Data Management at the National Space Science Data Center

J. I. Vette* and N. Karlow†
NASA Goddard Space Flight Center, Greenbelt, Md.

The primary function of the National Space Science Data Center (NSSDC) is to provide the means for the dissemination and analysis of space science data beyond that provided by the original experimenter. The purpose of NSSDC is not that of a data processing center. However, the recognition of important secondary uses of data generated for operational needs is vital. Since the requirements placed on operational centers to accommodate this secondary function are minimal, the experiences of NSSDC in the secondary uses of space science data will be discussed. With a present satellite generation rate of 10^{12} bits per year of space science information, the data management problems become significant, and considerable data processing is required before maximum utilization of the data base can be realized. In addition, the Data Center must be concerned with an information system to handle documentation, performance data, instrument calibrations and characteristics, and management information.

Introduction

THE National Space Science Data Center (NSSDC) was established in 1965 with the primary function of providing the means for the dissemination and analysis of space science data beyond that provided by the original experimenter. To fulfill this mission, the Data Center is responsible for the active collection, organization, storage, announcement, retrieval, dissemination, and exchange of space science data obtained from satellite experiments, sounding rocket probes, and high-altitude aeronautical and balloon investigations. Thus, as can be seen from this description, the Data Center is not an operational data processing center in the strictest sense of the word. However, considerable data processing is required to obtain maximum utilization of the data base. For example, with a present satellite generation rate of 10^{12} bits of space science information per year, the data management problems become very important.

In this context, it is critical to understand the secondary uses of these data and what is involved. Both space science and environmental data may be collected for a number of reasons. On the one hand, there may be an operational mission that must be supported. On the other hand, the primary reason may be one of basic research in which an attempt is made to find out what is there and how it varies with time and space, as well as to understand its properties in terms of fundamental processes and principles. Regardless of the initial reason, much of the data, either in the fundamental or converted form, may be very useful to others for entirely different reasons. Although the initial use of these data may have been exploited, the preservation of such data for secondary use is important for at least two reasons: 1) in many cases these data are very expensive and time-consuming to obtain, and 2) the actual volume of such data has become so large that it would be impractical to publish all of it.

Various secondary user groups include space scientists other than the principal investigators, scientists in related fields, engineers and systems planners, and educational activities. To satisfy the needs of these groups, specialists in the various space science disciplines, systems analysis, computer programing, data processing, technical writing, publication procedures, and reproduction are required. In short, both an automated data processing system and an information system are required to handle the numerous magnetic tapes, cards, pictures, microfilm, and copies of written, graphical, and tabular materials, and to prepare the Data Center holdings for effective secondary use by others interested in conducting space science investigations. This preparation often involves independent analysis on the part of Data Center scientists to properly service the total user community.

Because the requirements placed on operational centers to accommodate this secondary function are minimal, the purpose of this paper is to discuss NSSDC experiences with the secondary uses of space science data, the data processing system developed to process these data, the interfaces in the use of primary and secondary data, and the integrated information system to support the use of the data, current and future, at NSSDC. Hopefully, these experiences will prove useful to those concerned with operational systems who will be interfacing with data centers such as ours.

Secondary Uses of Space Science Data

Two of the biggest problems facing those concerned with data processing are the tremendous amount of data produced by satellite measurements that must be processed and analyzed and the wide diversity of space experiments. These data serve many different purposes: engineering measurements are vital to the advancement of techniques and procedures that lead to more sophisticated and reliable spacecraft systems; applications satellites are concerned with communications, navigation, weather, and Earth resources; biomedical experiments assist the manned flight program; and scientific experiments measure those quantities which will lead to a general understanding of the natural phenomena.

One must, however, understand the different philosophies behind the collection of these categories of data. The useful lifetime of engineering data is generally much shorter than than for scientific data, since rapid strides in technology soon produce new devices with different and more desirable properties. Likewise, operational data is used in near real time for most purposes. Consequently, the dissemination of this type of information must occur rapidly. The nature of space science measurements, on the other hand, requires that a much larger active lifetime be provided for the data. The processes being observed are not yet completely understood, and they interact with each other in many different ways. As a result, there are significant variations with time and location. This

Presented as Paper 70-320 at the AIAA Earth Resources Observations and Information Systems Meeting, Annapolis, Md., March 2–4, 1970; submitted March 12, 1970; revision received June 22, 1970.

* Director.

† Technical Information Specialist.

requires large volumes of data to obtain relations, patterns, and interactions. As new ideas develop in understanding the phenomena, scientists will want to take a new and different look at the existing data, and these analyses will lead to new results which may have little resemblance to the original intent of the individual experiments. Thus, this fund of new scientific knowledge will continue to grow.

But much work must first be done to prepare the data for independent analysis in the future by other scientists. The first step is to collect and preserve this data in the proper form. Raw data transmitted directly from the satellite is not the best form because it is impossible to acquire all the supporting information that is necessary for its independent use. Reduced data records are the basic records acquired by the Data Center. A discussion of the processes involved in going from raw to analyzed data is given in Ref. 1. However, to insure an understanding of terms, a definition of reduced data records, as viewed by NSSDC, is given here.

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.

For some scientific uses of the data, it is not necessary to reanalyze or re-evaluate. In most cases the interpretations given by the original investigator are the most valid and respected ones. These analyzed results are often incorporated with those of other experiments in order to gain new understanding of the various phenomena. For this reason, the Data Center must also collect analyzed data records which are defined as follows.

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 is usually too large to be published in its entirety.

To assist secondary users such as engineers or system planners, a new product—the space environment—must be developed from the raw materials which flow into the Data Center. Most knowledge of the space environment comes directly from the scientific measurements. However, only a small segment comes from each individual experiment. It is necessary to study the analyzed and/or reduced data from many experiments in order to obtain a fairly comprehensive description of the space environment. This translation or synthesis into useful data summaries, compilations, or environments is a natural professional activity for those space scientists associated with the Data Center. It represents a type of analysis not done very often by the original investigators and contributes a useful product for dissemination to all user groups, including space scientists.[2]

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 new model could also serve to identify certain data as no longer useful. Thus, these data subsets could be retired from the active data base or purged completely.[3]

The type and amount of data that will flow into the Data Center from a particular satellite will depend upon its mission and the number of experiments carried onboard. Of the 1106 successful experiments flown as of October 31, 1969, some data has been acquired by NSSDC for 222. The actual time involved in the flow from source to the Data Center may range from weeks in the case of photographs to years in the case of some satellite data. In the latter case, scientists must be given sufficient time to plan and conduct the experiment and the subsequent primary analysis of the data.

These data can arrive at the Data Center on 1) magnetic tapes, 2) microfilm, 3) microfiche, 4) photographic positives or negatives, 5) graphs and roll charts, 6) computer-generated plots, or 7) printed materials. To give you an idea of the data at NSSDC, Table 1 shows the growth of the data base at NSSDC. More will be said on this subject later. Of course, having the data is not enough. We must also have an information system which can retrieve facts about the data, satellite orbits, on-orbit performance, instrument characteristics, transmitter frequencies, and other supporting information such as funding information. Access and retrieval of this information in a variety of ways serves the management community as well as the other users previously described. The value of having such supporting information and the reduced data to incorporate into a university graduate program has been pointed out by Dessler.[4] He stated that the cost of a space-hardware program could perhaps be reduced from about $400,000 per Ph.D. to about $100,000 by carrying out one or two space experiments and supplementing this with the analysis of data obtained from NSSDC. Another example of the secondary use of space science data can be seen by the growth in the number of users of data held in NSSDC as shown in Fig. 1. At the present time, we are processing over 1800 requests per year for secondary users of space science data. More important, this number has been increasing at a steady pace, and, according to all indications, will continue to climb.

To meet the current and anticipated needs of secondary data users, the Data Center must provide a wide range of services. One guiding principle is that if the data cannot be handled by a diversified spectrum of users with a minimum of

Table 1 Growth of the data base at NSSDC

Medium	August 1967	May 1970
Sheets and bound volumes, sheets	175,000	207,537
Digital magnetic tapes, ½ in. × 2400 ft	291	8,870
Microfilm, 100-ft rolls	7,800	15,410
Photographic films:		
9½-in. width, linear ft	14,000	18,000
70-mm width, linear ft	33,200	245,502
35-mm width, linear ft	0	759,680
4 × 5 in., each	2,100	5,471
8 × 10 in., each	0	2,410
16 × 20 in., each	93	93
20 × 24 in., each	2,200	8,005
Photographic prints:		
9½-in. width, linear ft	0	9,000
70-mm width, linear ft	0	22,000
8 × 10 in.	600	4,898
11 × 14 in.	200	500
16 × 20 in.	93	93
20 × 24 in.	2,200	3,200

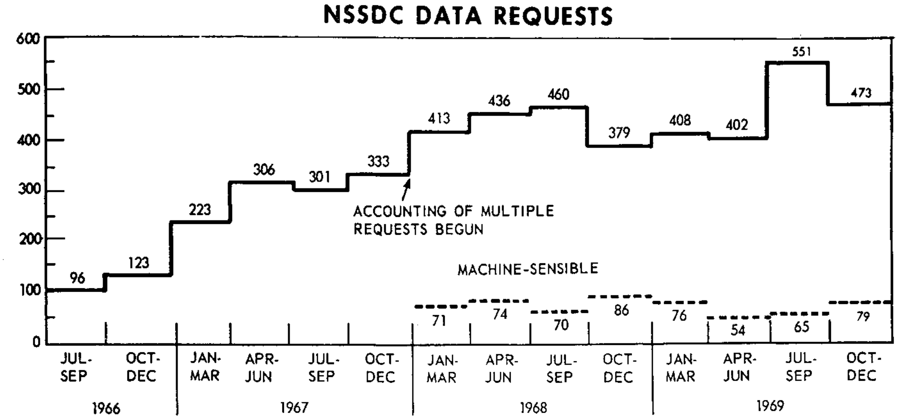


Fig. 1 Growth of requests: as of January 1968, requests requiring machine processing were identified apart from other requests, and single requests requiring different forms of data were treated as multiple requests.

effort, they should remain with the original investigators and be noted as available. To provide the necessary services, NSSDC must have the following capabilities: 1) a data processing system, 2) an information system about both the data comprising the data base as well as the availability of specialized data collections that exist in other locations, 3) 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, 4) a specialized technical library and automated document retrieval system, 5) a professional staff in the scientific disciplines that carries out analysis and synthesis of the data, and 6) a professional staff in the computer and information sciences that develops and upgrades information systems, analysis routines, and storage and retrieval techniques based on the latest technology. The complexity of the job to be done together with the huge volume of data that must be handled and processed required the adoption of a total systems approach and the automation of the Data Center.

Flow of Data and Information

It was within this framework that the Data Center planned and developed its current integrated information and data processing system. To oversimplify the mission of NSSDC, one must first arrange for obtaining the space science data and understand the form/format of incoming data. Once the data begin to arrive, there must be a central source of information concerning these data. This need is satisfied by the subsystem called Automated Internal Management (AIM). Upon arrival, one must process the machine-sensible data, prepare it for retrieval, and be able to handle special types of data in different forms and formats—this is done through the Machine-Oriented Data System (MODS). (The steps for processing nonmachine-sensible data are generally analogous.) In their work, the professionals at NSSDC must have ready access to the documentation relating to appropriate satellites, experiments, and data—the Technical Reference File (TRF) serves this purpose. Finally, statistics must be kept on the processing and use of data, and management must have a variety of reports in this area—this is greatly facilitated by the use of the Request Accounting Status and History (RASH) file. These subsystems are being tied together to form the General Automated Internal Management (GAIM) system and are supported by five additional special-purpose files: Computer Program Status Report, Magnetic Tape Unit Record, Computer Program File, Rocket File, and Distribution File.

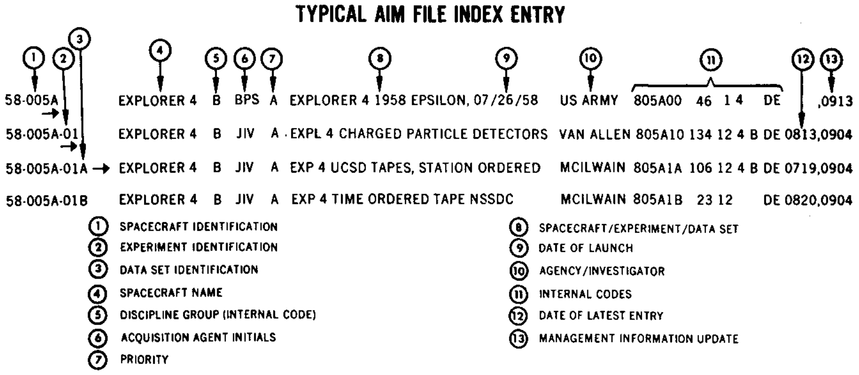
To obtain a better understanding of the NSSDC system and to get a broad picture of what happens during this process, it will be helpful to follow the path of information flow from the experimenter to the system. First, a space data scientist is assigned to each satellite/experiment/data set, as appropriate. He then obtains advance prelaunch information from such sources as the satellite project office, news releases, bulletins, reports, and personal contacts. This and subsequent information is entered into a working acquisition file, and, at this time, an AIM entry is generated. The agent establishes contact with the experimenter and his data processing personnel to arrange for bringing in data and related documentation. It should be pointed out that long periods of time are normally involved between this first stage and getting the actual data into the NSSDC system. This usually takes two or more years after launch.

Once the preliminaries are over, the acquisition scientist remains in constant contact, through visits or phone, with the experimenter and his data processing staff to help solve problems relating to the submission of data and documentation. Thereafter, the data and information come in almost on an automatic basis, except where special problems arise. The first items that should arrive at the Data Center are usually calibration curves, unpublished information, instrument descriptions, and data processing documentation. These are analyzed and selectively entered into the acquisition file and TRF, and notices are placed into the AIM file for subsequent use in processing incoming data and preparing publications.

Next, the reduced data, consisting mainly of magnetic tapes, arrive. At this time, the acquisition agent, together with a programer, as required, verify and analyze these data, prepare duplicates of the tapes, and prepare data set catalogs (indexing) using the MODS subsystem to accomplish these tasks. At this point, tape reformatting often has to be accomplished. The agent then feeds appropriate information into special-purpose files such as the tape and program status files. The AIM entry is brought up to date. (These subsystems will be discussed later.)

The analyzed data, normally made up of plots, graphs, and tables, arrive quite a bit later. Of course, for older experiments that are not yet in NSSDC, analyzed and reduced data may arrive in any order. The acquisition agent again goes through a similar processing cycle as in the case of machine-sensible data (data normally adaptable to computer processing). Data are verified, analyzed for information content, logged, indexed, and copied or microfilmed, and the information entered into the AIM information subsystem.

The working relationship with the experimenter is beneficial for the information transfer in other ways. Through this association and contacts at professional meetings, the acquisition agent receives copies of appropriate talks, reports, preprints, and published papers, as well as gaining a deeper understanding of the experiment and the implications of the data. (These items are supplemented by a thorough screening of the current literature.) Each of these documents is carefully analyzed, keyworded (by the acquisition agent), and entered into the TRF. Appropriate information is extracted for entry into the AIM.



TYPICAL AIM FILE INDEX ENTRY

```
58-005A         EXPLORER 4  B  BPS  A  EXPLORER 4 1958 EPSILON, 07/26/58   US ARMY    805A00  46  1 4    DE    ,0913
58-005A-01      EXPLORER 4  B  JIV  A ·EXPL 4 CHARGED PARTICLE DETECTORS  VAN ALLEN  805A10 134 12 4 B DE 0813,0904
58-005A-01A →   EXPLORER 4  B  JIV  A  EXP 4 UCSD TAPES, STATION ORDERED   MCILWAIN   805A1A 106 12 4 B DE 0719,0904
58-005A-01B     EXPLORER 4  B  JIV  A  EXP 4 TIME ORDERED TAPE NSSDC       MCILWAIN   805A1B  23 12     DE 0820,0904
```

(1) SPACECRAFT IDENTIFICATION
(2) EXPERIMENT IDENTIFICATION
(3) DATA SET IDENTIFICATION
(4) SPACECRAFT NAME
(5) DISCIPLINE GROUP (INTERNAL CODE)
(6) ACQUISITION AGENT INITIALS
(7) PRIORITY
(8) SPACECRAFT/EXPERIMENT/DATA SET
(9) DATE OF LAUNCH
(10) AGENCY/INVESTIGATOR
(11) INTERNAL CODES
(12) DATE OF LATEST ENTRY
(13) MANAGEMENT INFORMATION UPDATE

Fig. 2 Typical AIM file index entry.

It should be clear by now that the acquisition scientist spends a great amount of his time studying and working with the data to put all the necessary information together so that it may be useful to others. Using the information from the AIM and TRF subsystems and special-purpose files, the acquisition agent and publications staff prepare, as necessary, *Data Announcement Bulletins* and entries for the *Catalog of Satellite Experiments.* This does not necessarily mean that all data from a particular satellite experiment have arrived or have been completely processed. Many other contacts and correspondence with the experimenter may still be necessary. The preparation of a *Data Users' Note* concerning a particular experiment normally occurs after the final stage of data acquisition and processing. This document shows where the supporting information is, in what forms the data are available, what literature of previous work relating to the experiment is available, and offers a key to the use of the data.

The information flow is not complete without a mention of the RASH subsystem. The acquisition and request agents work through the RASH subsystem in satisfying users' requests on a daily basis. These requests may involve copies of data or publications, logical searches of the information files, or may even require further detailed data analysis on the part of the acquisition agent to help solve a particular problem.

Information and Data Processing System

Automated Internal Management (AIM)

AIM, as the centralized source of information, is the hub around which the other subsystems revolve. It is built upon detailed descriptions of the data, experiment, and spacecraft, along with the status of acquisition activity. The purposes served by the AIM subsystem are three-fold: 1) it performs logical searches to answer queries, 2) it indicates workload/volume of expected data, and 3) it provides action reminders.

Two by-products of the AIM file are the Acquisition Management Report (AMR) and the AIM File Index. The AMR provides information on spacecraft/experiments/data sets, responsible agent, priority, stage of acquisition, hours expended, and next contact. The AIM File Index provides a listing of the spacecraft, experiments, and data sets.

The contents of the AIM file are organized into a hierarchical structure. The most significant level is the spacecraft. Information which relates to the spacecraft is included here. The second level relates to the experiment. All experiment identification, detector descriptions, and commentary about a single experiment are contained in this section. The third level deals with a single data set.‡ These levels are generally tied together in the following manner, depending on identification of experiment and availability of data: the satellite-level entry will be followed by all experiment-level entries which pertain to that spacecraft; similarly, the data set-level entries are associated with the experiment. This concept can be perhaps better visualized by examining the typical AIM File Index entry shown in Fig. 2. Within each of the levels in a full AIM entry, there are specified categories of information concerning personnel, objectives, telemetry, instrumentation, acquisition information, experiment performance, data set contents, etc.

As noted earlier, AIM is also used for providing management information. Based on the same levels just discussed, detailed information is provided concerning spacecraft, experiment, and data set. The various categories of information are explained in Ref. 2.

A few examples of the current activity of the AIM file may be appropriate at this time. Orbital parameter data for 85 Russian satellites have been coded. The AIM file maintenance consists of about 1000 general changes per quarter, and this file is currently accounting for about 1200 satellites, 1600 experiments, and 350 data sets.

Machine-Oriented Data System (MODS)

To be responsive to the users who request data in digital form, as well as to those who provide the original data, NSSDC must have the flexibility to accept the data in any format and to provide it in any format. Since both the giver and taker will have restrictions imposed by their existing computer hardware and software, the Data Center facility must provide the "impedance matching." MODS is used for processing the data into the NSSDC computerized data base, for data set analysis, generation of data set catalogs, tape reformatting (when the interchange of information is inhibited by the diversity of hardware), and production of allied reports. Perhaps the best way to examine the composition of this subsystem is to follow incoming machine-sensible data sets through their processing cycle and then look at the tape reformatting process.

Processing Incoming Data Sets

All magnetic tapes received by NSSDC are first entered into the storage records by filling out the proper forms and assigning a unique accession number. At this point, an acquisition scientist, to be assisted by a programmer as necessary, is assigned to the data set for preliminary analysis.

The joint objectives of the acquisition scientist and programmer in the preliminary processing are 1) ability to read the entire tape in its logical format, 2) ability to list out any function or special record, 3) ability to detect errors (logical and physical), and 4) verification of the acquisition agent's understanding of the contents.

During this preliminary processing, the programmer writes all the necessary routines to manipulate the data and reformat it, if necessary. These programs are entered into the Com-

‡ Defined as a body of data reduced by one group of investigators in one specific way in a form, format, or organization that uniquely describes it. It can be a unit of machine-sensible or nonmachine-sensible data, which can contain one to several hundred magnetic tapes, rolls of film, etc.

puter Program File. The preliminary analysis stage is completed when NSSDC has the ability to use and interpret all data in the data set. This may require additional contacts with the experimenters.

At this time, the acquisition agent and the programmer define the format of the data set catalog, the functions of which are to provide an index to the contents of the data set, provide a series of error checks, calculate bounds or distributions of functions, provide a useful tool to the request agent for identifying data, and provide a coarse description of the information in the data set.

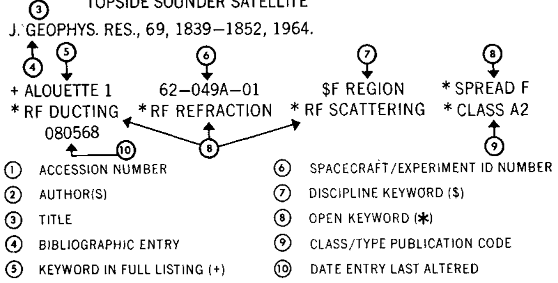
After the requirements are defined, the programmer writes a program to produce the data set catalog. This routine should also produce a copy of the original tape or a reformatted version. After the program has been checked and turned over to the computer people, the rest of the tapes in the data set are processed.

Tape Reformatting

The processing of normal machine-sensible data is well taken care of by using the procedures just outlined. However, experience has shown that people will not use data if it takes a lot of time and effort to convert it to a format which allows for direct entry into their own computer. The Data Center currently has routines available to read tapes generated by a number of operating systems,§ as well as BCD (binary coded decimal) tapes with arbitrary and variable record sizes, physically formatted binary tapes, and FORTRAN-generated tapes. For achieving compatibility with systems using 9-track tape, NSSDC uses other computers at GSFC. The hub of the MODS subsystem is a package called PIFT (Package for Information Formatting and Transformation) which will accomplish the functions just outlined and at the same time will produce densely packed machine- and media-independent data sets that may be accessed in the man-machine mode. PIFT, as a problem-oriented language, allows junior-level programmers to perform manipulations in a straightforward manner.

Technical Reference File (TRF)

The Data Center professionals must have internal documentation support and a tool for satisfying the bibliographic needs of space science data users. This is why the TRF comes into the system. It provides access to documents used for evaluating and verifying acquired data and for publishing catalogs, *Data Users' Notes*, and bulletins. It provides a useful record of the documentation available at NSSDC, as well as that which exists in the published literature. The references include published and unpublished documents relating to the spacecraft, experiment, or data set which are or will be preserved at the Data Center.



1
2
A00115 CALVERT, W. SCHMID, C. W.
SPREAD—F OBSERVATIONS BY THE ALOUETTE
TOPSIDE SOUNDER SATELLITE
3
J. GEOPHYS. RES., 69, 1839—1852, 1964.
4
5
6
7
8
+ ALOUETTE 1
62—049A—01
$F REGION
* SPREAD F
* RF DUCTING
* RF REFRACTION
* RF SCATTERING
* CLASS A2
080568
10
8
9
1 ACCESSION NUMBER
2 AUTHOR(S)
3 TITLE
4 BIBLIOGRAPHIC ENTRY
5 KEYWORD IN FULL LISTING (+)
6 SPACECRAFT/EXPERIMENT ID NUMBER
7 DISCIPLINE KEYWORD ($)
8 OPEN KEYWORD (*)
9 CLASS/TYPE PUBLICATION CODE
10 DATE ENTRY LAST ALTERED

Fig. 3 View of TRF entry.

The computer can display pertinent information in a variety of ways. Open (subjective) and controlled keywords are used to cover standard satellite/rocket identification, type and content of publication, and discipline-oriented keywords. A typical TRF entry is presented in Fig. 3. As is shown here, category 9 identifies the type and content of a publication. Here, the letter A indicates a journal article, and the number 2 designates a scientific paper containing experimental results.

To overcome the common gap between indexer-selected terms and user-selected terms, the acquisition agents themselves, who are space data scientists and the prime users of this subsystem, review the literature, select the entries, and keyword the inputs.¶ Thus, each member of the acquisition staff devotes a portion of his time to building up the TRF base and verifying the output. In this manner, up to 120 items per week are entered into the file, which now contains well over 4000 entries.

Considerable effort is presently being devoted to the production of a notebook-sized TRF output. Once this is fully implemented, NSSDC will have the capability of producing space science bibliographies ordered by author, discipline, experiment, or spacecraft. To produce special bibliographies upon request, a logical routine has been integrated into the TRF program which allows for the usual Boolean logic searches of AND, OR, and NOT among the entries. Present and additional keywords are also being studied to derive eventually a meaningful thesaurus.

Request Accounting Status and History (RASH)

At this point, the data from the space science experiments have been obtained, entered into the system, and prepared for retrieval. The next step is to facilitate the acquisition and request agents' contacts with the users of these data—this need is satisfied through the RASH subsystem.

This subsystem provides much valuable information. It is used to aid in keeping track of the progress of requests received by the Data Center and providing management with bookkeeping information. Specifically, RASH is designed to display up-to-date information relating to the number of requests, their status, estimated and actual costs, processing time, and necessary action reminders.** This variety of information can be retrieved by data set, requester, affiliation, date of request, date filled, request agent, status of request, and so forth.

Generalized Automated Internal Management (GAIM)

The various subsystems just described need to be tied together so that each file can be accessed. For example, a realistic requirement may call for information on both available data and documentation. GAIM will serve this purpose. The development of GAIM will occur in three phases. During phase 1, the various subsystems comprising GAIM (AIM, TRF, RASH, etc.) are being combined into a single homogeneous data base. The data base will be structured in such a manner as to be responsive in an interactive environment. A generalized command language to facilitate file maintenance, support searches, and generate special reports is being developed. The language must be suitable for interacting with acquisition scientists, data technicians, and other users of Data Center facilities.

§ Bell Operating Systems, Applied Physics Laboratory Operating Systems, and IBM-Direct Couple System.

¶ Document storage and retrieval is based upon randomly assigned accession numbers.

** This subsystem also can aid in the construction of a model of the NSSDC by supplying viable information describing the user community, types of requests and responses, and data sets most likely to be used.

Phase 2 will center about the development of a prototype interactive analysis system operated from remote terminals and supported by graphic devices. Included will be tools for general modeling, unit conversions, standard transformations, statistical packages, etc. Display techniques, both on-line and off-line, will be provided to assist the secondary data user in his research. Also, a natural command language will be developed which is responsive to the scientific user untrained in computer applications.[5]

The third phase associated with the development of GAIM is the definition and specification of the next generation system needed to support the mission of the Data Center. This includes both hardware and software. Consideration will be given to CRT devices, large mass storage devices (photographic, laser, electron), archival media (film images, inexpensive magnetic devices, etc.), computer-compatible microfilm equipment, computer-oriented publications, syntactic analyzers (to interpret natural written English inputs), procedures for analysis of space science data, and photo recognition techniques.

Interfaces between the Operational and Secondary Uses of Data

Up to now, we have described the secondary uses of space science data, the flow of data and information, and the information and data processing system at the National Space Science Data Center. No doubt many similarities between operational data processing centers and our activities have been noted. At this time, the need for both operational and secondary use of the same data in terms of the big picture should be re-emphasized.

First let us recap some of the distinctions and interfaces between these uses. A project is usually set up to generate data for a specific need. This could be an economic requirement in the area of weather, agriculture, R & D, etc. The key item is that such data are used as generated almost on a real-time basis to satisfy the identified requirements. This is fine, but the costs and time involved in obtaining, processing, and analyzing these data dictate that they should possibly be preserved for future use. And as we have tried to show, there are many valid secondary uses for such data.

In many instances, the major secondary users do not require the data per se, but, instead, need products 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. And, as mentioned, the users of such products may not be the scientists intimately involved in the particular discipline. More commonly, they include such groups as scientists in related disciplines, engineers and designers, planners, management, educational activities, recreational activities, commercial activities, and the general public.

It should be clear by now that there is a valid need to consider both uses of data. Figure 4 details the generalized flow of data. A few observations may be helpful at this time in following this flow. First, once a data generating program is approved, acquisition scientists backed by data processing specialists from the data center that will be acquiring the data for secondary use should start working with the generators during the time that data reduction plans are being formulated. There are a number of reasons for this involvement: 1) the primary data generators and users may not be aware of all secondary data uses; 2) data processing compatibility should be considered; 3) the proper data from each experiment/operation should be acquired; and 4) interface schedules should be established.

Following the data flow in Fig. 4, when the data arrive at the reduction or operational analysis point, they begin the cycle of serving related, but different purposes. When arrangements are made for sending the data to a data center, supporting documentation must also be collected to make these data useful for secondary users. It should also be noted that for secondary uses, there is no requirement for real-time links—delays are acceptable and normal. However, it is imperative that each center have some options in deferring or accepting past and currently available data. It should have the prerogative of determining what data are important in terms of known or potential secondary user requirements and on what they should expend their limited resources.

On the subject of resources, we do not believe that it is feasible for data centers to be totally self-sufficient in the same sense that research and development efforts are not. The agency responsible for the data-gathering program should provide funds for the experiment or operational programs to make the data and necessary documentation available to the center. Conversely, the agency responsible for the operation of the data center should fund for the internal operation and for its portion of the acquisition costs. Thus, it would appear to be appropriate for a fraction of the agencies' R & D and/or operational budget which supports the data-gathering programs to be used for supporting data center activities. In the case of space science data, about 1% of the funds expended to generate the data for primary use would be sufficient to perform the functions and services that we have been discussing.

Next Phase

We have tried to show the many different types of secondary data users, to tie together some of our common data pro-

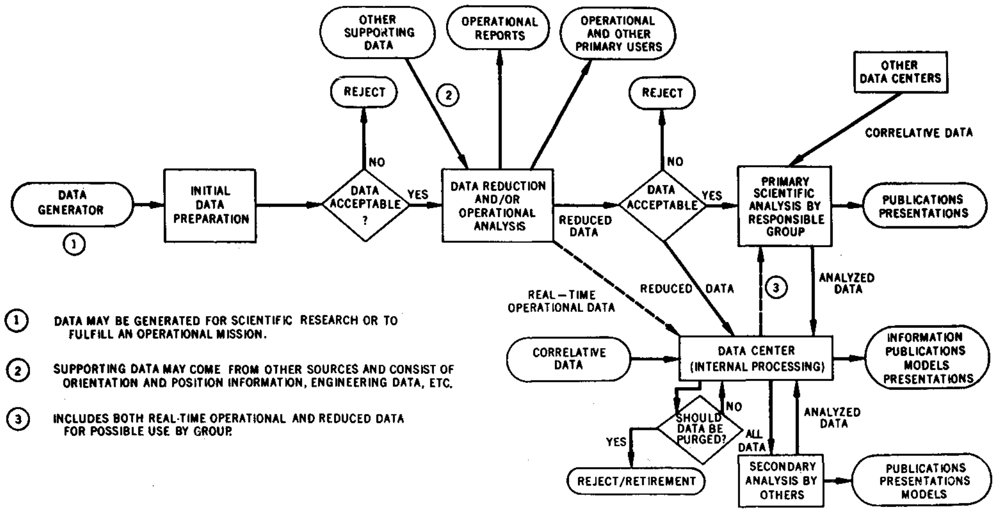

OTHER SUPPORTING DATA
OPERATIONAL REPORTS
OPERATIONAL AND OTHER PRIMARY USERS
OTHER DATA CENTERS
REJECT
2
REJECT
CORRELATIVE DATA
NO
NO
DATA GENERATOR
1
INITIAL DATA PREPARATION
DATA ACCEPTABLE ?
YES
DATA REDUCTION AND/OR OPERATIONAL ANALYSIS
REDUCED DATA
DATA ACCEPTABLE
YES
PRIMARY SCIENTIFIC ANALYSIS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

Fig. 4 Generalized data flow.

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.

cessing and information problems, and to give some insight into the system that we have developed and are continuing to develop. More important, some of our present data management problems based on our experiences with secondary data users point quite vividly to the direction we will be going in the future. In hopes of furthering the solution of some of our common problems, we would like to state some of our views based on these experiences with secondary data users.

As the Data Center grows, so must its information system. It must be able to handle large varieties and amounts of data and prepare them for a multitude of different uses. It must rely on new and better software and hardware to effectively perform these tasks, bearing in mind the guiding principle of furthering the effective use of data from space science experiments. The present system software must be upgraded with respect to processing incoming tapes for verification of inputs and quality control—two goals are immediate detection of errors or omissions and standard maintenance and system quality control programs. Effective purging of the active data base will have to be accomplished. Consideration must be given to a good long-term archival medium as the lifetime of magnetic tape cannot compare with photographic or printed matter, although recent tests are more encouraging. Time-phased data compression will be another vital area of concern. Considerations include higher density storage techniques and the actual compression of data. This data compression can occur in various steps. The first step would involve the retirement of alternate forms of the data in which the most useful form would be retained. Then, even this most useful form of data could be subjected to the removal of derived variables, which are computed from basic positional and attitude information. This would still permit recalculation of these variables at a later date, should this prove necessary. At this point, no reduction in the basic information content has occurred. However, if one wishes to use this data, more time and resources will have to be utilized than previously. One is balancing this cost against the maintenance cost of keeping all the derived bits in the active data base. There is a break-even point depending on data usage. As one starts the irreversible process of destroying information content, a sensible approach would be to separate the background information (ambient, quiet time) from the event information (disturbed time). This will permit time averaging the background information, say over hours or days, for subsequent use in determining long-term changes. A sizable reduction in the number of bits for a given data set will occur in this process, and yet the information most likely to be used in future studies will still be available. Clearly, data of historical significance would be preserved as long as possible.

In other words, certain points should be considered when planning for the retirement of data: 1) large volume plus cost of maintenance plus fixed resources dictate the orderly retirement of data; 2) early in its life, various forms of the data are useful, e.g., time-ordered, space-ordered, etc.; 3) data can be reduced without losing information content by eliminating certain forms of data, by removing derived variables, and by keeping only the significant number of bits, not the full computer words; and 4) information content of data can be reduced by breaking out event data from background data, by averaging the background over suitable time intervals, by preserving only special data for historical purposes, by preserving only outstanding geophysical event data, and by compressing data into analyzed forms so that general understanding of phenomena is retained. In short, data can shrink in size and in information content, but knowledge of it never disappears from the scene.

Some thought, as we have previously mentioned, is already given to the next generation of the NSSDC information system. One consideration is to provide the Data Center with much greater flexibility and capability by developing varied analysis programs which can be readily applied to the data. Although complete requirements have not yet been defined, it is envisioned that scientists, experimenters, and acquisition agents should be able to interact on-line, through a computer, to data bases and data sets held at NSSDC. It is also anticipated that in this way the resulting dialogue between two or more scientists can be used to synthesize new information in the process.

These concepts are not too far from reality. With the progression of time, the central processing facilities are performing more work on the raw data before it is sent to the experimenters for analysis. In the beginning, the raw data was sent directly. Now, tapes are digitized and edited, noise flags are inserted, time overlaps are removed, and decommutation is performed. There is interest at present in having the orbit and attitude information merged with the data before it is sent to the experimenter. As high-speed data links become available across the country, there will be no need to send the data to the experimenter. Instead, standard processing will be performed up to the point where detailed analysis can begin. The data in this form could be sent directly into the Data Center where it could be reached via high-speed terminals and manipulated on large computers by the principal investigators using many standardized analysis programs. Special-purpose analysis programs would be constructed on-line by the individual users as the needs arise. At that point, the processing facility and the Data Center will have blended into one operation. There exists today an on-line retrieval system with a 10^{12} bit capacity.[6] This is capable of handling a year's worth of space science data at the present rate of generation. Within 5 years 10^{15} bit systems seem to be feasible. This technology advancement will permit the development of a truly interactive system with the whole space science data base plus correlative ground-based measurements. It is quite clear that this new type of facility could be a reality in 5 to 10 years.

In closing, we would like to say that 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 and solution of problem areas, the reduction of unnecessary overlap, and the development and spread of technological advances in data storage, manipulation, and retrieval.

References

[1] "The Operation of the National Space Science Data Center," NASA/NSSDC 67-41, Oct. 1967, NASA.

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

[3] Fava, J. A., "A Framework for Future Data Centers," *Proceedings of the American Society for Information Sciences*, 1969, pp. 417–421.

[4] Dessler, A. J., "The Role of Space Science in Graduate Education," *Transactions, American Geophysical Union*, Vol. 49, No. 3, Sept. 1968, pp. 519–554.

[5] Shapiro, A., "Requirements for the National Space Science Data Center Information System," NASA/NSSDC 69-04, March 1969, NASA.

[6] Kuehler, J. D. and Kerby, H. R., "A Photo-Digital Mass Storage System," *Proceedings—Fall Joint Computer Conference*, Spartan Books, 1966, pp. 735–742.