X-601-69-414

PREPRINT

NASA TM X-63700

APPROACHES TO SEARCHING AND RETRIEVAL AT A DATA ANALYSIS CENTER*

BRUCE I. BLUM

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. This work was performed under contract NAS 5-8060 to NASA by Wolf Research and Development Corp.

X-601-69-414

APPROACHES TO SEARCHING AND RETRIEVAL AT A DATA ANALYSIS CENTER

Bruce I. Blum
National Space Science Data Center

September 1969

GODDARD SPACE FLIGHT CENTER
Greenbelt, Maryland

ABSTRACT

Because of the wide scope of activities carried on at a data analysis center, there will frequently be a need for a variety of information systems using different data bases, file organizations and search strategies. In this paper, work done at the National Space Science Data Center is used to illustrate this situation. Two major information systems are discussed together with the factors which led to their design; other retrieval systems used at NSSDC are also briefly mentioned.

CONTENTS

ILLUSTRATIONS

APPROACHES TO SEARCHING AND RETRIEVAL AT A DATA ANALYSIS CENTER*

INTRODUCTION

The techniques employed in the retrieval of data and information are a function of the form of the material, the size of the data base, the use to which the response will be put, and the ability to describe the data. An approach which is satisfactory for one application may be totally irrelevant in another situation. Consequently, an organization may be forced into adopting several searching techniques, each tailored to a specific problem. In this paper we shall discuss the various approaches used at the National Space Science Data Center. We use NSSDC because it illustrates a new class of information retrieval problem and - at the same time - provides an example of a generalized retrieval system which may be modified to meet many specific requirements.

The NSSDC has as its data base the processed results of space science experiments which were conducted on-board satellites. These data sets (as they are called) are the outputs of the experimenter's reduction and analysis. The result is an extremely heterogeneous data base which tends to be quite large. It is the NSSDC mission to collect, announce, distribute, and provide facilities for the analysis of these data. In the discussion which follows we shall identify several of the retrieval problems which confront the NSSDC and outline the solutions which were adopted. It should be noted that while all the systems discussed in this paper are currently operational, much of what is presented represents the first phase in the development of an integrated system for the NSSDC.

What Data Are Available?

There are approximately 1200 different satellite experiments which have collected or are collecting data of interest to the space science community. Of these, approximately 200 experiments have one or more data sets deposited at NSSDC. The present anticipated acquisition rate is 50-100 data sets per year.

One class of retrieval requirement imposed upon NSSDC, therefore, is the ability to define which data are currently available, which are in the process of being acquired (and when), and which experiments will not produce data that may be acquired. An information system to answer these questions is required in order to service requests for information and data. The system is also essential

*This work was performed under contract NAS 5-8060 to NASA.

to the NSSDC management and staff charged with the responsibility for identifying and acquiring data.

The specifications for an information system to satisfy this requirement is a function of:

- The kind of information to be contained in the data base.
- The number of different people who need to access the data.
- The depth of query capability desired.
- The other uses to which the data base will be put.

For this system, it was decided that the data base should include all information which might be useful in the description of a satellite, experiment and the resultant data. This would include fixed elements such as dates, funding agents, and phenomena observed, as well as free text descriptions of the objectives, design, performance, and results of the investigations. The data base was also designed to produce data products and query responses for the NSSDC staff - including management, scientific, and non-professional personnel - and any visitors to the NSSDC. The data base specifications required that the information retrieval system must be able to query any item in each entry - both fixed items and descriptive information in free text. Finally, it was decided that the data base would also be used to provide announcement material, support a management information system, and maintain records relating to internal NSSDC processing.

Once the data base requirements were established, NSSDC set out to produce an information system to operate upon it. The resulting Automated Internal Management System (AIM) is now fully operational. The data base upon which it operates, however, has not yet been fully implemented. In the next section we shall describe the organization of the data base and the strategies which have been used to search for and retrieve information. Because some of the categories of information are not yet stored in any quantity, we have not been able to evaluate the effectiveness of this approach to information retrieval.

The AIM Data Base

The AIM data base is broken into three hierarchical levels - satellites, experiments, and data sets. Figure 1 illustrates the relationships among the elements in this structure. Each entry (i.e., set of information describing a single satellite, experiment, or data set) is subdivided into categories of information. A category may be a fixed format (i.e., given character positions always

SATELLITE	EXPERIMENT	DATA SET
EXPLORER 18	RETARDING POTENTIAL ANALYZER	PLOTS OF ENERGY VS VOLTS
	FLUXGATE MAGNETOMETER	5 MIN AVG OF MAGNETIC FIELD
	COSMIC-RAY PROTONS	RATES + PH REDUCED DATA
	COSMIC-RAY	GM HOURLY RATES
	ENERGETIC PARTICLE	GM+ION CHAMBER DATA
		GM+ION CHAMBER (CHRONOLOGICAL)
		GM+ION CHAMBER (GRAPHS)
	ELECTROSTATIC ANALYZER	PLOTS OF FLUX VS TIME
	FARADAY CUP	3 HR AVG OF PLASMA PARAMETER
		IRREGULAR SAMPLE OF PLASMA
		PLOTS OF PLASMA CURRENT

Figure 1. Structure of AIM File as Illustrated by the Explorer 18 Satellite

contain the same kind of information), a free text structure, or a combination of the two. Figure 2 presents the presently defined categories for a satellite entry. Experiment and data set entries are similarly structured.

Category	Contents
1	Name and alternate names
2	Personnel, e.g., Principal Investigator, etc.
4	Fixed information, e.g., orbit, funding agent, etc.
5	Brief description (free text)
6	Objectives (free text)
7	Description (free text)
8	Performance (free text)
A	Action reminders (free text)
B	Remarks (free text)
D	Acquisition information (fixed format)

Figure 2. Categories Used in AIM Spacecraft Level Entry

The number of categories for any level is optional and easily changed. Because the creation of accurate and complete entries requires considerable effort, the system has been designed to operate efficiently upon a data base containing only partial entries. Thus, each entry may contain any or all categories of information. A complete entry is normally around 1000 words long.

The fixed format information is carried as a character string. The format of each category is encoded in the program as a set of tables and routines. Hence, if one knows the external name of an item, e.g., LAUNCH for the date of launch, then it is a simple process to identify the category and character positions which will contain that information.

Retrieval of information in free text form is not as direct. For example, consider the Satellite objectives category (SOBJ) for the Explorer 18 satellite listed in Figure 3. The information presented here has been structured for readability and announcement. Some of the information in this paragraph can also be identified as useful for retrieval purposes. For example, someone interested in the "radiation environment of cislunar space" would certainly want to retrieve information about Explorer 18. In the present operational system, each of the following inputs would be able to recognize this string:

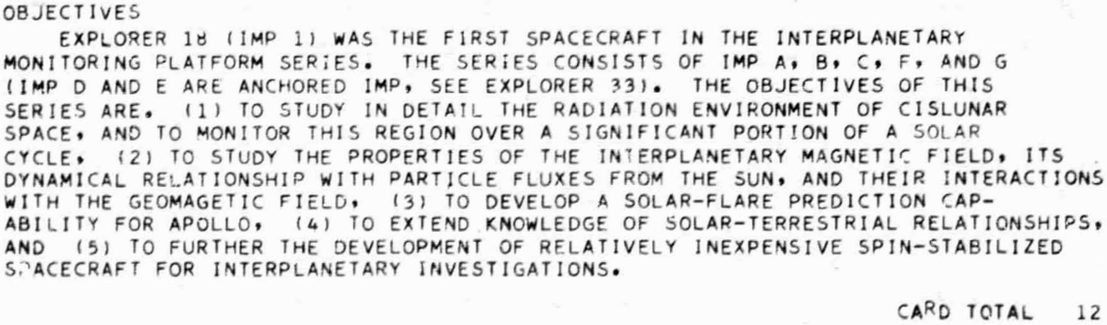
(a) SOBJ RADIATION ENVIRONMENT OF CISLUNAR SPACE

(b) SOBJ RADIATION ENVIRONMENT
SOBJ CISLUNAR SPACE

(c) SOBJ CISLUNAR

It is obvious that (a) requires an exact match in the satellite objectives category. Statement (b) is slightly more general at the loss of some precision, and (c) is the most general and the least precise.

There are, of course, many techniques which could be used to link the terms together to indicate their syntactic roles. In the context of the current NSSDC



```
OBJECTIVES
    EXPLORER 18 (IMP 1) WAS THE FIRST SPACECRAFT IN THE INTERPLANETARY
MONITORING PLATFORM SERIES.  THE SERIES CONSISTS OF IMP A, B, C, F, AND G
(IMP D AND E ARE ANCHORED IMP, SEE EXPLORER 33).  THE OBJECTIVES OF THIS
SERIES ARE.  (1) TO STUDY IN DETAIL THE RADIATION ENVIRONMENT OF CISLUNAR
SPACE, AND TO MONITOR THIS REGION OVER A SIGNIFICANT PORTION OF A SOLAR
CYCLE,  (2) TO STUDY THE PROPERTIES OF THE INTERPLANETARY MAGNETIC FIELD, ITS
DYNAMICAL RELATIONSHIP WITH PARTICLE FLUXES FROM THE SUN, AND THEIR INTERACTIONS
WITH THE GEOMAGETIC FIELD,  (3) TO DEVELOP A SOLAR-FLARE PREDICTION CAP-
ABILITY FOR APOLLO,  (4) TO EXTEND KNOWLEDGE OF SOLAR-TERRESTRIAL RELATIONSHIPS,
AND  (5) TO FURTHER THE DEVELOPMENT OF RELATIVELY INEXPENSIVE SPIN-STABILIZED
SPACECRAFT FOR INTERPLANETARY INVESTIGATIONS.

                                                                 CARD TOTAL   12
```

Figure 3. Sample OBJECTIVES Entry (Satellite Level)

operation, however, a more extensive system would not seem justifiable. Nevertheless, we have implemented several features to speed search time and lessen this requirement for an exact match.

In the example given, we have selected keywords embedded in the descriptive text. Yet, only a relatively small number of words are applicable for searching. There are two approaches one may use to identify these keywords: they may be isolated and repeated in an index or they may be flagged in the text stream. We have chosen the second technique for the following reasons:

- Our vocabulary is open and unstructured.
- Our file is linear.
- Keeping the keywords in the text facilitates the listing of hits with the keywords used in context.

In identifying keywords, we have recognized the problem of multiple word terms. For example, consider INTERPLANETARY MAGNETIC FIELD. Ideally, we would like to consider both that set of words and MAGNETIC FIELD as keywords. To do this we have defined three special flagging characters. One ($) is used to indicate the start of a keyword item. The second (+) indicates that the following word is to be associated with the previous string, and the third (=) has the effect of both the $ and +. Hence, we would write

```
$INTERPLANETARY=MAGNETIC+FIELD
```

and this would define both desired keyword terms. In a normal listing, the special flags are printed as blanks. When desired, the system can also search each word in a category and ignore the keyword flags. However, the use of these keywords speeds scanning time and facilitates the construction of a concordance.

A second modification was implemented to lessen the exact term match restriction. For example, one would not want to lose the reference to PARTICLE FLUXES simply because PARTICLE FLUX was requested. To bypass this problem, we use an asterisk to indicate that no matching should be done beyond this character. Hence, the command

```
SOBJ        PARTICLE FLUX*
```

would accept either item.

Through the use of these techniques, we can rapidly identify any fixed format item and any keyword in a free text stream. Each term has an external name such as LAUNCH (date of launch), SOBJ (satellite level, objectives category keyword), etc. The AIM system uses a command language which links an external identifier with a relationship and a string of terms which must satisfy that relationship with the elements in the data base. For example, a request for all satellites which were launched in the last six months of 1968 and which were in some way associated with the ionosphere might be written:

```
LAUNCH        .BT.,070168, 123168

SOBJ          IONOS*
```

In this example, the between relationship is written .BT. and the match (or equal) relationship is implied. Other relationships are "less than," "greater than," and "not." Individual commands may be grouped together or combined with an "and," "or," or "not" relation. In this way, the AIM system allows a full logical search utilizing all meaningful elements in its data base.

What Documentation Is Available?

Through use of the AIM system, NSSDC is able to extract information and produce a variety of reports concerning satellites, experiments, and data available at NSSDC. In addition to this system, NSSDC requires an information retrieval system for the documents which are required to acquire, announce, and analyze the data stored at NSSDC. To satisfy this need, a Technical Reference File (TRF) was established. To provide the necessary document accounting, bibliography preparations, and query capability, the AIM system was modified to operate upon a data base containing TRF citations and index terms. The resultant system is called the TRF system.

Although the TRF system inherently contains all the features of the AIM system, we have adopted different techniques based upon the contents of the TRF data base. For the NSSDC's purposes, the TRF must respond to the following kinds of queries:

- What documents relate to a given satellite, experiment or data set?
- What documents relate to a specific space science discipline?
- What documents were authored by a given investigator?

In addition to satisfying such queries, it is important that the NSSDC be able to identify documentation by source, e.g., journal article, university report, etc., and by content, e.g., scientific results, instrument description, calibration document, etc.

To satisfy these demands, the TRF file was setup as a single level file with each entry containing the following categories:

A	Author(s)
T	Title
B	Bibliographic citation
K	Keywords

The first three categories are free text fields, with the author category using a double blank between authors for a co-authored document. The system is capable of searching on an author's name; for the NSSDC application, there is no current need to search on either the title or bibliographic citation. Nevertheless, the latter is structured to facilitate searching should it be required.

The basic search item is the keyword. Keywords are of a fixed length (23 characters) and are preceded by an identifying character. A blank is used to indicate that the keyword is the NSSDC ID for a satellite, experiment, or data set. The dollar sign denotes a discipline keyword, and an asterisk flags a keyword which is part of an uncontrolled vocabulary. This last class of keyword is used to facilitate access to special documents. At the same time, it is hoped that the use of this kind of keyword will provide experience which may culminate in the creation of a controlled vocabulary or space science thesaurus.

Finally, there is the "*CLASS" keyword which is used to identify the source and contents of a document. Figure 4 lists the codes which are used with this keyword; for example, *CLASSD3 would indicate a university report describing an instrument.

The search routine for the TRF program is essentially the same as that in AIM. Search may be on AUTHOR, DISCIPLINE (i.e., $ keywords), ID (i.e., keywords with a leading blank), KEY (i.e., *keywords), and CLASS (i.e., *CLASS keywords). A search for all theoretical or scientific papers by L. R. Davis or N. F. Ness relating to the fluxgate magnetometer experiment on Explorer 18 (NSSDC ID 63-046A-02) could be written as follows:

AUTHOR	DAVIS, L*bbNESS, N*
ID	63-046A-02
CLASS	1, 2

PUBLICATION CODE	CONTENT CODE
A JOURNAL ARTICLES	0 BIBLIOGRAPHIES
B BOOKS	1 THEORETICAL PAPERS
C GOVERNMENT PUBLICATIONS	2 SCIENTIFIC PAPERS - EXPERIMENTAL RESULTS
D UNIVERSITY REPORTS	3 INSTRUMENT DESCRIPTION PAPERS
E INDUSTRY REPORTS	4 DISCIPLINARY REVIEW PAPERS
F MAGAZINES, PRESS RELEASES AND NEWSPAPER ARTICLES	5 SATELLITE & MISSION DESCRIPTION
G PROCEEDINGS, SYMPOSIA AND OTHER COLLECTIONS	6 NEWS RELEASES
H UNPUBLISHED	7 DATA PROCESSING PAPERS
	8 WORKING PAPERS, MINUTES, ETC.
	9 DATA TABULATION

Figure 4. Codes Used with the *CLASS Keyword

The form used in the AUTHOR command has a special format because the normal command delimiter - the comma - is generally part of the search string. In this case the double blank (bb) is used to indicate the end of a single search item. Since DAVIS, L. R., and DAVIS, LEO R. are standard entries for the sample person, the asterisk is used to make either form acceptable.

Other Extensions of AIM

In addition to providing information on satellite experiments and documentation, the AIM program has also been modified to:

- Maintain supporting and descriptive information relating to photographs of the lunar surface taken from satellite experiments, e.g., Lunar Orbiter, Apollo 8, etc.

- Maintain an address file of all NSSDC users. This file is currently used for all standard distributions; it will be extended in the near future to support an SDI capability.

- Maintain a production control, management-information system for NSSDC request activities.

- Maintain separate systems for both rocket data and information about correlative ground based data.

In each of these cases the organization and contents of the data base is quite different. Consequently, the retrieval requirements and search strategies also vary. Nevertheless, each system allows a full search capability on all definable items in an entry. The items actually searched upon vary according to the structure, content, and function of the file.

Each of the seven systems mentioned above is operational in an IBM 7094. Each is a simple extension of a simple basic program. The AIM program developed at NSSDC is not unique. There are a variety of file management systems currently in operation. Many are manufacturer supplied, others may be rented, still others are distributed by the originating organization. The available software does not differ markedly. There are few features of the NSSDC system which are unique. Nevertheless, the real challenge which the NSSDC faced was not how to develop programs, but how to adopt a system which would be responsive to its needs. And this involved the creation of seven major files with associated retrieval keys and search strategies.

A LOOK TO THE FUTURE

At the present time, the systems described above operate upon linear files in the batch mode. The NSSDC is now beginning to convert these systems to a third generation computer which will support direct accessing of data and user interaction with the data base, i.e., an on-line system. As this is implemented, certain of the off-line features will be revised to allow for conversational retrieval. The files will also be structured to allow referencing across files, and a single integrated system - the Generalized AIM System (GAIM) - will be used to maintain, search, and produce reports from the data base.

The systems discussed in this paper are all extensions of the same basic program. Yet each data base differs as to function, structure, size, and use; and the approach toward the retrieval of elements in the data base has varied accordingly. There are, however, other retrieval problems which cannot be solved by use of the GAIM system. For example, there is the task of retrieving specific information in a data set contained on several hundred magnetic tapes - a problem NSSDC is solving in a very different way. And, this simply reinforces the central statement of this paper: that a modern data analysis center must be equipped with a variety of search and retrieval capabilities with no single system capable of satisfying all of its requirements.