W., "Characteristics of a National Marine Data Management Program Required to Meet National Marine Needs," Contribution 128 of Chesapeake Bay Institute, The Johns Hopkins University, 1968.

7. Simpson, G. S., Jr., "The Evolving U.S. National Scientific and Technical Information System," Battelle Technical Review, pp. 21-28, May-June 1968.