SECTION II

DATA BASE

Kenneth H. Cook and Lynn Trump

This section describes the development, organization, and growth of three separate data bases used by SUPARS/DPS II. In order to give a broad and general overview of how data flows through the system, Appendix A presents a description of the utilization of data as it is channeled through the various programs.

The three data bases that were developed were: (1) Document Data Base: includes bibliographic citations and abstracts (documents) from Psychological Abstracts (January 1969 - June 1971), (2) Vocabulary Data Base: all unique terms identified through the free-text processing of documents and stored in the DPS inverted file, and (3) Search Data Base: search inquiries originally entered by users to search the document data base, including the search words and Boolean operators.

A data base of documents was developed during the previous year's experimental work in 1969 - 71; the vocabulary and search data bases were developed as new searching algorithms for the 1970 - 71 research.

1. DOCUMENT DATA BASE

The document data base was developed from machine readable types of Psychological Abstracts rented from the American Psychological Association. The general rules for exclusion of common words, special character handling, sentence and paragraph endings were the same as reported in the July, 1971 report. (2)

Because of formating changes and character set changes made to the 1971 PA tapes by the American Psychological Association, modifications had to be made to the SUPARS/DPS programs in order to process data. First, program TRANSLATE had to be modified in order to deal with the standard IBM scientific character train using upper case characters rather than the previously used special multiple alphabet set of characters. Second, program REFORMAT had to be modified to accomodate (a) changes of document "fields" by APA and (b) the new use of new, fixed length fields to be used as pointers to the actual variable length fields of data, rather than the previous use of all variable length fields.

A description of the general logic and a flowchart of the TRANSLATE and REFORMAT programs is given in Appendix B. A chart of the fourteen fields, such as author, title, source, abstract, etc. used to organize the data for each document are also given in Appendix B. Figure 1 of Appendix B lists the fields as found on the original APA tapes before translation; Figure 4 of Appendix B, gives the translated fields that have been reformatted and

are ready to be processed (loaded) by the SUPARS/DPS programs.

After proceeding through a VALIDATE program, which checked each document for maximum length allowable, and deleted those which were too long, documents were ready for processing, or loading. Documents from January 1969 through November 1970 were processed through the two-phase DPS program that developed an inverted file of free-text words and documents.

The loading history of the SUPARS/DPS II document data base is shown in Table I. A total of 46,828 documents were loaded into the data base. The number of unique words and terms derived from free-text indexing of these 46,828 documents amounted to 106,702. Documents were loaded into 6 separate batches as shown in the six rows of Table I, and included the months of January 1969 - June 1971.

A graphic representation of the growth of the data base is shown in Figure 6, "Growth Rate of SUPARS/DPS Data Base." The six numbered points on the chart represent a separate batch of documents that were loaded, and correspond to the six batches of documents shown in Table I.

Another indication of the size of the document data base is the number of tracks used on the main storage devices, the 2314 and the 3330 disk pack. The track usage is shown in Table II. Three separate files are maintained in SUPARS/DPS: (a) Dictionary, containing each unique word and the document postings, (b) Vocabulary, containing unique words, and all document numbers which contain the word, and (c) Master, containing a coded representation of the entire document. For each file, comparisons are given of the number of tracks allocated and the number of tracks used on the 2314 disk pack (used with the 360 operating system) and the 3330 disk pack (used with the 370 operating system.)

During two separate periods of time during the growth of the data base, the vocabulary file had to be reduced in size, or restructured, in order to fit into the available storage room on the 2314 disk packs. A restructure program allows for a more compact and efficient storage of the large strings of document numbers listed after each unique word in the vocabulary.

The first restructure program was accomplished while the 360/50 operating system and the 2314 disk packs were in use and was done on the data base of January 1969 - April 1970. As a result of this restructure, the amount of storage space needed on the disk packs was reduced by 45% in the vocabulary. The smaller document number strings contributed to both an increased search efficiency and a greater amount of available work space.

A second restructure was necessary when the 370/155 operating system was installed in December, 1971 and the 360/50 system was removed. The different relative track addresses on the new 3330 disk packs from the old 2314 packs necessitated a restructure of data for the vocabulary, dictionary, and master files. The reduction in size of the three files from the 2314 to the 3330 in terms of the number of tracks used is given in Table II. For

TABLE I

LOADING HISTORY OF SUPARS/DPS II DOCUMENT DATA BASE

1 Batch	2 Dates	3 Documents Loaded	4 Total Documents Loaded	5 New Words Added	6 Total New Words Added	7 (5/3) Avg. No. New Words Per Document
#1	Jan-June, 1969	8,859	8,859	34,239	34,239	3.86
#2	July-Sept, 1969	4,607	13,467	8,621	42,860	1.87
#3	Oct-Dec, 1969	4,601	18,068	6,853	49,713	1.49
#4	Jan-Apr., 1970	5,805	23,873	27,913	77,626	4.80
#5	May-Sept, 1970	9,920	33,793	11,614	89,240	1.17
#6	Oct '70-June '71 (*)	13,035	46,828	17,462	106,702	1.34
	TOTALS	---	46,828	---	106,702	

(*) Does not include March or May, 1971 (omitted from original tape version.)

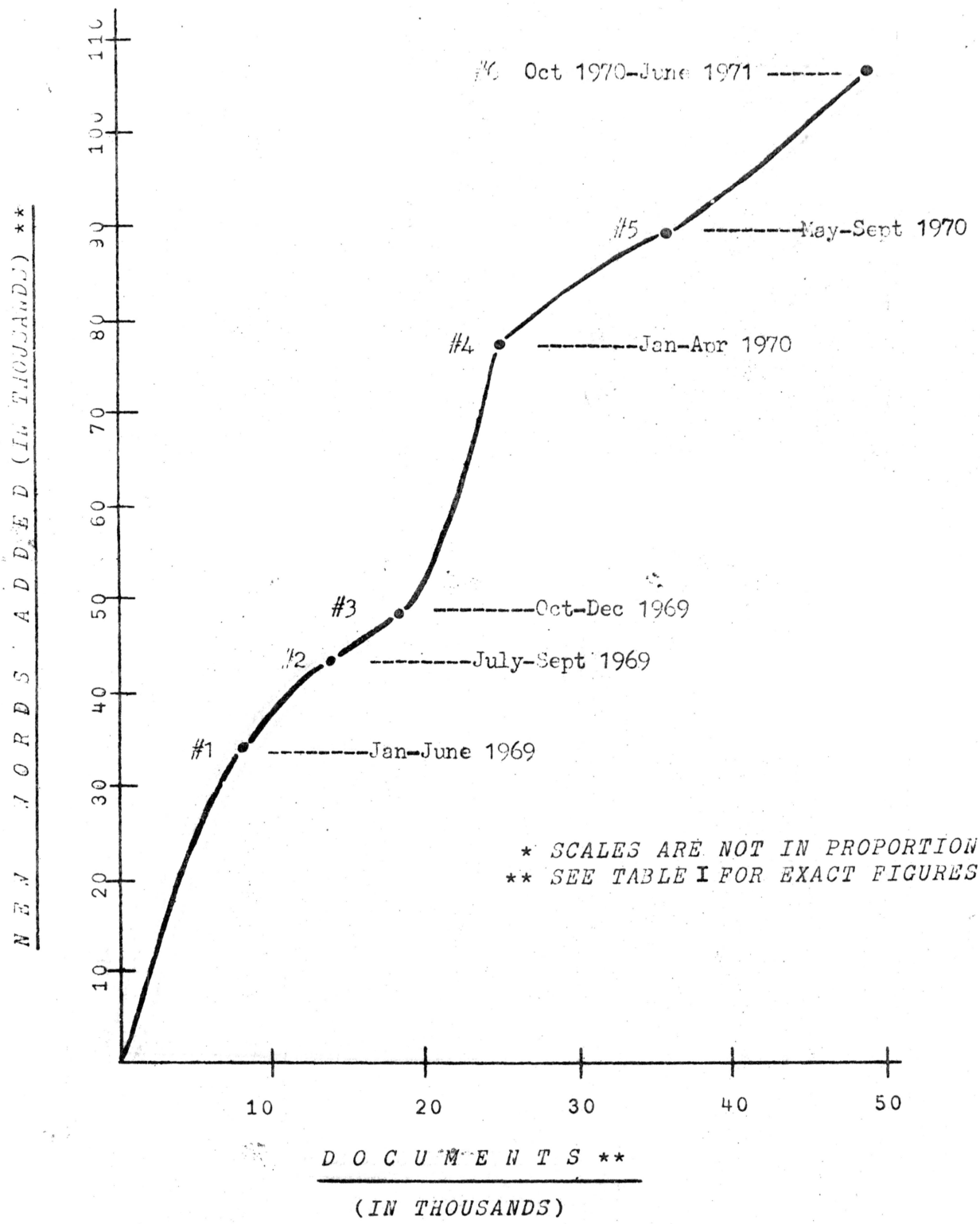


Figure 6. Growth Rate of SUPARS/DPS II Document Data Base*

TABLE II

TRACK ALLOCATION AND USAGE OF DOCUMENT DATA BASE

Data Set Name	Document Range	Total Number of Items Stored	Number of Tracks Allocated		Number of Tracks Used	
PSYCHOLOGICAL ABSTRACTS			2314	3330	2314	3330
DICTIONARY	Jan/69-Apr/70[1]	59,415 words	600	200	286	156
	Jan/69-June/71	106,702 words	---	300	---	235
VOCABULARY	Jan/69-Apr/70	59,415 words[2]	3000	1350	1684	1083
	Jan/69-Jun/71	106,702 words	---	2650	---	2434
MASTER	Jan/69-Apr/70	23,873 documents	3980	3000	2866	1581
	Jan/69-Jun/71	46,828 documents	---	3200	---	3110

[1]Until data files were moved to 3330 disks, users were restricted to a data base of documents from January, 1969 through April, 1970.

[2]Dictionary and vocabulary files each contain 1 entry for each unique word. However, the vocabulary occupies more space as each record is a list of all documents in which the word appears.

example, in the dictionary, for the documents January 1969 - April 1970, the number of tracks used on the 2314 was 286, while the number of tracks used on the 3330 disk pack was only 156 after the restructure.

By the time the total restructure was completed, only 1 1/2 3330 disk packs as compared with four 2314 disk packs were used to store the document data base, the search monitor, and DPS search modules, and the search data base (explained below in sub-section 3, "The Search Data Base").

2. VOCABULARY DATA BASE

A second major data base that was made available on-line to users of SUPARS/DPS was the entire free-text vocabulary. The vocabulary contains all the unique words taken from the free-text processing of documents and stored in the inverted file. The vocabulary, called the "dictionary" in the original DPS programs, is actually a by-product of the free-text processing.

The total size of the vocabulary reached 106,702 words by the time documents from January 1969 - June 1971 were processed. Figure 7, "Cumulative Growth Rate of the SUPARS/DPS Vocabulary Data Base" shows the cumulative growth of the Vocabulary data base for each batch of documents processed.

In order to make use of the words contained in the vocabulary and have them accessible in an on-line mode for users to query, special search programs had to be developed. The term "delta V" is used to describe these types of searches, because the character "delta" and "V" are typed by the on-line user to access the vocabulary data base.

a. Special Programs Developed for a Vocabulary Data Base

Two new search modules had to be developed to handle the Vocabulary Data Base. The programs were necessary in order to access the vocabulary as a data base rather than have it operate in its normal capacity as a special handling procedure for the DPS Dictionary file.

The vocabulary control program is a combination, with a great deal of modification, of the existing programs of the standard DPS search routine which is called by the search control program for all Delta V searches. The vocabulary control program reads user input line-by-line, storing the first user-entered search word. If the user has not established a limit to the number of words to be printed as output, the program sets it to 100. That is, a maximum of 100 items (words) will be allowed to be outputed. The program then calls the interface program, which is a modification of the DPS program which actually locates the words in the file of all free-text words.

Two output options are currently available. The first simply calls for a frequency count of documents in which the user-entered word appears and returns the record address of the word to the control program. In the case of a search where the user enters a word prefix, or root, and wants all words containing the root to be printed out, a maximum of 100 words at a time are located that contain the root. The string of data record addresses of these

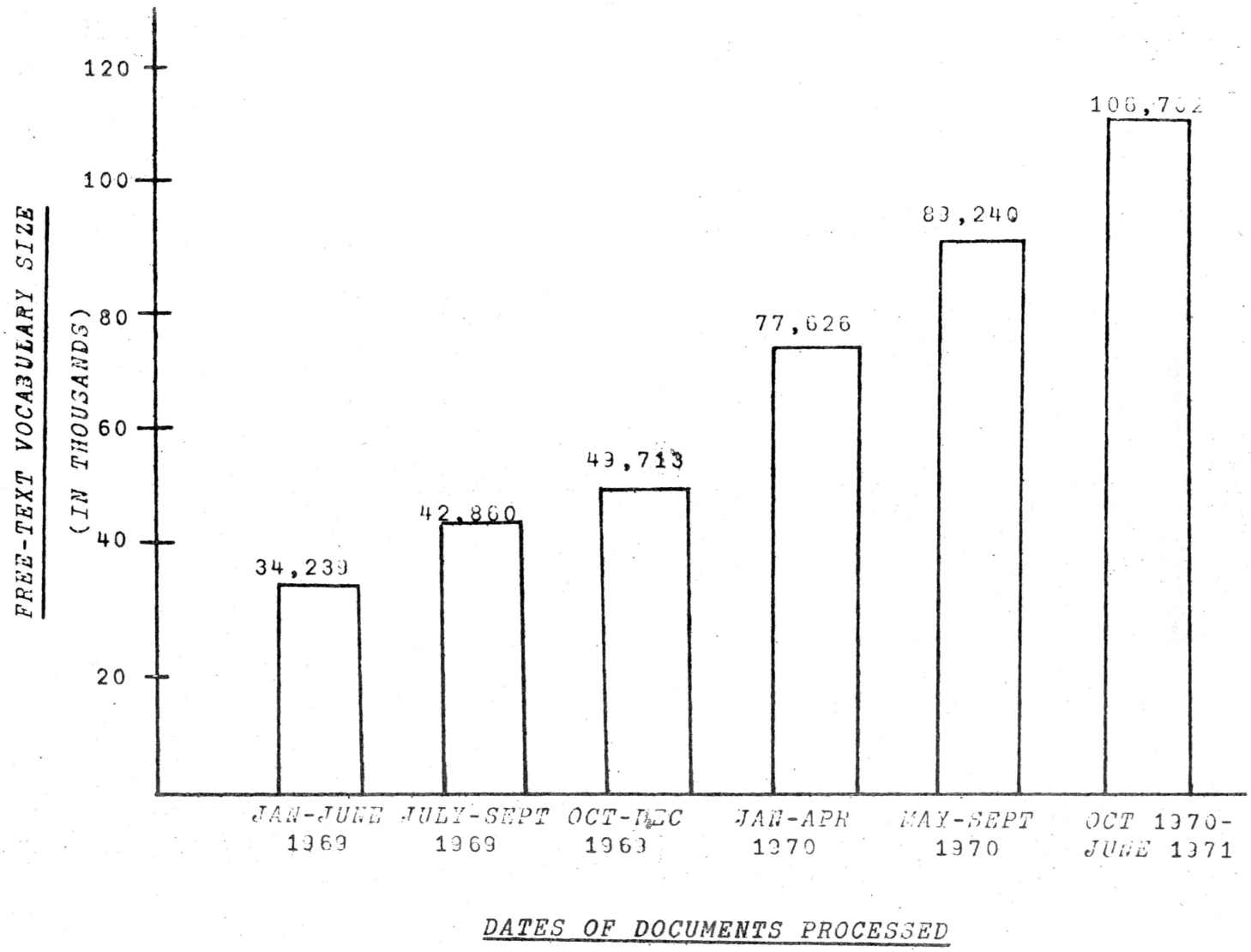


Figure No. 7 Cumulative Growth of the SUPARS/DPS Vocabulary Data Base

words are then returned to the control program. In both cases, when control is returned to the vocabulary program, it reformats the output for the user and collects statistics that are stored for later retrieval.

The third option, which would allow the user to request the printing of words proceding and following a specified word in the vocabulary, was never implemented, although programming was begun. Time restrictions did not allow the programming work to be completed, although initial efforts were begun. The major difficulty in implementation rests in the complicated structure of the DPS dictionary file. The embedded master and block index records greatly complicates the problem of reading and collecting words proceding a specified word, although the task of reading words following the word is not as difficult.

In addition to the problems of developing new search modules to access the vocabulary data base on-line, another major programming effort in the development of the vocabulary data base was the expansion of the total word size capacity of the vocabulary.

b. Expanding the Capacity of the Vocabulary Data Base

One limiting factor in the original DPS program was that the size of the vocabulary (dictionary) file was limited to 65,534 free-text words because of the half-word, or 16-bit coding scheme, i.e. $2^{16} = 65,534$. When this limit was reached, a new file had to be defined. However, the new file would collect all new words, including those in the first file. Each file of documents would have to be searched separately with its own new vocabulary. If a larger capacity could be developed for the vocabulary, only one file would be necessary which would make for less cumbersome and more efficient searching of the data base. A second advantage would be the opportunity for the user to access a complete file of all free-text words that would be available to him in developing search inquiries.

To increase the size of the vocabulary and maintain it in one file, the half-word, or 16-bit coding of a word identifier, was changed to accept full-word coding of 32 bits, i.e. 2^{32}. This change increased the capacity of the vocabulary file from approximately 65,000 words to over 4 billion words, as 2^{32} equals 4,294,967,296.

In order to process 32-bit word identifiers, the format of all data records containing the word identifier field (WID) and all programs creating and referencing those records had to be altered. The WID fields in the Dictionary, Master, and Master Identification Update records and those in the dictionary record area in the search DSECT were increased from two byte (half-word) to four-byte (full-word). Consequently, the relative addresses of all subsequent fields in these records were displaced. The loading programs (PRELOAD, SORT and LOAD), two search modules (KEYWORD and POSITIONAL NOTATION PROCESSOR) and all three versions of the dictionary interface program (DICTIO) had to be changed to handle the new record formats. Programming included converting pertinent half-word instructions to full-word instructions, changing the displacement values for references to all fields following WID in the data records, and altering record-length calculation routines to take

into account the new field length.

In addition to the changes necessitated by the change to full-word word identifiers, substantial changes were made to all search modules. First, the unnecessary coding was deleted to save space. All instructions and variables referring to the synonym, equivalent, and text files were removed, as were unused routines handling weighted keywords and unlabelled search statements. Second, in order to save space in the search monitor and facilitate handling the many different output requirements, the output formatting formerly handled by the search monitor was incorporated into the search programs and new formating routines were written where necessary. Finally, all data entered by the user or written to him by the search monitor program were written into separate intermediate disk files by the search and monitor programs and file access routines were added to all affected programs.

3. THE SEARCH DATA BASE

The third major data base which was developed for SUPARS/DPS II was the search data base which contained the collection of search inquiries that had been previously submitted to the system and subsequently stored during October-December, 1970, and November-December, 1971. The development of this data base simply meant processing each search through the SUPARS/DPS loading programs in the same manner as the document had been. This loading process created a data base containing its own dictionary, vocabulary and master file. The searches were loaded in two separate batches: those of the 1970 period of system operation and those from the 1971 period. A summary of searches loaded and the number of free-text words generated from those searches is shown in Table III, on page 26.

Table IV gives an indication of the size of the three files contained in the complete search data base by the number of tracks used on the two different disk storage packs. The DPS Dictionary file contains all unique words followed by a string of document numbers in which a coded representation of the entire document is entered into the system.

TABLE IV

GROWTH OF SEARCH DATA BASE

Batch	Dates	Searches Loaded	Total Searches Loaded	New Words Added	Total New Words Added
#1	Oct-Dec '70	2,409	2,409	12,016	12,016
#2	Nov-Dec '70	1,826	4,235	5,477	17,493
	TOTALS	4,235	---	17,493	---

TABLE III

TRACK ALLOCATION AND USAGE OF SEARCH DATA BASE

DATA SET NAME	DOCUMENT RANGE	Total Number of Items Stored	Number of Tracks Allocated		Number of Tracks Used	
			2314	3330	2314	3330
STORED SEARCHES						
Dictionary	1970	12,016 words	500	200	59[1]	32
	1970 & 1971	17,493 words	---	200	---	46
Vocabulary	1970	12,016 words[2]	2500	750	271[1]	149
	1970 & 1971	17,493 words	---	75o	---	235
Master	1970	2,409 searches	3380	1000	146[1]	80
	1970 & 1971	4,235 searches	---	1000	---	142
		TOTAL TRACKS: PARTIAL DATA BASE (JAN 69-APR 70+1970 searches)			5312	3081
		COMPLETE DATA BASE (JAN 69-JUNE 71 + 1970-71 stored searches)			n/a	6202

[1]Exact file size not available. Figures listed are estimates.

[2]Dictionary and vocabulary files each contain 1 entry for each unique word. However, the vocabulary occupies more space as each record is a list of all documents in which the word appears.

The development of this search data was made in three stages which are discussed in more detail in the paragraphs below: (1) definition of a data base description, including conversion specifications for document data fields, and punctuation conventions, (2) modifications of the DPS search module and reformat programs in order to load and search the data base of searches in an on-line, interactive mode, and (3) specification of the various forms of search output available to the user.

a. Definition of Data Base Description

Definition of this data base description (DBD), especially the character handling statements for the loading programs, required special care to insure that routine SUPARS/DPS processing of input records did not result in the loss of essential data. Specifically, this essential data included routine treatment of the Boolean operators AND or OR, which DPS defines as common words, and certain punctuation marks (i.e. , ; ()), which standard processing defines out of existence. These problems were solved by (a) adding special characters to the words or converting the symbols we wanted to retain so that the system would store them, and (b) by deleting the extra characters and reconverting the symbols at output time so they appeared normal to the user.

Example 1, Figure 8 contains a sample of a user's previously entered search before any input processing. Example 2 shows how the same search would be stored by DPS if no special reformatting were performed before input processing. Note that all parentheses and Boolean operators would have been deleted and the relationships among the search words would be lost. In Example 3, the search appears in the form it would have taken when the reformat program was run before input processing. In the bibliographic field, dummy characters, which are unprintable, were inserted around Boolean (AND, OR) and other logical operators (?, +, ;) to preserve them from automatic elimination in the standard DPS processing. These dummy characters are converted to parentheses and blanks at output time, as seen in Example 4. In the text field, which contains the actual characters to be searched on, all the added characters are brackets. The adding of brackets around Boolean and other operators forms a new word, i.e. [OR] instead of OR. In this way, these new bracketed words become acceptable search words that can be entered in a search inquiry by the user. This development of new searchable words allows searching of the stored-search data base using Boolean operators as keywords, and allows the system operator to study and analyze the various types of logic used by previous users.

The complete Data Base Description can be found in Figure 1 of Appendix III.

b. Modifying Search Module and Reformat Programs

The search reformat program was written as two modules which are combined by the linkage editor to form a single load module. Appendix III contains flowcharts and program descriptions of each module. In addition, the output record description found in Figure 3 (Appendix III) gives an annotated list of all fields in the search data base records.

Example #1: Previously Entered Search

```
L1  Cat(?)
L2  Rat OR L1
L3  Lab AND Animal(≠1)
L4  L2,L3;
```

Example #2: After Standard DPS Character Handling

```
L1  Cat ?
L2  Rat L1
L3  Lab Animal ≠1
L4  L2 L3
```

Example #3: Stored Search After Reformatting*

Bibliographic Field

```
L1  Cat/?/
L2  Rat //OR// L1
L3  Lab //AND// Animal /≠1/
L4  L2 //OR// L3 //;//
```

Text Field

```
L1  Cat [≠1]
L2  Rat [OR] L1
L3  Lab [AND] Animal [≠1]
L4  L2 [OR] L3 [;]
```

Example #4: Printed Output

```
L1  Cat (?)
L2  Rat OR L1
L3  Lab AND Animal (≠1)
L4  L2 OR L3 ;
```

* Characters / and // represent dummy characters

Figure 8. Modification of Characters in Search Data Base

In the flowchart of Appendix III, the main reformat program, BIBFLDS, reads as input the search statistics records collected by the STATPAC program. It formats all bibliographic fields except NDOCPR and those fields which must be extracted from the user's search entries: keyword and list statements, logic, and a list of all keywords entered. These fields are prepared when the second module, FORMSRCH, is called. This program also creates the text format of the keyword lines and list statement. Control is then returned to the main program which writes this first output record and formats and writes the second record, consisting of the remaining text data.

c. Specification of Output Forms

Output processing of the two forms of output available to the user proceeds under the control of specially written routines. These routines allow the user to request (a) LIST SEARCH which yields a list of all search inquiries that were retrieved, and (b) LIST WORDS, which yields a listing of all the unique words contained in the searches that were retrieved.

The output of the LIST SEARCH option in a search inquiry of the search data base is designed to closely approximate its input format. Because each labeled line of input, such as L1, L2, etc. in Figure 8, appears as a separate line in the printed output, it becomes necessary to check the SRCHA bibliographic field to determine the internal end of line indicators which are inserted by the reformat program.

The LIST WORDS option prints for the user all words in order of decreasing frequency that were found in the searches that are retrieved from a search inquiry. For this option a sort routine accumulates the WRDA and WRDB fields, sorts each word in frequency order, deletes duplicates, and formats output lines for printing.

4. SUMMARY

This section has described three major data bases that were developed for the SUPARS/DPS II research. On-line access was made available to (1) a document data base, (2) a vocabulary data base, and (3) a search data base. Data was presented to show the growth rate of the units contained in each of the three data bases which were documents (bibliographic citations and abstracts) for the document data base, free-text words for the vocabulary data base, and search inquiries for the search data base. In addition, programming changes that were necessary to make the vocabulary and search data bases accessible and searchable in an interactive, on-line mode were explained.

5. IMPLICATIONS AND PROJECTIONS

The successful development and implementation of interactive algorithms based on SEARCH and VOCABULARY data bases has implications for the future use of human intelligence to augment the user's searching negotiations.

For example, under the present experimental system, a user has the ability to search through the list of free-text terms in the vocabulary for potential search words. Once the user has a sufficient number of terms, search inquiries are entered and documents retrieved. One alternative that could be considered is the development of what is known as a synonym-equivalent list (SEL). In standard DPS, terms considered to be synonyms are appended to the free-text terms and are used in retrieval much like an individual manually uses "See also" references in a controlled index. For example, in the batch-mode version of DPS, the SEL would take a user's search word in an inquiry, and its equivalent or synonym words initially supplied by the system operator, and process the inquiry using not only the user submitted terms but the equivalent terms as well. However, this process requires the intervention of a human indexer or subject expert to determine synonym words and build a list of equivalent words.

One projection of the synonym-equivalent idea using human intelligence without the intervention of a decision by a selected group of indexers or experts would be the following:

Because all search inquiries previously submitted by users are stored in the search data base, the word associations in this body of inquiries could be available as data to input and automatically generate a synonym-like equivalent list.

For example, if an inquiry originally entered the intersection of the term "Achievement" and the term "Motivation," a synonym-like term related to "Achievement," as the user originally decided to associate the two words, would be "Motivation."

These word associations from inquiries could be automatically incorporated into the SUPARS/DPS inverted file for the document data base using the available SEL module. The user would then have the option of entering an inquiry and selecting or not selecting the synonym equivalent function. By keeping the synonym equivalent words in their order of decreasing frequency, the user could also be able to select a "cutoff" figure of the first "x" terms available in the list, rather than having to have his search inquiry processed on the basis of all the terms in the listing. Initially, batch-mode updates of the equivalent list would be periodically made from the search inquiries of the search data base.

Another related projection of this idea of an automatically generated synonym-equivalent list based on human intelligence would be the ability of the user to interactively query the list. In effect this option would form a "synonym-equivalent data base" that could be interrogated by the user during the developing of alternative inquiries, or could be employed as an automatic means of extending the original search inquiry.

Some of the advantages of using a synonym-equivalent list developed in this way would include the fact that (1) human intermediaries, experts, or indexers would not be needed to make decisions about word associations and links, (2) the equivalent list would represent a current, and up-to-date

option based on the cumulative intelligence of the user population, and (3) the user would have the ability to interactively query the equivalent list as an additional source of potential keywords for a search inquiry.

The potential disadvantages of a humanly generated equivalent list would be the necessity for continual updating of the list, and the fact that terms in the list might possibly not be considered realistic synonyms by some users.

The capability for developing and implementing these projections of the current work is available in terms of the programming and systems changes necessary. The larger issue in the overall system is the cost-performance level of the equivalent list in relation to the other interactively available algorithms in the SUPARS/DPS system.

APPENDIX I. SYSTEM OVERVIEW

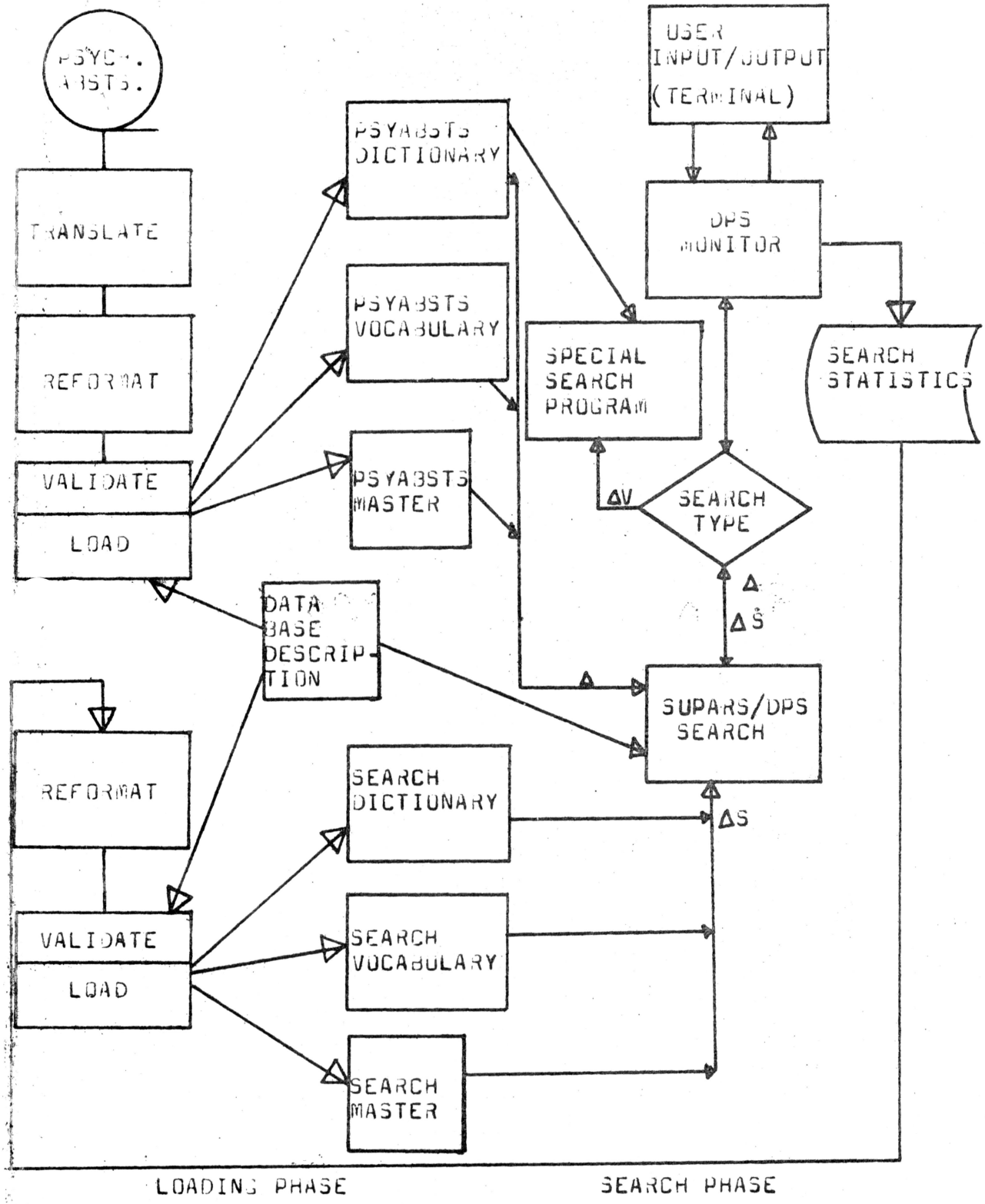


Figure 1. System Overview

APPENDIX II

TRANSLATE PSYCHOLOGICAL ABSTRACTS

TRANSLATE is a BAL program written to run on the IBM S/360 Model 50. It examines tapes of Psychlogical Abstracts records on a character-by-character basis for membership in a selected subset of valid characters. When possible invalid characters are converted to valid ones and the record is written to an output file; when this is not possible the record is written to an error file.

Computer Definition

1. IBM S/360 Model 50
2. Three 2400 tape drive facilities and 9-track tapes
3. Model 1403 Printer
4. Core requirements:
   a. Assembler 140K
   b. Linkage Editor 128K
   c. Program execution 34k

System Description

1. Operating System: Syracuse University Operating System (SUOS)
2. Assembler Level F translator program
3. Linkage Editor Level F program

Program Description

This program reads an input tape of Psychological Abstracts records and examines each byte to insure that it is a member of the character set defined as valid by the American Psychological Association. Certain characters are translated to more meaningful APL characters. Records containing invalid characters are written to an error tape and a message to that effect is printed. Valid records are written to a tape of translated documents.

Input

Each tape contains a copyright statement as the first record. It is followed by the data records and an end-of-file mark.

Data records are composed of 4 different types of fields: fixed length fields, directory fields which reference variable length fields, and variable numbers of fixed length fields.

Output

1. TRANSLATED tape - see Input record description

Displacement	Data Item	Comments
	Fixed Fields	
0- 3	Record length	Binary control field
4- 5	Generation code	
6- 9	Year	
10-11	Volume number	
12-13	Issue Number	
14-18	Abstract number	
19-20	Type of publication	Numeric code for 1971 documents only--blanks for 1970 documents
21-26	Journal title code	
27-30	Language	blank or FRGN
31-34	Availability	
	Directory Fields	
35-38	Subject Index Codes	All directory fields are right justified numerics, blank-filled on the left. The fields contain the displacement of the first byte of the corresponding variable length data field relative to the first data byte (generation code) of the record.
39-42	Subject index phrase	
43-46	Author Subdirectory	
47-50	Designator other than author	
51-54	Affiliation of first author	
55-58	Publication title	
59-62	Source document title	
63-66	Source document description	
67-70	Abstract	
71-74	Abstractor's name	
	Variable Number of Fixed Length Fields	
75-76	Number of classification codes	This is the last field which begins at a fixed displacement.
	Classification codes	Each is 6-digit code
	Number of subject index codes	
	Subject index codes	Each is 5-digit code
		All are left justified
	Subject index phrase	
	Author Subdirectory	2 characters: right justified count of number of authors
	Author(s)	2 characters each: Number of characters in each author's name
	Designator	
	Affiliation	
	Publication title	
	Source document title	
	Source document description	
	Abstract	
	Abstractor's name	

Figure 1. Input Record Description

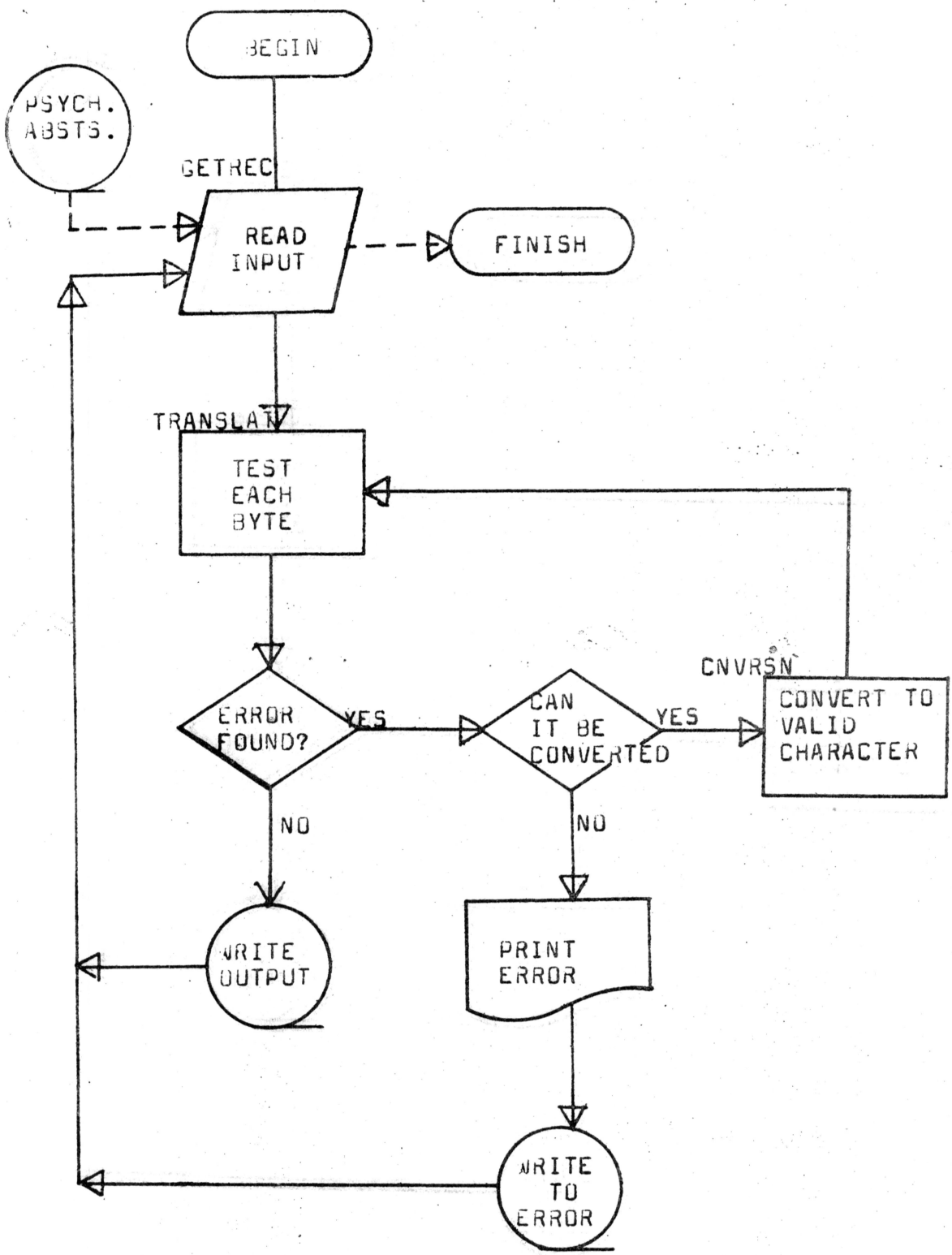


Figure 2. Translate Psychological Abstracts Logic Diagram

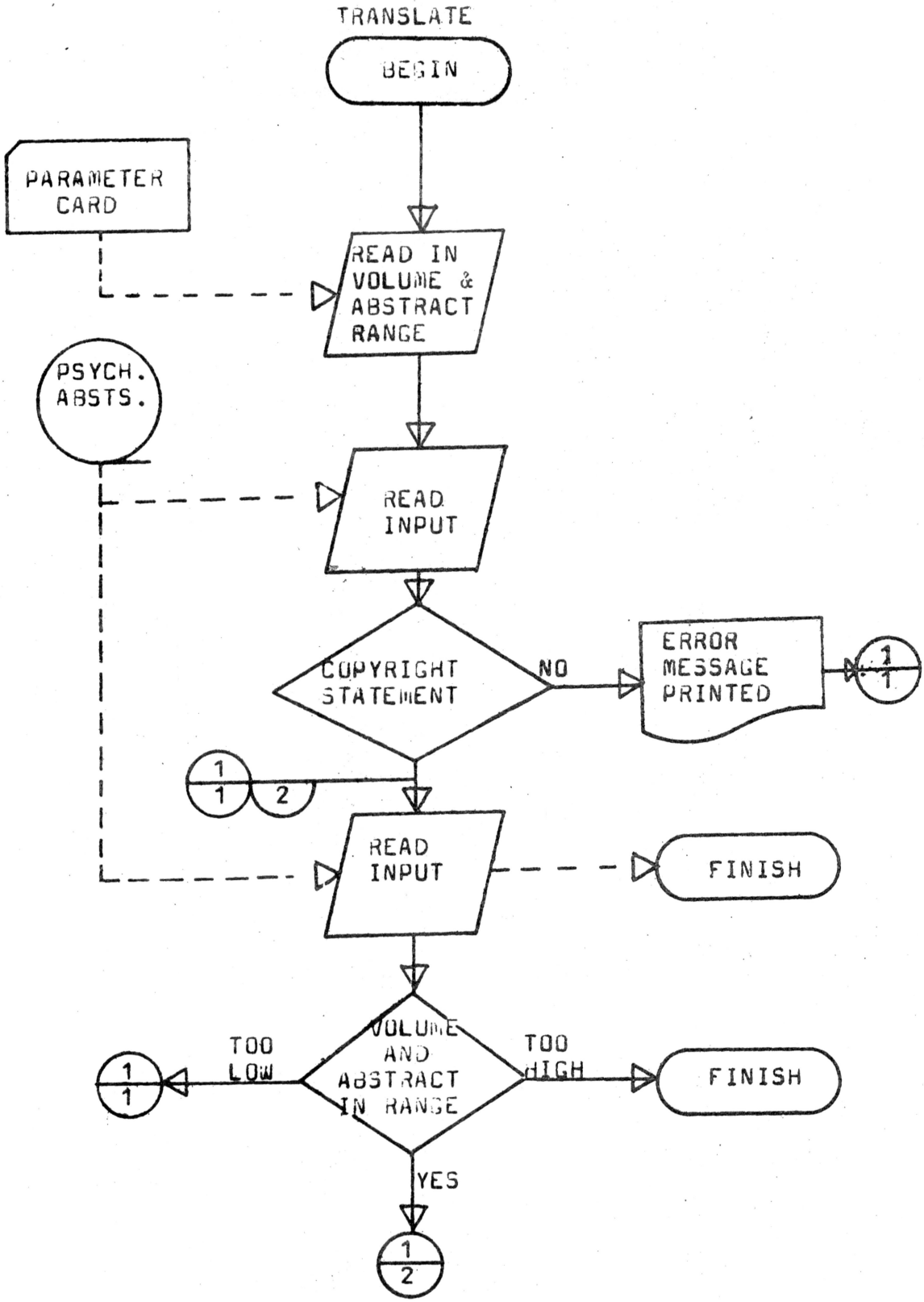


Figure 3. Translate Psychological Abstracts Flow Chart

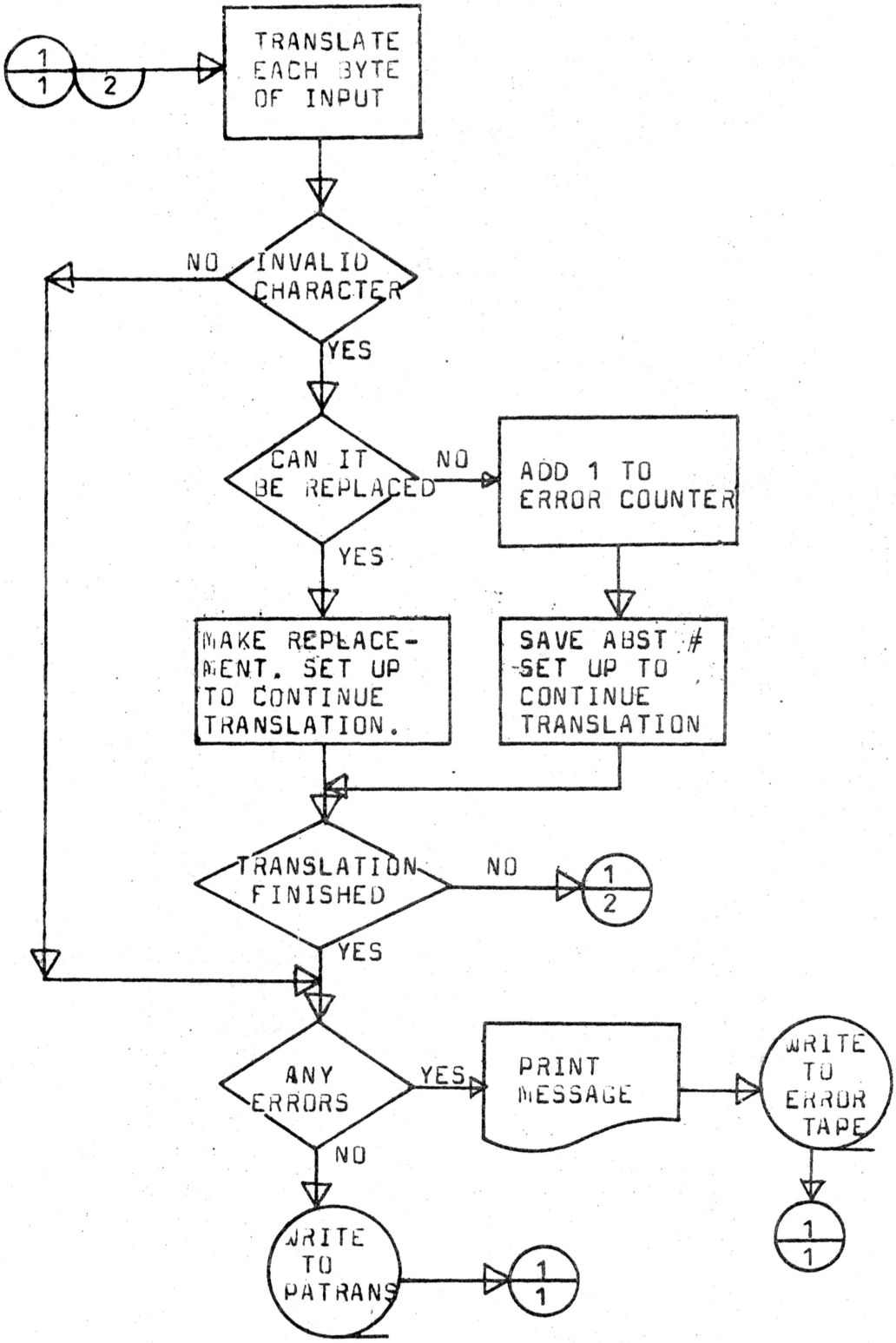


Figure 3. (Continued)

2. ERROR tape - see Input record description

3. PRINTED OUTPUT

The following messages may be printed:

(a) TAPE HAS NO COPYRIGHT MESSAGE

(b) ABSTRACT NUMBER nnnnnn HAS xxxxxx ERROR(S)

where nnnnnn and xxxxxx are the abstract number and error count as computed by the processing program.

Program Name: REFORMAT

ABSTRACT

REFORMAT is a BAL program written to run on the IBM S/360 Model 50. Its purpose is to reformat the data contained in each Psychological Abstracts record into a format that is compatible with the SUPARS/DPS input record description. It accomplishes this by rearranging the fields, inserting termination characters, truncating field which exceed the maximum acceptable length, and designating sentence terminators.

Computer Definition

1. IBM S/360 Model 50
2. Two 2400 tape drive facilities and 9-track tapes
3. Model 1403 Printer
4. Core requirements:
   a. Assembler 140K
   b. Linkage editor 128K
   c. Program execution 20K

System Description

1. Syracuse University Operating System (SUOS)
2. Assembler Level F translator program
3. Linkage Editor Level F program

Program Description

This program takes as input the TRANSLATED Psychological Abstract tape. It processes 1 input record at a time and produces either 2 or 3 output records for each. Each document is assigned a DPS assension number. Then the fields are broken down and reconstructed into a format suitable for DPS processing. The first record contains all bibliographic fields with their termination identifiers and, if there is room, the reformatted

abstract. The abstract is rewritten so that the character handling statements will process the punctuation properly. If the abstract will not fit in the first record, it is outputted as the second record for that document number. If the abstract is too long for a record it is truncated to the maximum length allowable for an output record -- 1646 characters. The last record for each document is the text portion, paragraph B, sentences 1 through 4.

Input

See input record description for Translate Program (Figure 1.)

Output

1. Printed output

The following messages may be printed when this program is run:

COPYRIGHT STATEMENT MISSING

LENGTH ERROR FOR ABST NUMBER nnnnn - if a field or the entire record exceeds limits by DPS.

ERROR IN DIRECTORY FIELD OR AUTHOR SUB-if a non-numeric field found

END OF PROCESSING - for successful termination of job

LAST DOCUMENT NUMBER ASSIGNED WAS xxxxxx

2. REFORMATED DATA

For each input record, 2 or 3 output records are produced. If there are 2 output records, the first contains the bibliographic data and Text Paragraph A; otherwise the bibliographic data and Text Paragraph A data are separate records. Text Paragraph B is always the last output record. Each output record is preceeded by a 4 byte control field containing the record length. (See Figure 4.)

Field Length	Data Item
	Bibliographic Data
6	DPS Ascension Number
4*	Year
2*	Volume Number
5*	Abstract Number
Variable***	Author
	Editor
	Affiliation
	Article Title
	Source document title
	Source document description
	Language
	Type of Work
	Random Number
	Abstract
	Text Paragraph A
4	Paragraph indicator
Variable	Abstract
	Text Paragraph B
6	Document Number
Variable***	Article Title
Variable***	Source Doc. Title
Variable***	Author
Variable***	Affiliation

*Field is terminated with the character ¬ to meet DPS requirements. This character is not included in the listed field length.

**Field length range is 1 to 255 characters. No field entry in the input document is indicated by one blank as the bibliographic entry; input fields exceeding the maximum are truncated to 254 characters and an asterisk is added as the final character to signal truncation.

***Entry is followed by a sentence indicator.

Figure 4. Reformatted Data

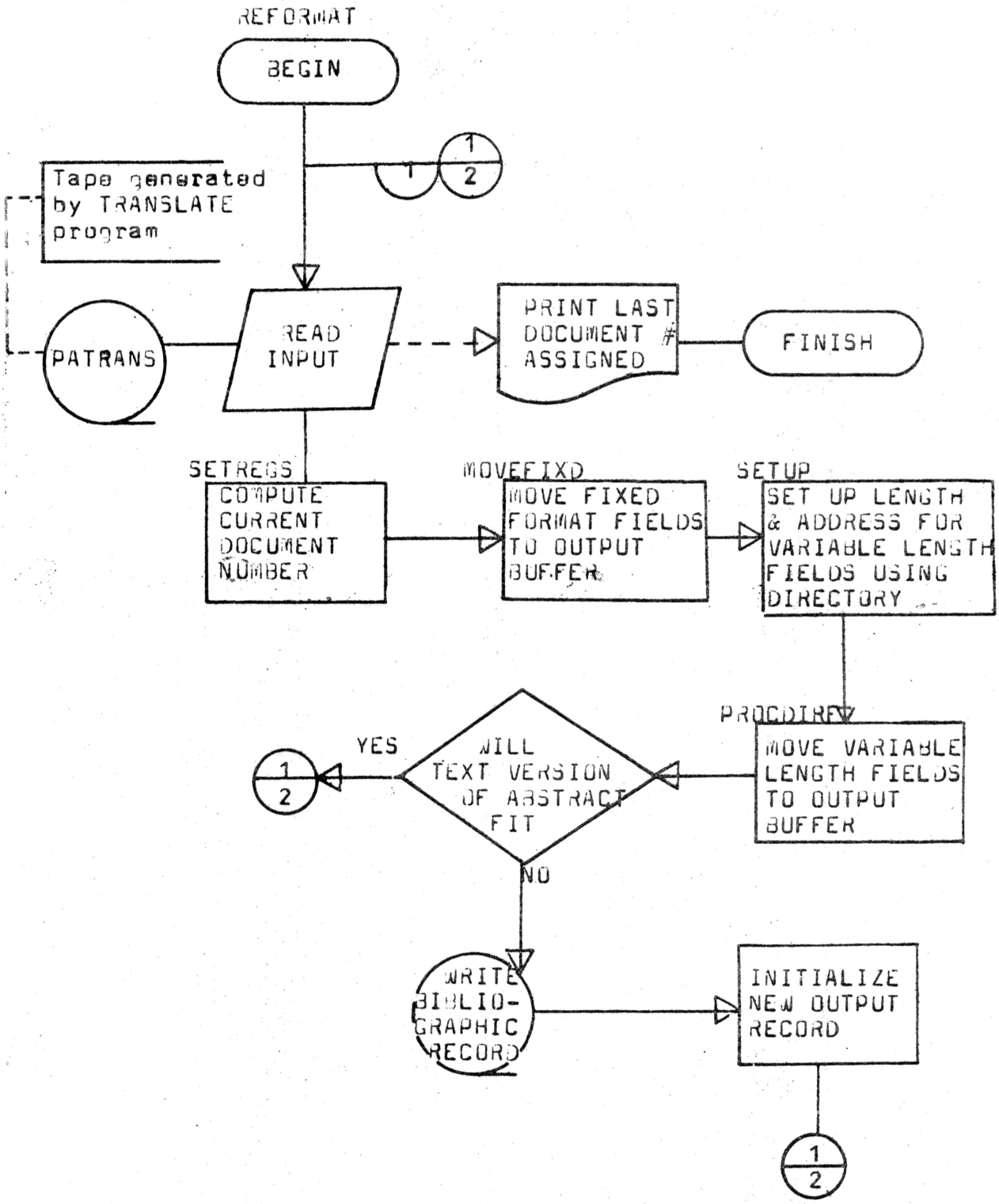


Figure 5. Reformat Psychological Abstracts Logic Diagram

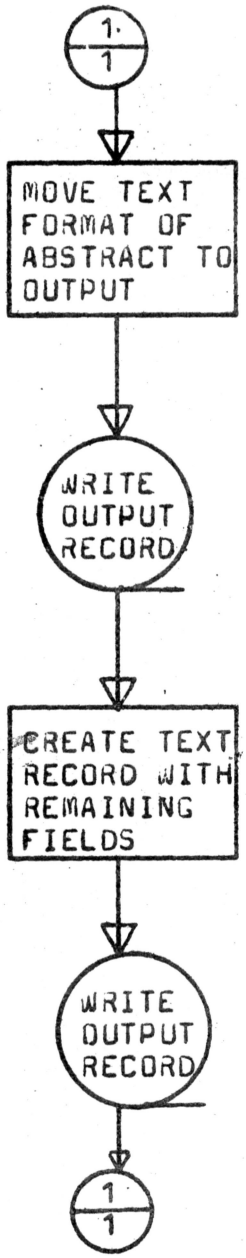


Figure 5. (Continued)

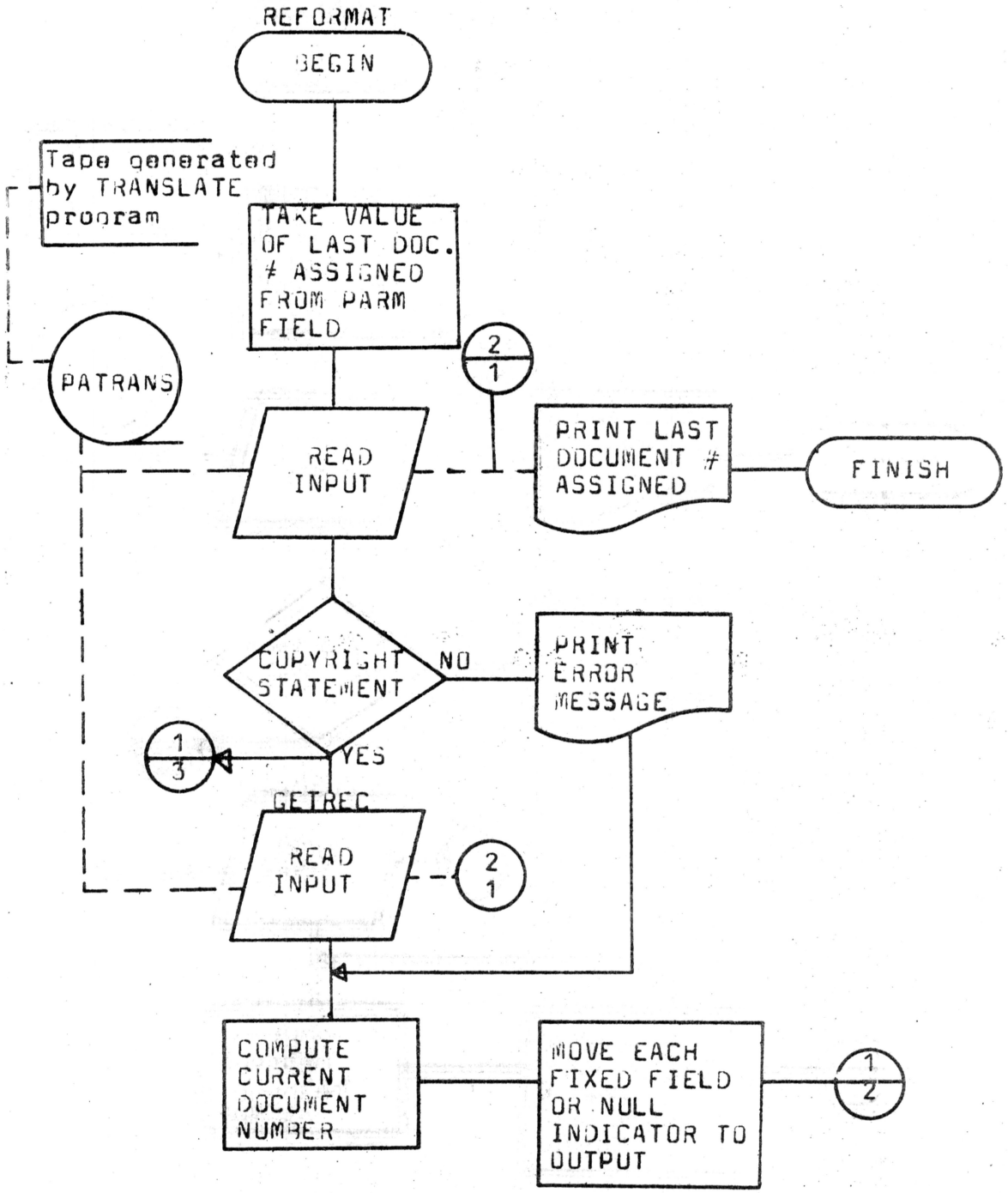


Figure 6. Reformat Psychological Abstracts Flowchart

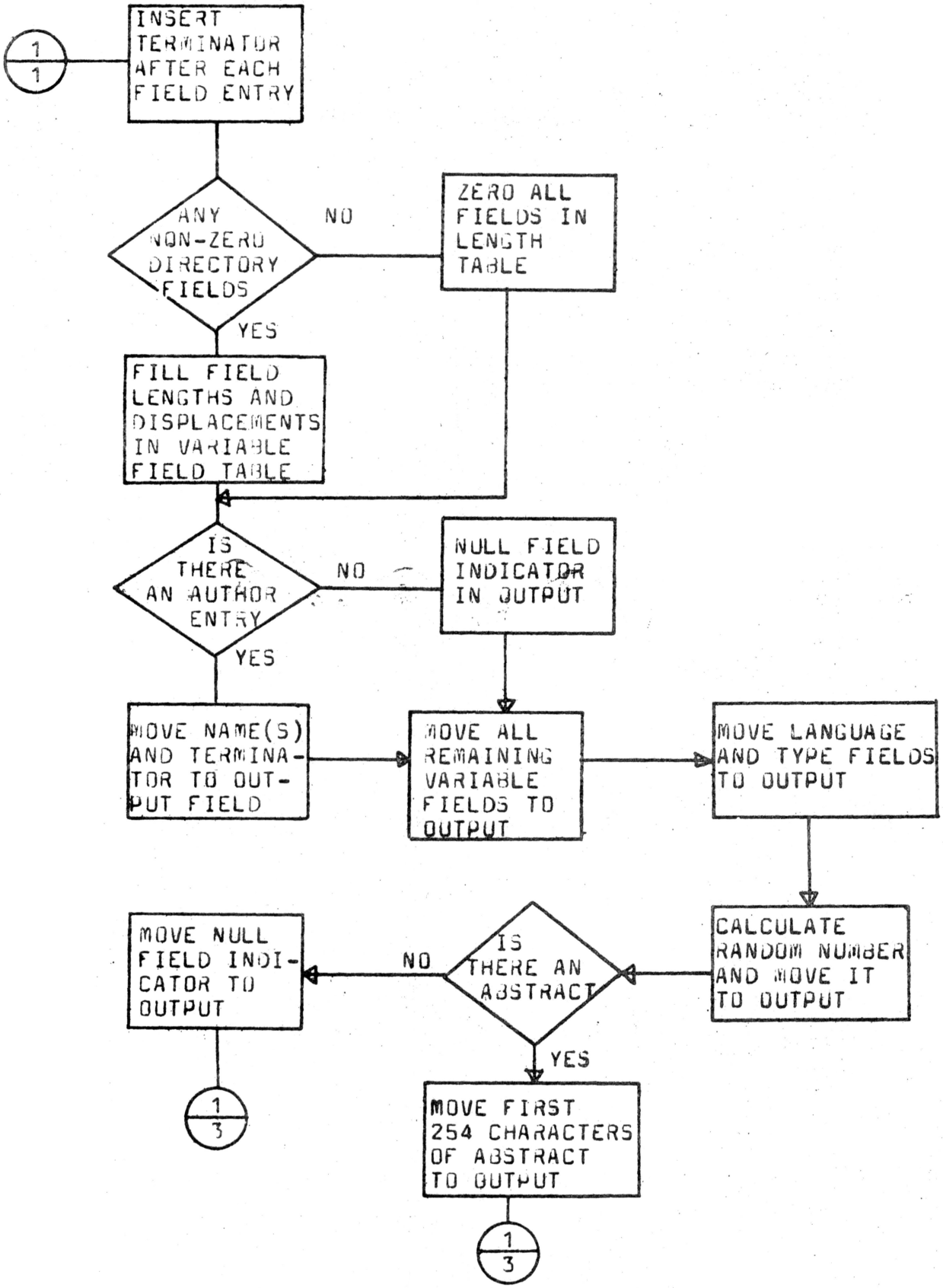


Figure 6. (Continued)

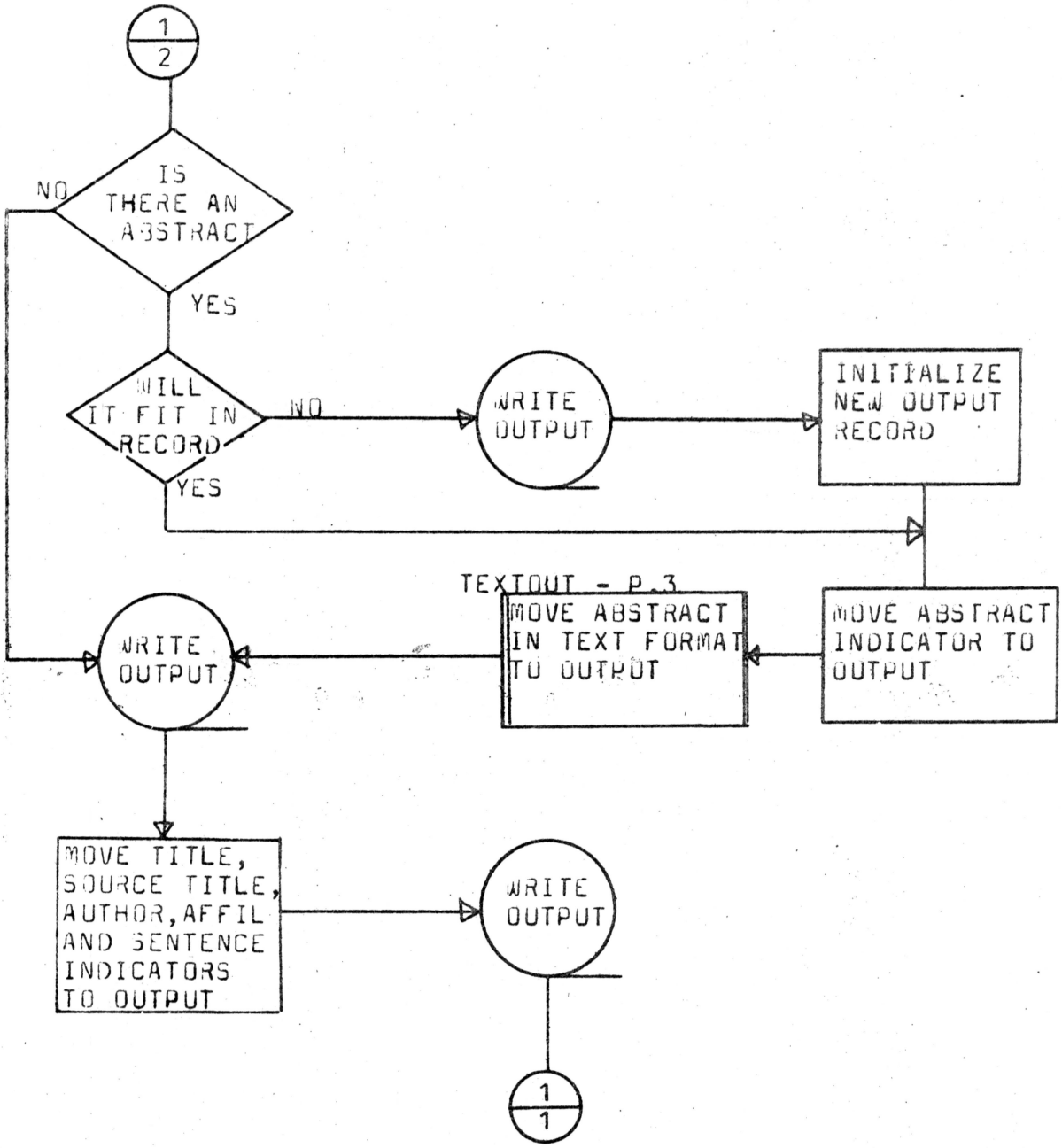


Figure 6. (Continued)

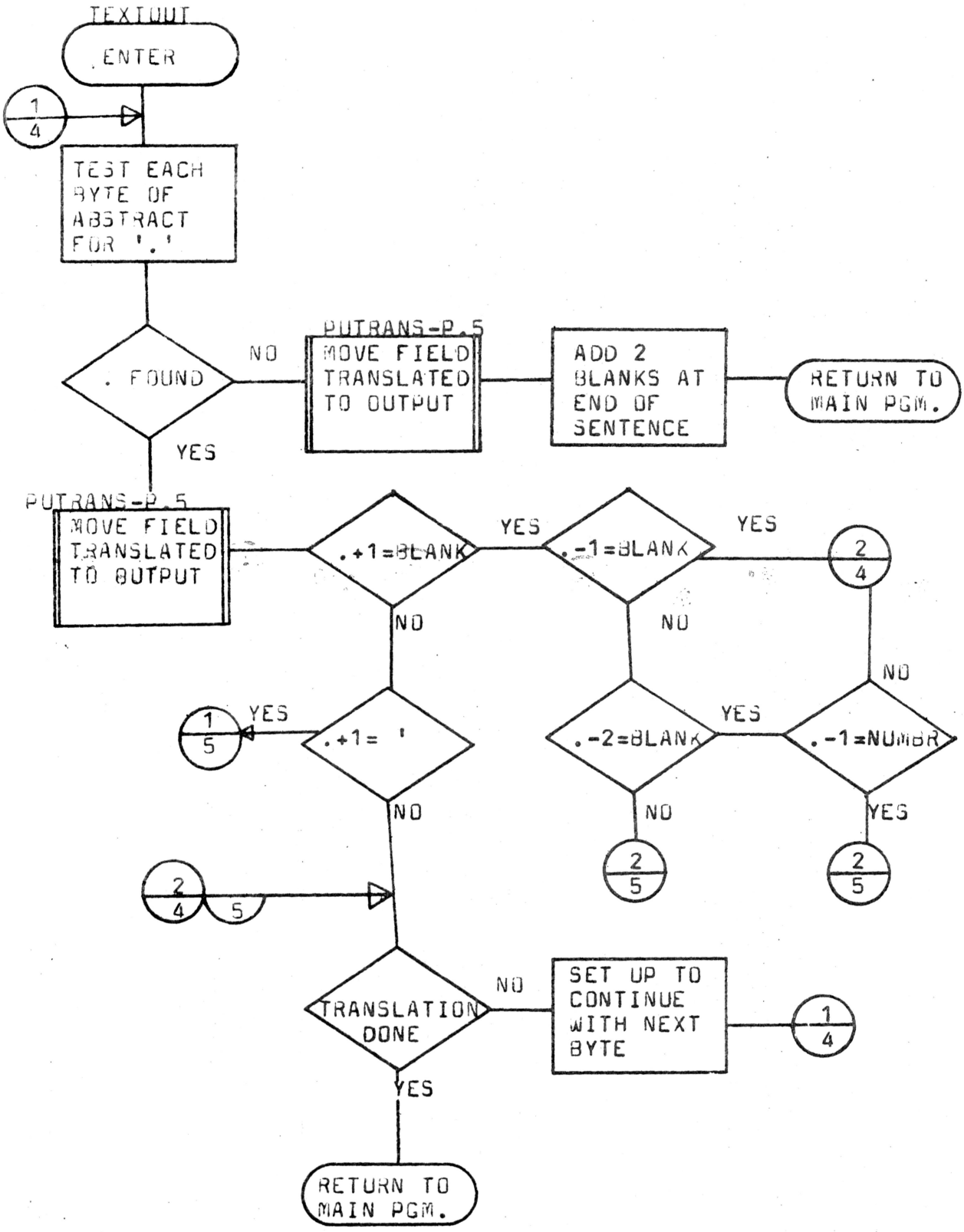


Figure 6. (Continued)

APPENDIX III

SEARCH REFORMAT

Program Name: SBCRFMAT

ABSTRACT

SBCFRMAT is composed of two BAL programs written to run on both the IBM S/360 Model 50 and the IBM S/370 Model 155. The modules reformat data collected by the SUPARS monitor into the format defined by the SEARCHES data base description, a format acceptable as input to the SUPARS/DPS loading and search programs.

Computer Definition

1. IBM S/360 Model 50 or S/370 Model 155
2. Two 2400 tape drive facilities and 9-track tapes
3. Model 1403 Printer
4. Core requirements:

    a. Assembler 140K
    b. Linkage Editor 128K
    c. Program Execution

System Description

1. Syracuse University Operating System (SUOS)
2. Assembler Level F translator program
3. Linkage Editor Level F Program

Program Description

This program is two BAL modules combined into one load module by the linkage editor. BIBFLDS is the main program. It reads as input the statistic records collected for it by the STATPAC programs, processing 1 input record at a time and producing for each 2 output records. Each document is assigned a DPS ascension number. Then the bibliographic fields are extracted from the raw data and moved to the output record in DPS format. Control is passed to the second module, SBCFRMAT, which reformats the remaining bibliographic fields and Paragraph A of the text portion of the record. Control is returned to the main program and output Record 1 is written. The remaining text data is formatted and written as Output Record 2. Processing continues until all documents have been reformatted. Before terminating a message is written out of the last document number assigned.

Figure 1, used in conjunction with the DPS Program Description and Operations Manual (H20-0477-1), pages 27-47, gives the complete data base description for the search data base.

```
CONCORD SEARCHES Y
FILE SEARCHES              V 1650 9999                    ;
FLD DOCNO        EBCD      6                              ;
FLD SSN                    9 TER ¬              1         ;
FLD LOGN                   6 TER ¬              1         ;
FLD TERN                   3 TER ¬              1         ;
FLD DATE                   8 TER ¬              1         ;
FLD ECPU                   6 TER ¬              1         ;
FLD ECLOCK                 6 TER ¬              1         ;
FLD LCOST                  7 TER ¬              1         ;
FLD MAXDOF                 6 TER ¬              1         ;
FLD SRCHA                255 TER ¬              1         ;
FLD SRCHB                255 TER ¬              1         ;
FLD LOGIC                 24 TER ¬              1         ;
FLD WRD                  255 TER ¬              1         ;
FLD WRDA                 255 TER ¬              1         ;
FLD NDOCPR                 5 TER ¬              1         ;
FLD TXT                   30 TXT                1         ;
SPCL 075;
SENT 080 094 107;
TRNS NONE     ;
DMDBD       DD   DSNAME=DEB.DBD,DISP=OLD,UNIT=2314
DMINPUT     DD   DDNAME=READR
DMREFUD     DD   UNIT=2314,SPACE=(TRK,(20,10)),DSNAME=SRCREF,VOL=SER=LB0005
DMCNCRD     DD   UNIT=2314,SPACE=(TRK,(50,10)),DSNAME=SRCCNC,VOL=SER=LB0005
DMTEXTS     DD   DUMMY
DMDICTN     DD   UNIT=2314,DSNAME=SRDICTN,DISP=OLD,VOL=SER=LB0005
DMVOCAB     DD   UNIT=2314,DSNAME=SRVOCAB,DISP=OLD,VOL=SER=LB0005
DMMASTR     DD   UNIT=2314,DSNAME=SRMASTR,DISP=OLD,VOL=SER=LB0005
DMWRKT1     DD   UNIT=2314,SPACE=(TRK,(50,10)),DSNAME=SRCWR1,VOL=SER=LB0005
DMWRKT2     DD   UNIT=2314,SPACE=(TRK,(50,10)),DSNAME=SRCWR2,VOL=SER=LB0005
DMWRKT3     DD   UNIT=2314,SPACE=(TRK,(50,10)),DSNAME=SRCWR3,VOL=SER=LB0005
DMSORTWK01  DD   UNIT=2314,DISP=OLD,DSNAME=SYS1.UT1
DMSORTWK02  DD   UNIT=2314,DISP=OLD,DSNAME=SYS1.UT2
DMSORTWK03  DD   UNIT=2314,DISP=OLD,DSNAME=SYS1.UT3
DMSORTWK04  DD   UNIT=2314,DISP=OLD,DSNAME=SYS1.UT4
DMSORTLIB   DD   DSNAME=SYS1.SORTLIB,DISP=OLD,DCB=(BLKSIZE=3265,RECFM=U)
DMSYSOUT    DD   SYSOUT=A
DMWORK1     DD   UNIT=2314,SPACE=(TRK,(50,10)),DSNAME=SRCWS1,VOL=SER=LB0005
DMWORK2     DD   UNIT=2314,SPACE=(TRK,(50,10)),DSNAME=SRCWS2,VOL=SER=LB0005
DMWORK3     DD   UNIT=2314,SPACE=(TRK,(50,10)),DSNAME=SRCWS3,VOL=SER=LB0005
END
```

Figure 1. Data Base Description

Displacement	Data Item	Type
0- 3	Record length	Binary
4- 8	Social Security Number	Packed Decimal
9-11	Log number	Packed Decimal
12-15	Date (YYMMDD)	Packed Decimal
16-19	Elapsed CPU (1/300 sec.)	Binary
20-23	Elapsed CLOCK (1/300 sec.)	Binary
24-27	Maximum Documents Dropped	Packed Decimal
28-29	List Length	Binary
30	Terminal Number	Binary
31	Δ Type	Binary
32-35	Cost	Binary
36-37	User I/P Length	Binary
38	Error Flag	Binary
39-41	No. Docs Printed	Packed
42	List type	Binary
43-n	User Input	Character
Var	User Output	Character
Var	List	Character

Figure 2. Input Record Description

Output: Two variable length records for each log number

Record 1. Bibliographic data and first paragraph of text data.

Field Name	Displacement	Length	Data item and comments
	0	4	Record length
DOCNO	4	6	DPS Ascension Number
SSN	10	9	Social security number (in alphabetic code)
LOGN	20	6	Log number
TERN	27	3	Terminal number (port number for hardwires, code for dial-ups)
DATE	31	8	Date (YY-MM-DD)
ECPU	40	6	Elapsed CPU time (mmm:ss)
ECLOCK	47	6	Elapsed clock time (mmm:ss)
LCOST	54	7	Search cost (*ddd.cc)
SRCHA	62	variable	Labels, keywords, and operators
SRCHB	variable	variable	LIST statement
LOGIC	variable	24	Frequency of occurance of each operator in same order as first 7 entries of LOGIC list for Record 2.
WRD	variable	variable	Length count and keywords from search
WRDA	variable	variable	Overflow field for WRD. Used only if keywords exceed 254 characters.
NDOCPR	variable	5	Number of documents printed
	variable	4	Paragraph A indicator
	variable	variable	Search, from L1 through Ln in text format, or NONESRCH
	variable	variable	LIST statement

Record 2. Second paragraph of text data.

	0	4	Record length
	4	6	DPS Ascension Number
	10	4	Paragraph B indicator
	14	9	Social security number
	23	variable	L+LOGN+L or LNONE
	variable	variable	T+TERN+T or TNONE
	variable	variable	D+MMDDYY+D or DNONE
	variable	variable	LIST type - LSTBRIEF,LSTRECORD, LSTOTHER, or NOLST
	variable	variable	LOGIC - from 1 to 7 entries depending on search
	variable	variable	Completeness of search
	variable	variable	User output in format VnnAnnnnn or V00ANONE

Figure 3. Output Record Description

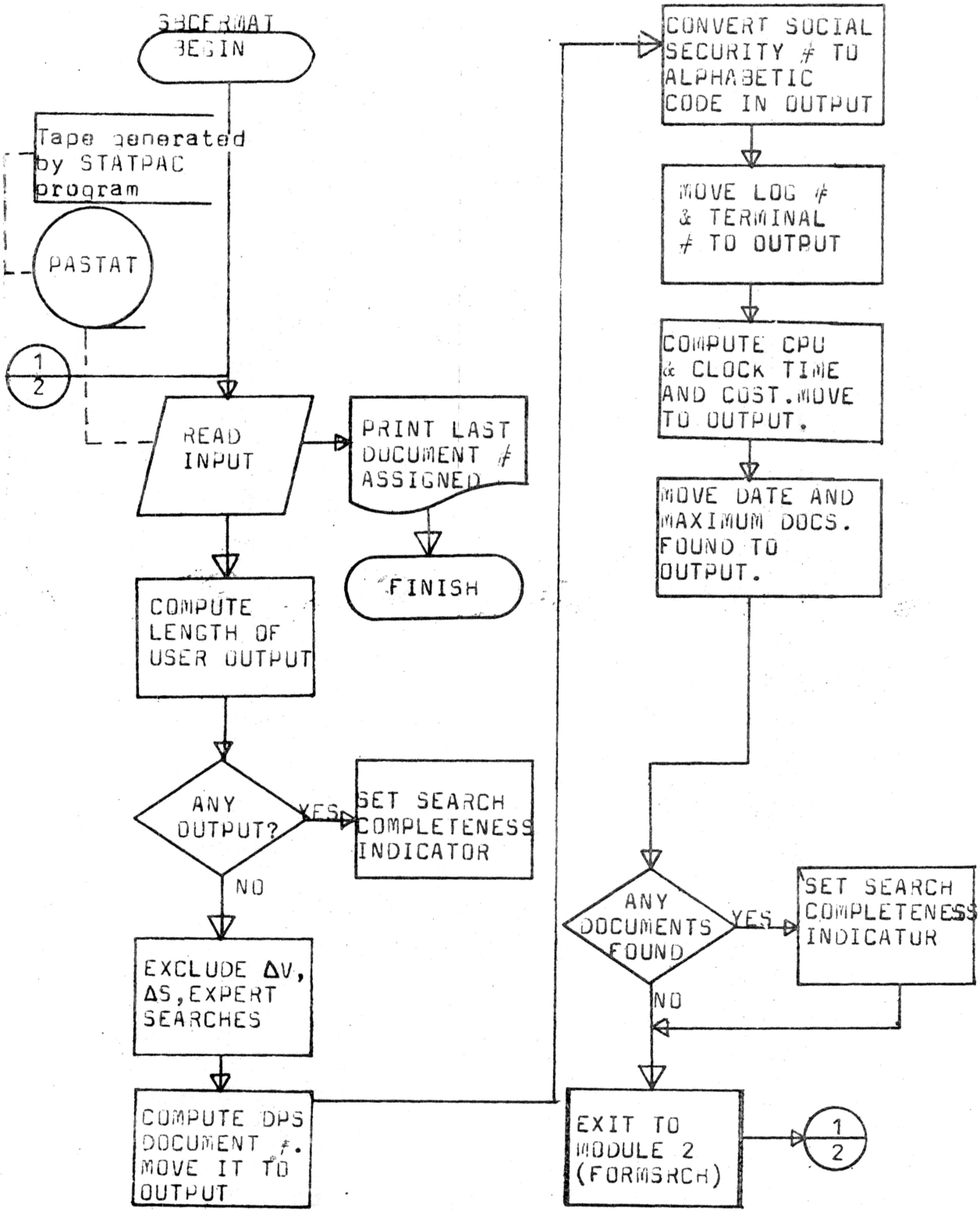


Figure 4. Reformat Searches Module 1 - BIBFLDS

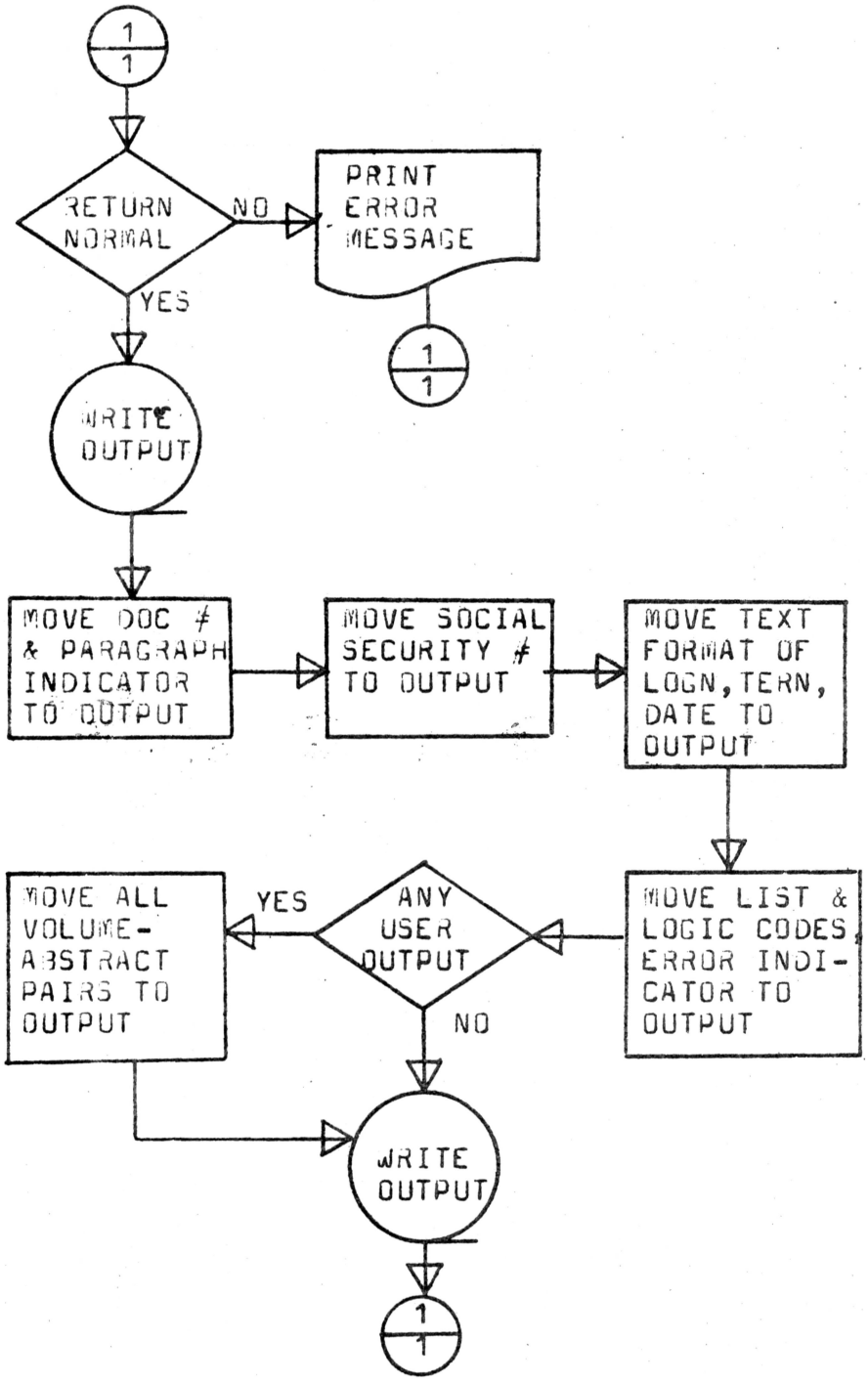


Figure 4. (Continued)

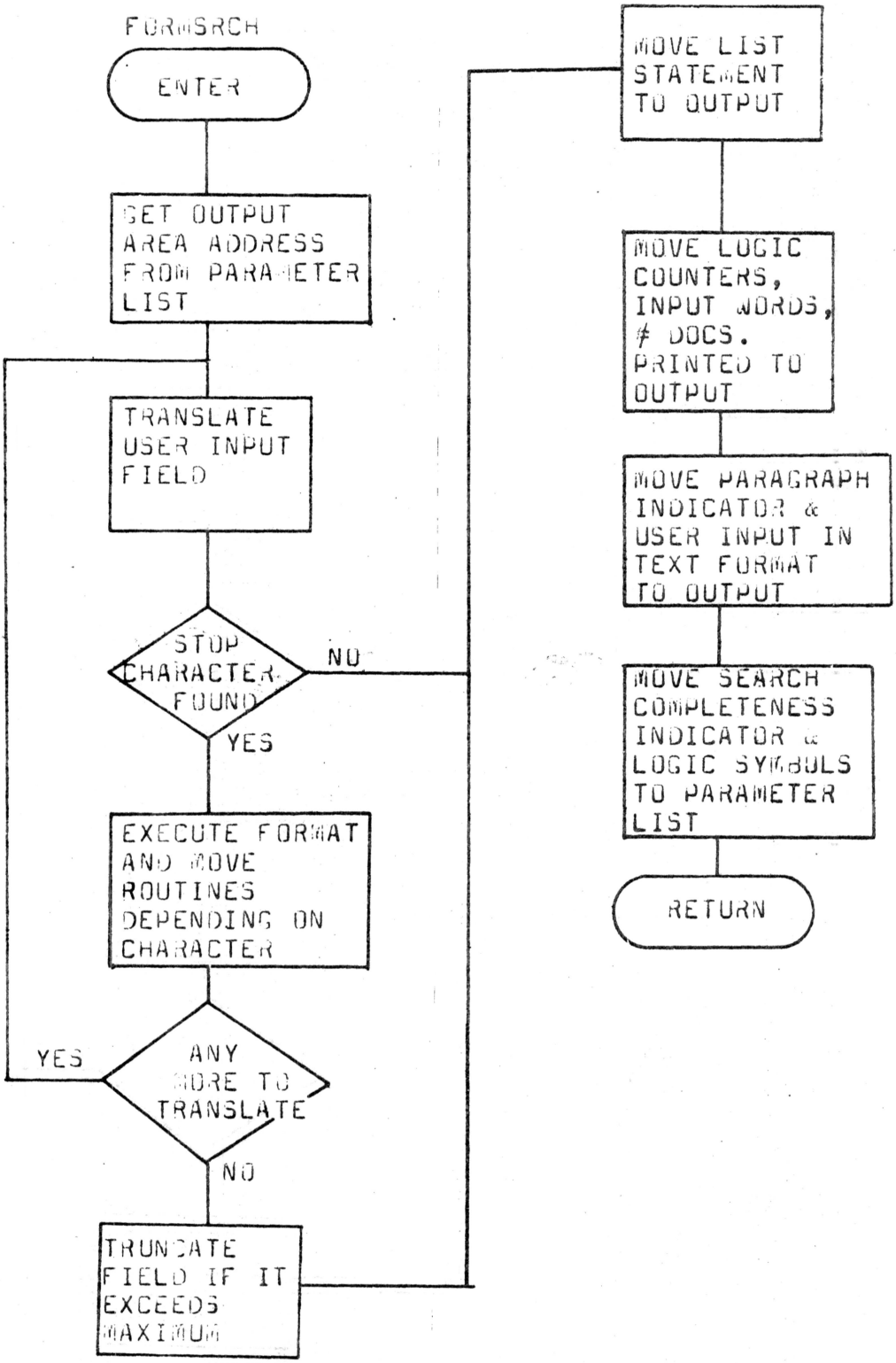
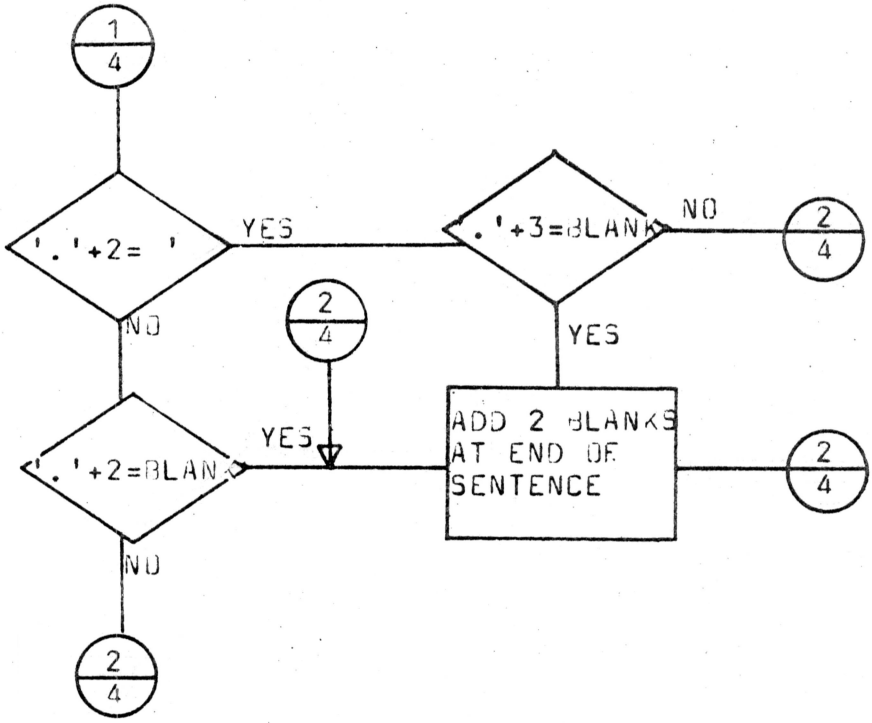


Figure 5. Reformat Searches Module 2 - FORMSRCH



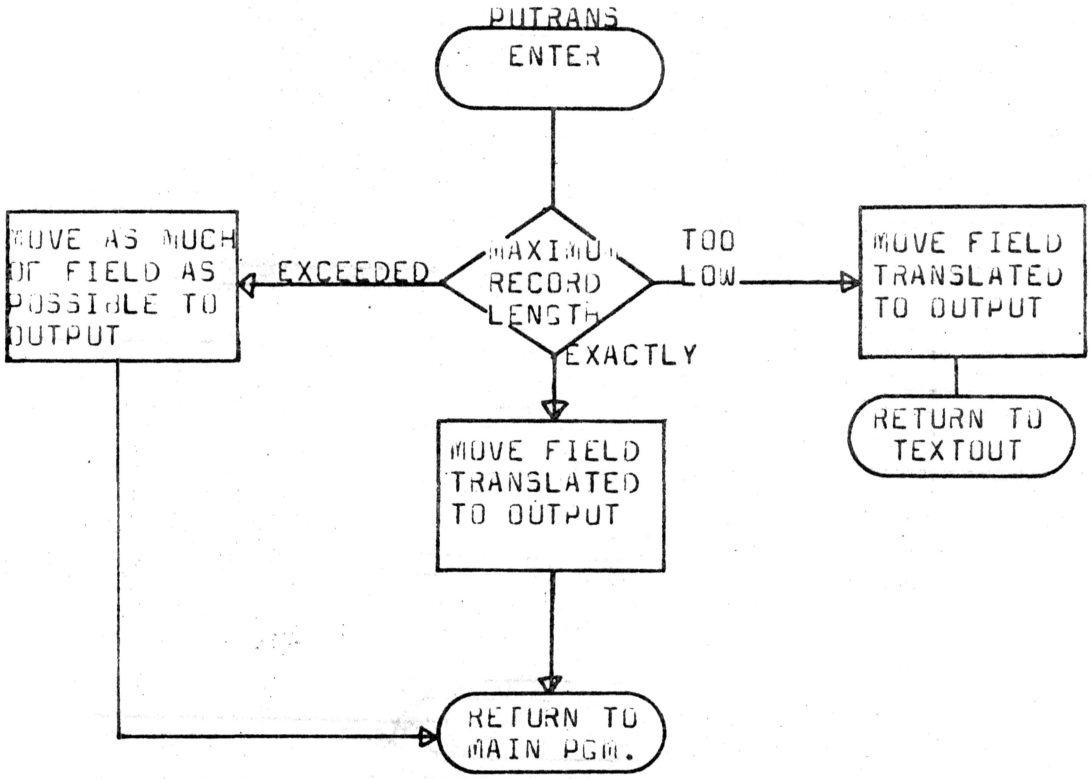


Figure 6. (Continued)