EPA-660/2-74-048
June 1974

Environmental Protection Technology Series

Implementation of A Computer-Based Information System For Mass Spectral Identification

Office of Research and Development
U.S. Environmental Protection Agency
Washington, D.C. 20460

RESEARCH REPORTING SERIES

Research reports of the Office of Research and Monitoring, Environmental Protection Agency, have been grouped into five series. These five broad categories were established to facilitate further development and application of environmental technology. Elimination of traditional grouping was consciously planned to foster technology transfer and a maximum interface in related fields. The five series are:

1. Environmental Health Effects Research
2. Environmental Protection Technology
3. Ecological Research
4. Environmental Monitoring
5. Socioeconomic Environmental Studies

This report has been assigned to the ENVIRONMENTAL PROTECTION TECHNOLOGY series. This series describes research performed to develop and demonstrate instrumentation, equipment and methodology to repair or prevent environmental degradation from point and non-point sources of pollution. This work provides the new or improved technology required for the control and treatment of pollution sources to meet environmental quality standards.

EPA-660/2-74-048
June 1974

IMPLEMENTATION OF A COMPUTER-BASED INFORMATION SYSTEM FOR MASS SPECTRAL IDENTIFICATION

By

James R. Hoyland and Maynard B. Neher

Grant No. R-800921
Project 16ADN 29
Program Element 1B1027

Project Officer

Dr. John McGuire
Chromatography & Mass Spectrometry Section
Southeast Environmental Research Laboratory
Athens, Georgia 30601

Prepared for
OFFICE OF RESEARCH AND DEVELOPMENT
U. S. ENVIRONMENTAL PROTECTION AGENCY
WASHINGTON, D. C. 20460

For sale by the Superintendent of Documents, U. S. Government Printing Office, Washington, D.C. 20402 - Price $1.10

ABSTRACT

Computer programs have been developed for a CDC 6400 and for a PDP8/e or PDP8/m for remote identification of mass spectra. Mass spectra generated on a Finnigan 1015 quadrupole spectrometer under control of the PDP8/e or PDP8/m computer are reduced to abbreviated form and transmitted over commercial telephone lines to the CDC 6400 for identification. Data may be transmitted by paper tape or by direct transmission through a Digital Equipment Corporation serial line interface and an acoustical coupler.

This report was submitted in fulfillment of Project Number 16ADN 29 Grant Number R-800921, by Battelle Memorial Institute under the Sponsorship of the Environmental Protection Agency. Work was completed as of September 30, 1973.

CONTENTS

FIGURES

SECTION I

CONCLUSIONS

The algorithm of Biemann et. al.[1] implemented on a CDC 6400 computer system results in an efficient and powerful program for interactive identification of mass spectra. Minimization of dialog and automatic transmission of an entire abbreviated spectrum via paper tape or directly from a mini computer via a serial line interface ensures rapid throughput, in contrast to other programs in current use which require data to be transmitted one peak at a time. Computation of a similarity index further enhances the value of the current system by providing the user with numerical data indicating goodness of fit.

An option allowing users to build specialized libraries of mass spectral data and to match unknown spectra against either these libraries or the main library provides a further element of flexibility not available in other programs.

SECTION II

RECOMMENDATIONS

Utilization of the basic Biemann algorithm results in an unresolved problem which appears not to have been recognized previously; namely, that the correction of computed intensity ratios by dividing each of these by the average ratio for strong peaks leads to similarity indices which are not synmetric on exchange of the known and unknown spectra. The difference can, in fact, be very significant. A mathematical procedure to overcome this difficulty should be investigated and implemented in the current program to eliminate this artifact.

The present system for storing spectra in the master library deliberately incorporates certain inefficiencies to provide maximum computational speed. It is now recommended that this situation be studied in more detail to provide a better compromise between storage and computational costs, particularly if the current library is expanded significantly.

A better record-keeping method is needed to provide an accurate tabulation of charges which are up to date rather than relying on a monthly dump from computer center records.

SECTION III

INTRODUCTION

Identification of water-borne pesticides or other pollutants is a prerequisite to their control and removal. Mass spectrometry has proved to be of significant value for such identification due to the inherent sensitivity of the method and the ease with which a mass spectrometer can be coupled to a gas chromatograph.

A further advantage of low-resolution mass spectral data is that it is highly amenable to computer reduction and manipulation. This has led to widespread utilization of computer matching techniques [1-8] for the identification of an unknown spectrum by comparing such a spectrum with that of known materials contained in a central database and computing the best possible match.

The availability of relatively inexpensive computer-driven instruments is another advantage of mass spectroscopy. These systems are usually designed such that the computer supervises data acquisition and plots or prints the spectrum. An example of such a configuration is the System Industries' System 150. This system consists of a PDP-8/E or PDP-8/M computer equipped with either DECtape, Diablo disk, or both. The PDP-8 is interfaced to a Finnigan Quadrupole Mass Spectrometer. Mass spectra are acquired under computer control and stored on either tape or disk for later analysis. It is therefore advantageous to utilize this feature in sending spectra to a remote source for identification by recovering spectra from tape or disk and either punching these or storing them back on disk or tape in a format which can be readily transmitted.

The research described in this report has been mainly program development along two distinct lines. Firstly, a program has been developed for the CDC 6400 computer system at Battelle's Columbus Laboratories for the remote processing of mass spectral data. Secondly, a great deal of

effort has been made to develop programs for the PDP-8/E computer for recovery of mass spectra, the formatting of these into a form suitable for transmission, and the punching or storing of the formatted data.

SECTION IV

MASS SPECTRAL MATCHING METHOD

The mass spectral matching program is based on the method of Hertz, Hites, and Biemann[1], which compares the unknown abbreviated spectrum with a library of known spectra, the latter also being abbreviated. (Mass spectra are abbreviated by taking the two most intense peaks in each fourteen-mass unit interval, beginning with the interval 6-19). Before such comparison, however, obvious mismatches may be culled by carrying out a presearch. The presearch eliminates known spectra from consideration on the basis of molecular weight, the number of peaks in the abbreviated spectra, (this number must be between Lx and Ux, where x is the number of peaks in the unknown abbreviated spectrum, L <1 and U >1) and on the basis of Biemann's rectangular array and a corresponding intensity array. The latter quantities are most easily defined as follows:

Let $S_j = \sum_k i\ (j+14k)$, j = 1-14, where i (m) is the intensity of the peak at mass m. The sum over k is such that the entire spectrum is covered. Now arrange the S_j in order of decreasing magnitude, i.e., $S_a \geq S_b \geq S_c \geq S_d$ - - - - -. Then the Biemann rectangular array is simply the set of numbers a,b,c,d - - - -. We consider only the first five such elements, a,b,c,d, and e, in the presearch. These, therefore, represent the rectangular array as utilized in this program. The corresponding intensity array is defined as

$N_j = (S_j \times 100)/ (S_a+S_b+S_c+S_d+S_e)$, j = a,b,c,d,e. Note that $\Sigma N_j = 100$.

Known spectra are eliminated in the presearch on the basis of the rectangular and intensity arrays by computing the following sum:

$$A = \underset{j \text{ paired}}{\Sigma} (N_j^u - N_j^k) + \underset{j \text{ unpaired}}{\Sigma} N_j^u + \underset{j \text{ unpaired}}{\Sigma} N_j^k$$

where superscripts u and k refer to known and unknown spectra. The sum A is bound below by 0 (a perfect match) and above by 200 (a complete mismatch in which there are no common N_j). If A is greater than a specified threshold, R, the known spectrum is eliminated as a possible match in the presearch. One further criterion is used to eliminate known spectra in the presearch. This involves the ratio of the intensity of a "search peak" in the known spectrum to the intensity of that peak in the unknown. This "search peak" is usually the 100 percent peak, unless this peak occurs at m/e = 41, 43, 55, 57, 91, or 105. In the latter situation, the "search peak" is taken to be the next most intense peak, provided its intensity is greater than 50 percent and that it is not another of the six peaks listed above. If these conditions are not satisfied, the 100 percent peak is taken as the search peak. For example, if the most intense peak in the known occurs at m/e = 91, and the second-most intense peak occurs at m/e = 77 with intensity 60 percent, then 77 is chosen as the search peak, with intensity 60 percent. If, however, the 100 percent peak occurs at m/e = 57, and the second-most intense peak at m/e = 43 with intensity 80 percent, then the search peak is taken to be 57 with intensity 100 percent. The criterion for eliminating known spectra is that the "search peak" must be present in the unknown spectra with an intensity at least 25 percent of that found in the known unless the mass of the search peak is greater than 350, in which case 12.5% is used rather than 25%.

A similarity index can then be computed for those known compounds whose mass spectra pass through the initial presearch screen. In describing this calculation, it is convenient to imagine the abbreviated spectra as being a series of two mass-intensity pairs occupying a "slot". The slot is further characterized by a number, such that slot 1 contains masses between 6 and 19, slot 2 contains masses between 20 and 33, and in general slot n contains masses between 14n-8 and 14n+5.

Spectral comparison is begun with the lowest commonly occupied slot. Let $m_j^u(1)$, $m_j^u(2)$, $m_j^k(1)$, and $m_j^k(2)$ be the masses of unknown (u) and known (k) peaks in the j^{th} slot, with corresponding intensities $i_j^u(1)$, $i_j^u(2)$, $i_j^k(1)$, and $i_j^k(2)$. Then the following situations can arise:

Case 1. $m_j^k(1) = m_j^u(\alpha)$ and $m_j^k(2) = m_j^u(\beta)$, where $\alpha, \beta = 1$ or 2. In this case set $r_j(1) = i_j^k(1)/i_j^u(\alpha)$, and $r_j(2) = i_j^k(2)/i_j^u(\beta)$. Further, let $I(1) = \text{Max}\ \left[i_j^k(1), i_j^u(\alpha) \right]$, with a similar definition of $I(2)$. Then set $f_j(1) = 12$ if $I(1) \geq 10$, $f_j(1) = 4$ if $2 \leq I(1) < 10$, and $f_j(1) = 1$ if $I(1) < 2$. The parameter $f_j(2)$ is similar-defined.

Case 2. $m_j^k(\alpha) = m_j^u(\beta); m_j^k(\beta) \neq m_j^u(\alpha)$. There is one matched pair in this case. Thus $r_j(\alpha) = i_j^k(\alpha) / i_j^u(\beta)$ and compute $f_j(\alpha)$ as above. Set $r_j(\beta) = 0$ and compute $f_j(\beta)$ using $I(\beta) = \text{Max}\ \left[i_j^k(\beta), i_j^u(\alpha) \right]$. Also note that $i_j^k(\beta)$ and $i_j^u(\alpha)$ are unmatched intensities.

Case 3. $m_j^k(1,2) \neq m_j^u(1,2)$. All intensities are unmatched, and both $r_j(1)$ and $r_j(2)$ are set to 0. $I(1)$ and $I(2)$ are determined by the two largest intensities chosen from $i_j^k(1)$, $i_j^k(2)$, $i_j^u(1)$ and $i_j^u(2)$. Note that all four intensities will be placed in the unmatched intensity group.

Special cases arise when there is only one peak of a known or unknown spectrum in a slot or a slot is vacuous. A missing mass in a slot is treated as a mass of 0 with intensity 0. Then all cases above go through except for a 0 - 0 match assigned r = 0 and f = 0, and is therefore ignored.

The average ratio of large peaks ($f_j(\alpha) = 12$, $r_j(\alpha) \neq 0$) is next computed and all ratios $r_j(\alpha)$ are redefined by dividing each by this average. All ratios greater than unity are then inverted. Finally, let $U = [\Sigma_{\text{unmatched}} i_j^k(\alpha) + \Sigma_{\text{unmatched All } j,\alpha} i_j^u(\alpha)] / \Sigma_{\text{matched}} [i_j^k(\alpha) + i_j^u(\alpha)]$, which is the fraction of unmatched

intensity. The similarity index is then defined as

$$S = [\Sigma_j \Sigma_\alpha r_j(\alpha)\ f_j(\alpha)] \ / \ [\Sigma_j \Sigma_\alpha f_j(\alpha)\ (1 + U)]$$

The problem mentioned in an earlier section concerning the non-symmetry of the similarity index with respect to exchange is easily recognized on the basis of the preceeding mathematical development. If the known and unknown spectra are interchanged, then all ratios, r, are inverted. Now, however, the average ratio for strong peaks, defined in terms of the original ratios, is $N^{-1} \sum_{\substack{j,\alpha \\ f_j(\alpha)=12, r_j(\alpha)\neq 0}} r_j(\alpha)^{-1} \neq N^{-1} / \sum_{\substack{j,\alpha \\ f_j(\alpha)=12, r_j(\alpha)\neq 0}} r_j(\alpha)$,

where N is the number of non-zero r values with f = 12. The best procedure to correct this recently recognized problem is to take the average ratio of large peaks as $\frac{N^{-1}}{2}[\sum_{\substack{j\ \alpha \\ f_j(\alpha)=12 \\ r_j(\alpha)\neq 0}} r_j(\alpha) + (\sum_{\substack{j\ \alpha \\ f_j(\alpha)=12 \\ r_j(\alpha)\neq 0}} r_j(\alpha))^{-1}]^{-1}$, which preserves symmetry.

SECTION V

PROGRAM DESCRIPTION

This section is intended to give a fairly extensive description of the current mass spectral matching program, together with schematic charts indicating the major logic flow. All routines other than drivers and those concerned with user input-output are written in COMPASS, the assembly language for the CDC 6400. Extensive utilization of COMPASS is necessary to take full advantage of the unique register structure of the CDC 6400, thus giving a program which is essentially optimal in so far as execution speed and file manipulation efficiency is concerned. Very little could be gained by further assembly language coding of those sections now written in FORTRAN.

The program itself is divided into three major sections, or overlays. The main overlay consists of a driver, subroutines which are common to the other overlays, and most of the file manipulation coding. The latter routines are of no real concern other than to point out their existence. They are written in assembly language and consist mostly of macro instructions which pass parameters to peripheral processors. These latter computers then supervise all further I/O operations.

The other two overlays are concerned with matching input spectra against either the main library or against a user's library. The user library overlay also contains routines for construction and maintenance of user files. Only one of these two overlays resides in core at any given time due to core restrictions. The main overlay contains logic to read these other overlays into core and transfer control to them.

Figure 1 illustrates the major steps and branches in the main overlay driver. After the user has logged in, control passes to this routine, which requests the user's laboratory number and name, attaches files, writes a header in user's record file, and queries the user as to whether he wishes to use the main library or his own library of known

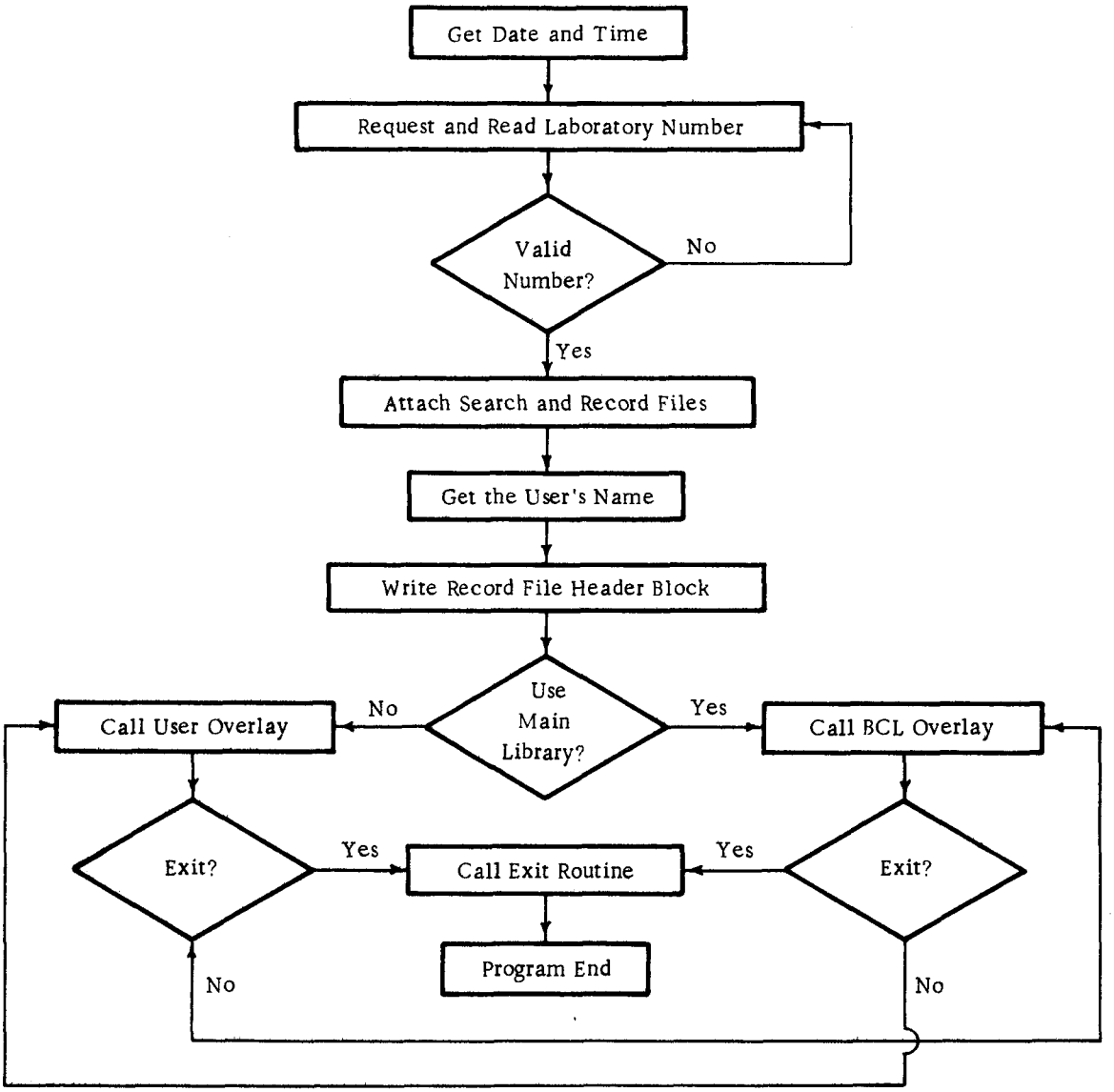

FIGURE 1. FLOW CHART FOR MAIN OVERLAY DRIVER

spectra. The appropriate overlay (designated BCL for the main library or USER for the user library) is read from a disk and control is transferred to the driver routine for that overlay. Upon return from either overlay, a test is made for the exit flag set. If this flag is not set, the other overlay is called into core and control is transferred to it. Otherwise, the exit routine, discussed later, is entered and the run is terminated.

Figures 2 and 3 are the flow diagram for the main library, or BCL, overlay driver. The presearch and main library files are positioned, and a buffered read request for the first presearch record is sent. This record is read while other operations are taking place. The input subroutine is called, and upon return a test is made to check if a spectrum has been entered. If not, control is transferred to the main overlay driver. The input spectrum is abbreviated and saved for possible future use, and presearch data is computed. The user is then requested to enter presearch parameters that determine the stringency of the presearch itself. The presearch subroutine is then entered, and a file key corresponding to the main library record is stored for each known spectrum passing the presearch screen. If no spectra passed the presearch screen, the message "NO HITS" is printed, the record file is updated, files are repositioned, and the input routine entered once more. The main search routine is entered for each possible presearch hit. Further screening occurs here (to be discussed later), and if passed, a similarity index is computed. Data for the best 20 hits are stored, unless fewer than 20 similarity indices were computed. If no similarity indices were computed, the "NO HITS" message is printed with further flow as diagrammed in Figure 3. The results are then printed five at a time. The user may halt output after any group of 5 has been printed, resulting in the run record being written, and a return to the beginning of the routine.

The input routine for the BCL overlay is diagrammed in Figures 4 and 5. The number of peaks in the input spectrum is set to 0 and the user is asked to select an option. Available options are search (S), Print a

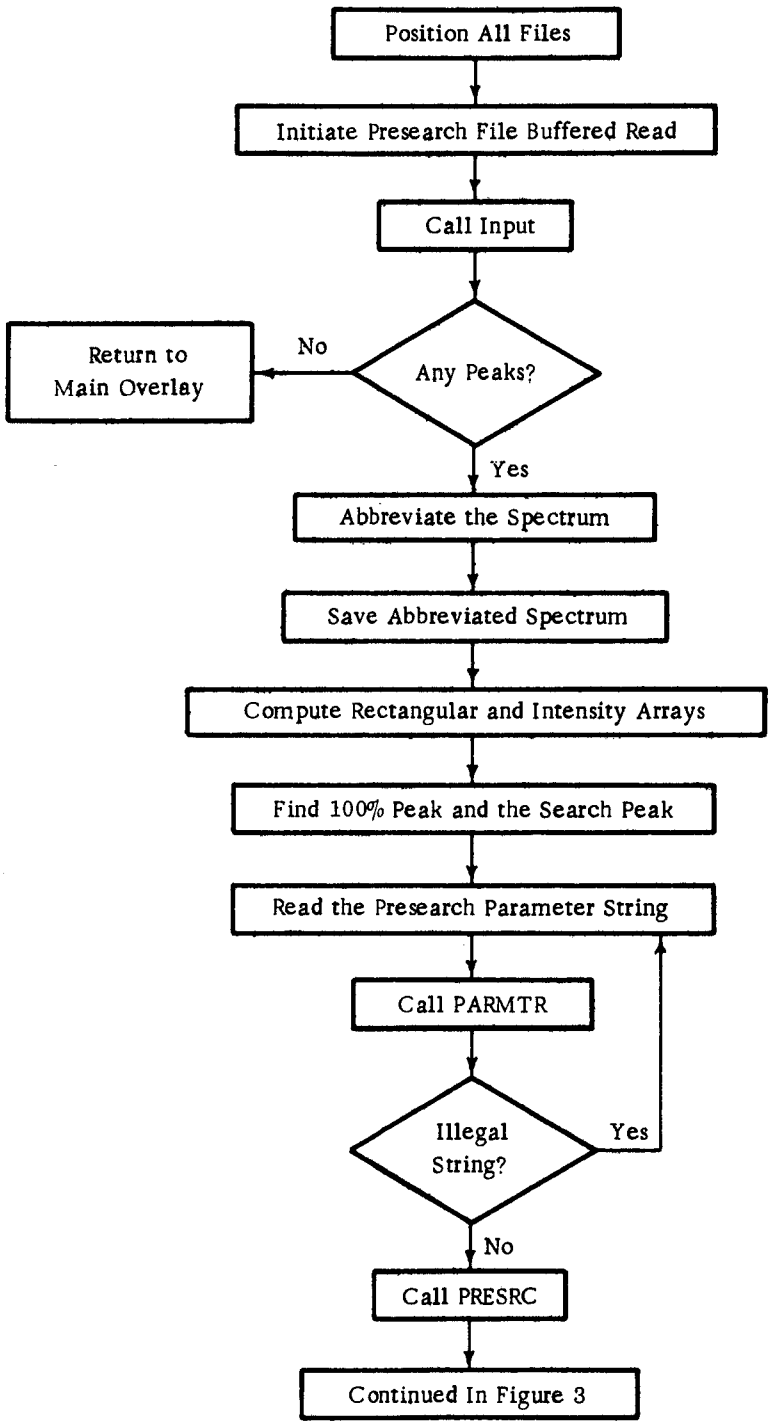

FIGURE 2. FLOW CHART FOR BCL OVERLAY DRIVER-I

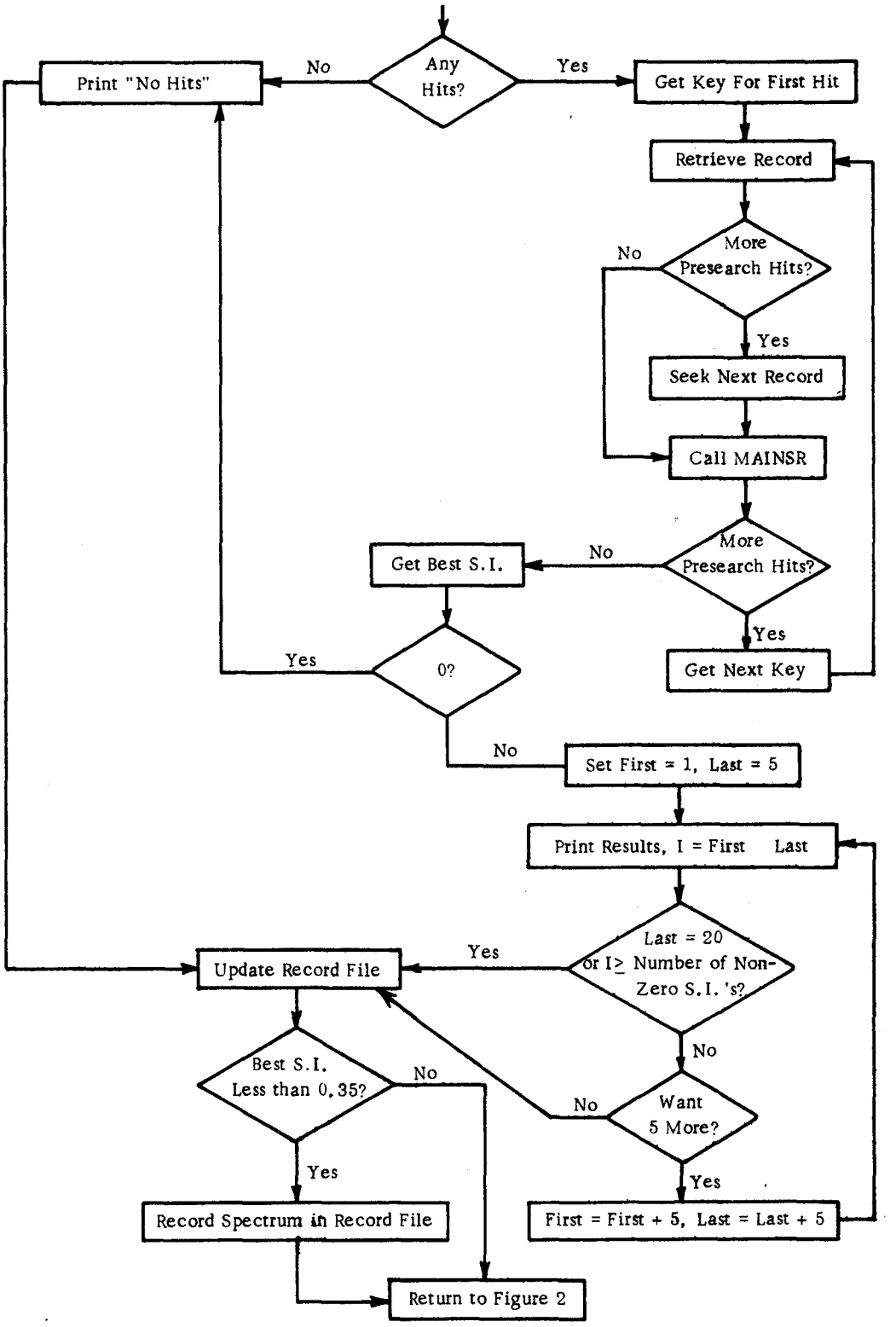

FIGURE 3. FLOW CHART FOR BCL OVERLAY DRIVER-II

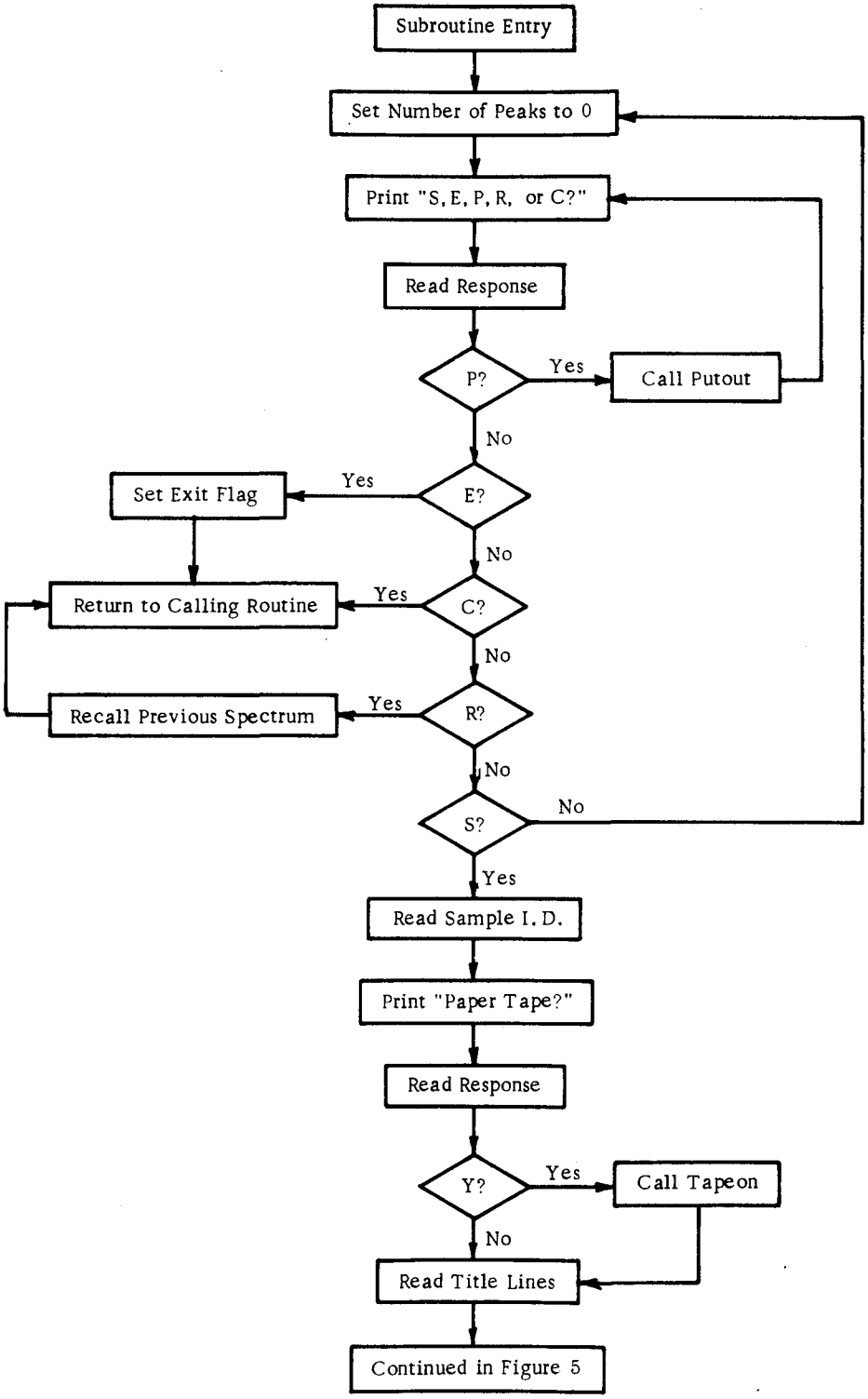

FIGURE 4. FLOW CHART FOR BCL OVERLAY INPUT-I

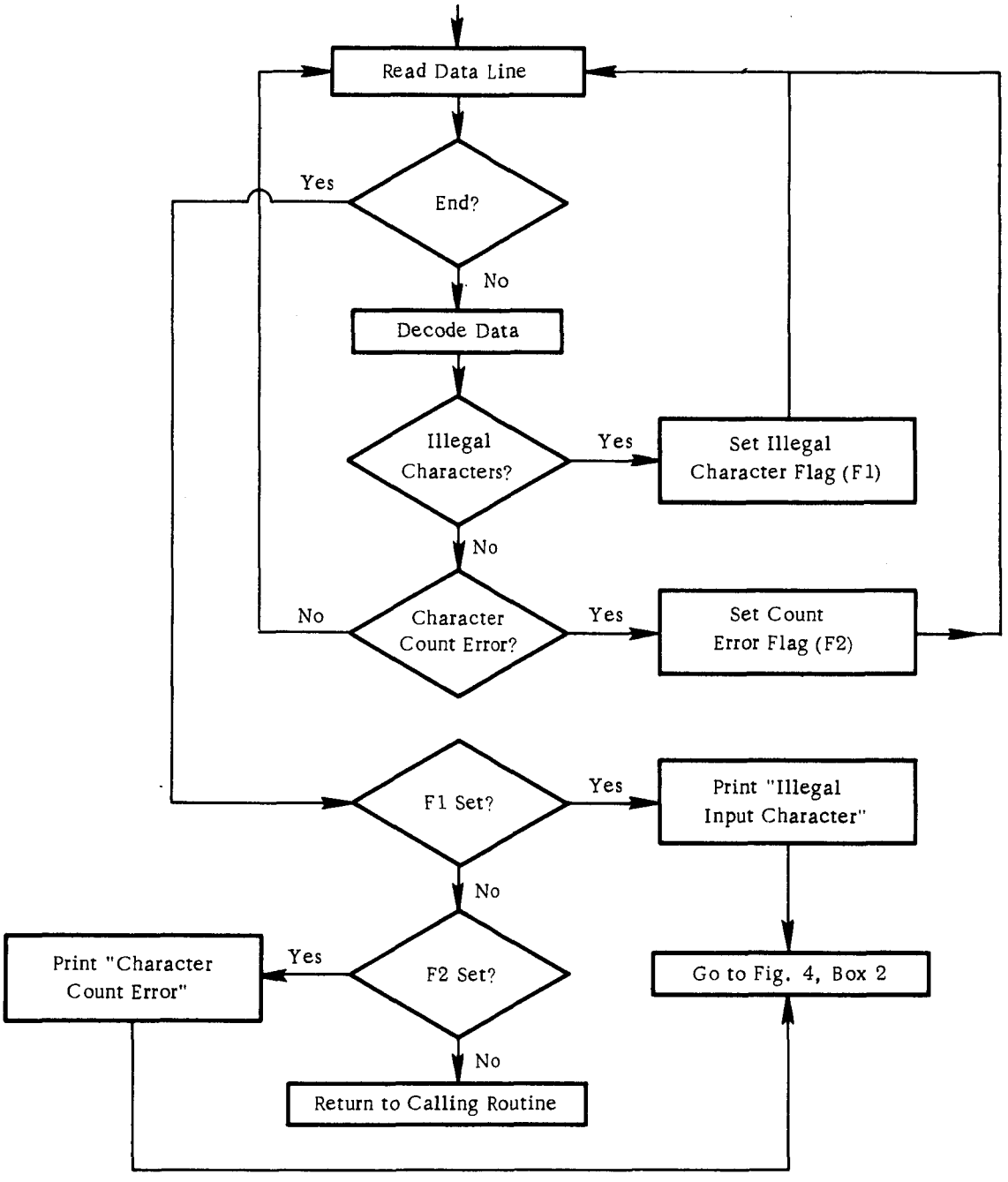

FIGURE 5. FLOW CHART FOR BCL OVERLAY INPUT-II

file spectrum (P), repeat the previously transmitted spectrum (R), change overlays (C) or exit (E). Selection of P transfers control to the print subroutine PUTOUT, and return to the option query after printing the desired spectrum.

If the user types E, the exit flag is set and control is returned to the BCL driver. The "change overlays" character, C, also transfers control to this driver, but without the exit flag set. The repeat option (R) causes the previously transmitted spectrum to be placed in the appropriate core locations, the peak count to be properly set, and control transferred to the BCL driver. The S option results in a query as to whether paper tape is to be used for input. If so, the subroutine TAPEON is called, which sets a peripheral processor paper tape bit and turns on the user's paper tape reader if it is equipped with the automatic tape on-tape off feature. Two lines of title information are read, followed by the spectrum. The input data is scanned for proper format, illegal characters, or character count errors resulting from dropped bits. The peak count is continually updated. If no errors were detected upon receipt of the three characters END, transfer is made to the driver. Otherwise appropriate error messages are printed and control is passed to the beginning of the input routine.

The subroutine, PUTOUT, sketched in Figure 6, is extremely simple. The user is requested to enter the file key for the desired library spectrum. The corresponding file record is retrieved and unpacked, the spectrum is printed, and control is returned to the input routine.

Figure 7 is a flow chart of the subroutine PARMTR which decodes an input parameter string. Up to four parameters may be entered by the user, although only M, a molecular weight range, is mandatory. Default values for others are set before entering PARMTR. The first parameter entered must be M, and thereafter any order may be used.

PARMTR initially tests that the first character is indeed M, and if so decodes the molecular weight range. Unrecognized characters, a second range value smaller than the first, or the first character not being M result in the error flag being set.

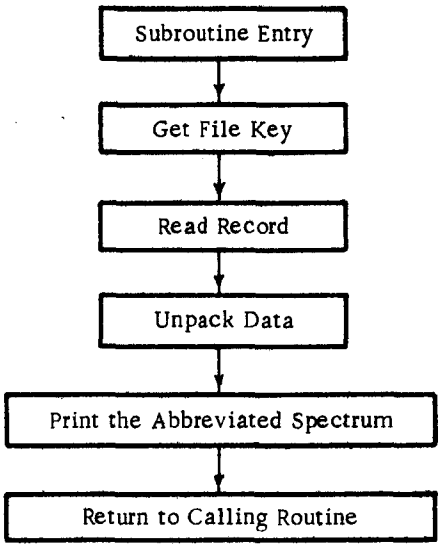

FIGURE 6. FLOW CHART FOR BCL OVERLAY ROUTINE PUTOUT

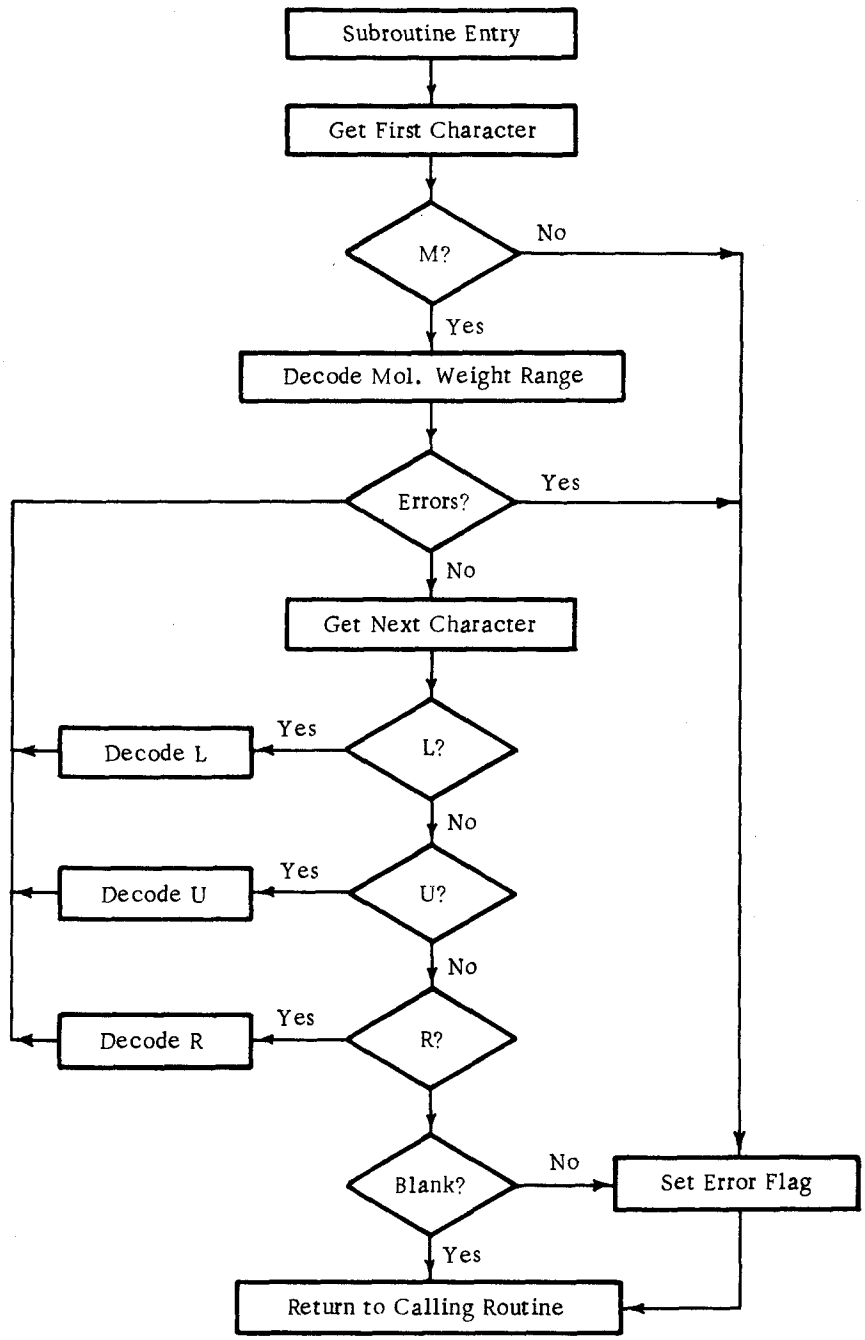

FIGURE 7. FLOW CHART FOR BCL OVERLAY ROUTINE PARMTR

The subroutine now loops on the remaining parameters, with a blank indicating termination of the string. The parameters L, U, and R are defined in Section IV. PARMTR exhaustively tests for illegal characters and out of range values. The current limits are $0.2 \leq L \leq 0.9$, $1.2 \leq U \leq 5.0$, and $40 \leq R \leq 140$, with default values of $L = 0.5$, $U = 2.5$, and $R = 120$. The value of R should probably be decreased to 100 as users become more familiar with the system.

A schematic flow diagram of the presearch routine, PRESRC, is shown in Figure 8. PRESRC utilizes a presearch file which contains two words of information for each known in the data base. This file is written in sequential binary format with 512 words per record. Information contained includes molecular weight, search peak, search peak intensity, number of peaks in the abbreviated spectrum, and the rectangular and intensity arrays. This information is readily packed into two words since the word length of the CDC 6400 is 60 bits.

The number of hits is set to zero and a counter is set to 1 upon entering PRESRC. A call to buffer in the next record is then made, and this record is read while the data from the first record is being analyzed. Thereafter the buffer in routine always performs the task of reading the next record while the previous record is being used to screen the known spectra.

PRESRC first tests for the molecular weight to be in range, since this test is done the most rapidly of any carried out. The test on the number of peaks in range is then done, followed by a test for the presence of the search peak with the proper intensity. If all tests are passed, the rectangular and intensity arrays are tested by the method described in Section IV. The Ith known is recorded as a presearch hit only if all tests are passed.

The coding in PRESRC has been very carefully arranged in order to minimize computation time. The central processor (CP) time required for a presearch naturally depends on the presearch parameters, but it is

found that on the average about 11,000 knowns can be screened in about 1 second of CP time. It is doubtful that this can be improved upon in any significant way.

The logic contained in the user overlay driver is shown in Figures 9 and 10. Note that a user may only attach a library during the first pass through this overlay driver. Thereafter, the file open condition is set and immediate transfer is made to the section of the driver concerned with mass spectral matching.

If no file is open, a master file which receives all new library spectra, regardless of origin, is attached and opened. The user is queried as to whether he wishes to create a new library. If so, a new file is opened and transfer is made to the subroutine NWSPEC which is concerned with entering new spectra into a library and the master file. If the user wishes to use an old library, he must enter the library name and his account number. An attempt is then made to attach this file. An attach error will occur if the library name or account number has been entered incorrectly, otherwise the library will be successfully attached. The file is positioned at the end of information (EOI) and the user is queried as to whether he wishes to add spectra to the file. If so, control is passed to NWSPEC.

Upon return from NWSPEC, or if no new spectra are to be appended to an old file, the input routine is called, and if no spectrum is entered, control is returned to the main overlay driver. The program procedes as in the BCL, or main library, case, except that no presearch is carried out since user libraries of necessity will probably be rather small. The only tests are for the presence of the 100% peak in either spectrum being present with proper intensity in the other. This is automatically done in the main search routine, MAINSR. The same main search routine is used for both overlays, and is described later. After all similarity indices are computed, the results are printed as in the BCL overlay, the record file is updated, the library rewound, and the input routine

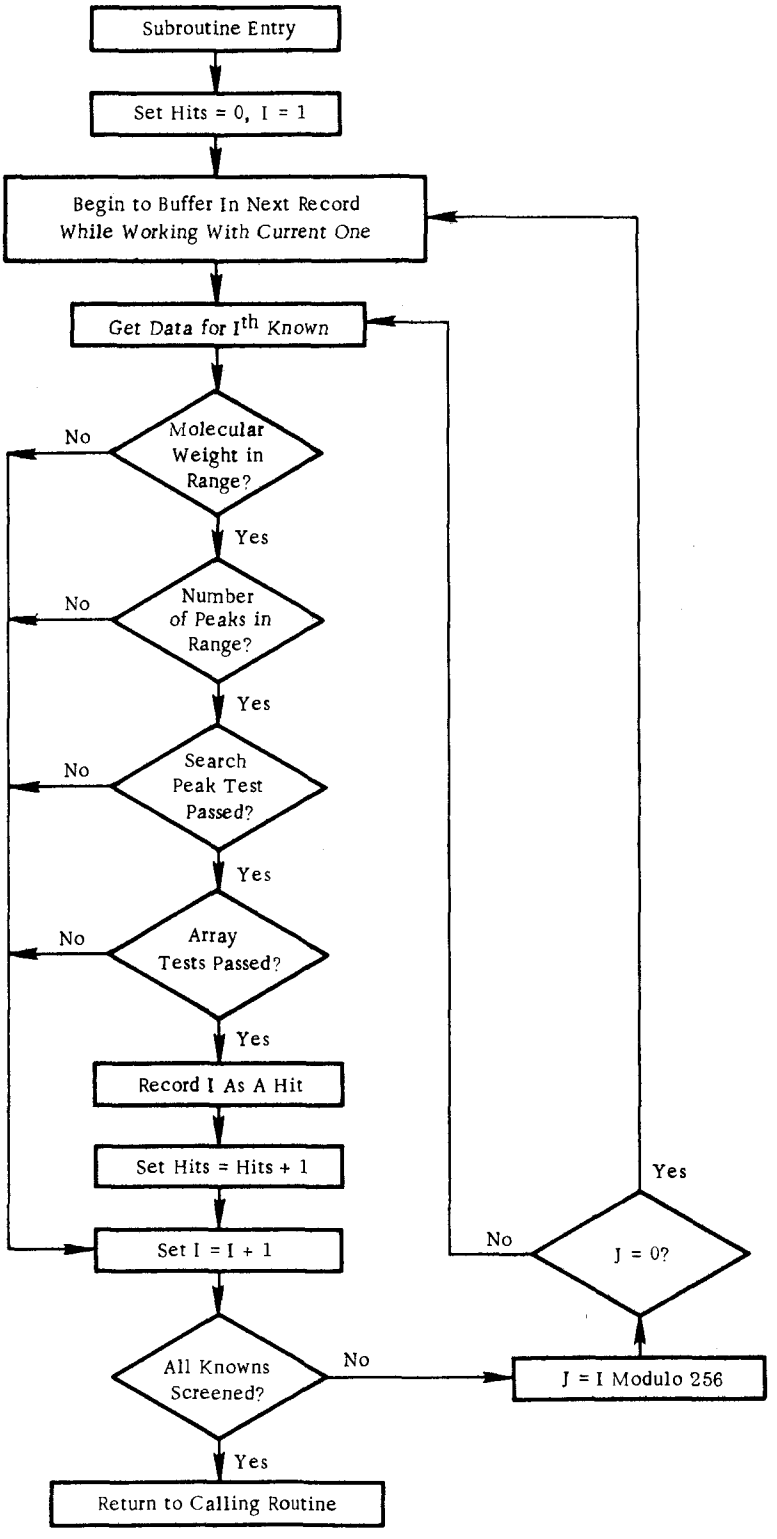

FIGURE 8. FLOW CHART FOR BCL OVERLAY ROUTINE PRESRC

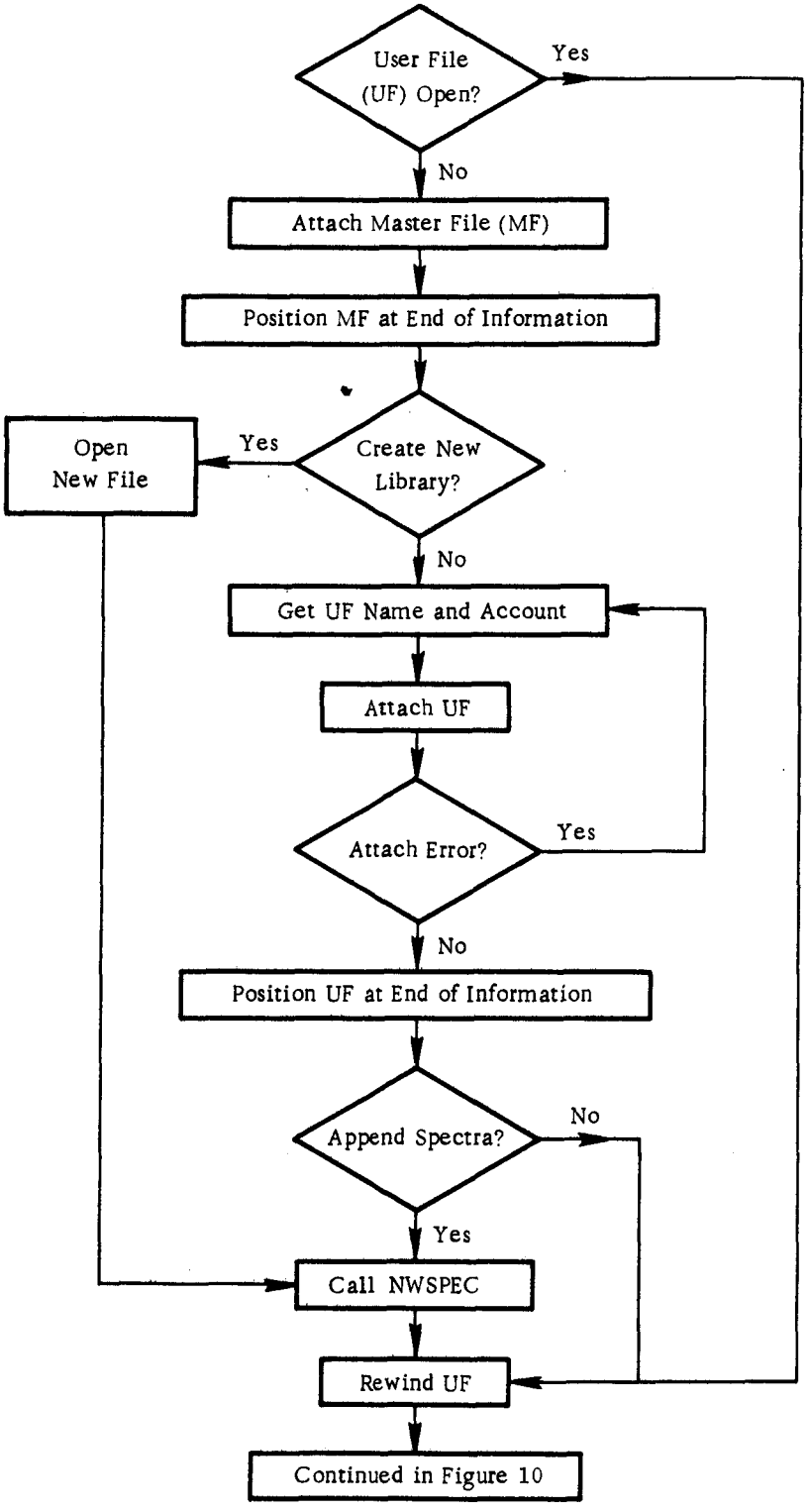

FIGURE 9. FLOW CHART FOR USER OVERLAY DRIVER-I

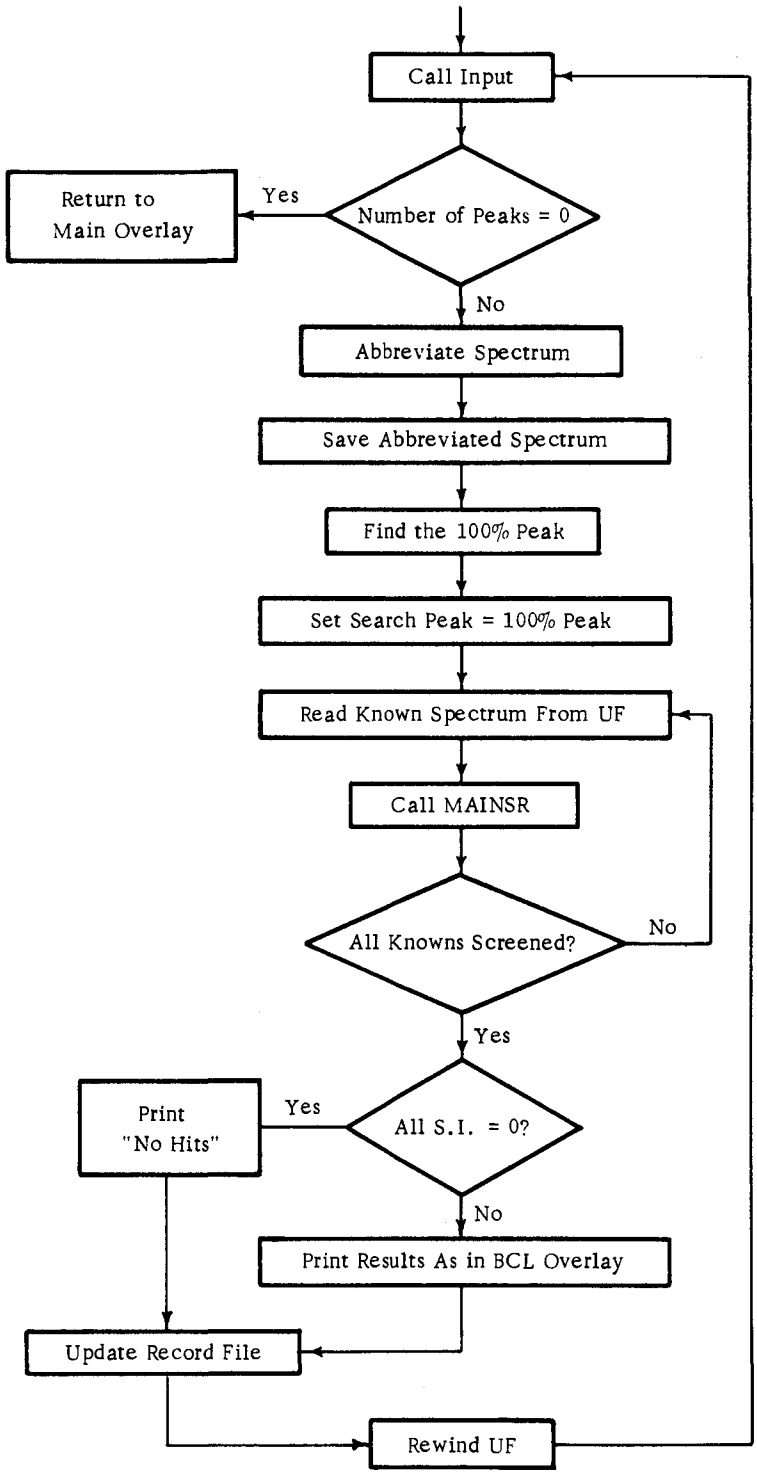

FIGURE 10. FLOW CHART FOR USER OVERLAY DRIVER-II

is called. Note that this latter routine is identical to the one used in the main library overlay except that the P option is not enabled. The subroutine NWSPEC is diagrammed in Figure 11.. The user has a choice of two commands, S (send a spectrum) or E (return to user overlay driver). If the S option is selected the user is requested to enter the compound name and a source identification, the Wiswesser line notation, the molecular weight, and the molecular formula. Spectral data is then sent either via paper tape or direct transmission. The spectrum is written in the master file, then abbreviated and written into the user's library, and control is transfered to the beginning of the routine.

Figures 12 and 13 contain the flow chart for the main search routine. After the known spectrum is read into core, a test is made to ascertain that the known search peak is present in the unknown spectra with sufficient intensity. If so, the 100% peaks of both spectra are tested to see if they are present with adequate intensity in the other. This test is automatically bypassed for 100% peaks with mass less than 33. If the above tests are passed, the intensity ratios, factors, and the sum of unmatched intensity are computed, followed by the computation of the total intensity in both spectra. The average ratio of large peaks is computed, the ratios are corrected by dividing by this average, ratios greater than 1 are inverted and the overall weighted average ratio, A, is found. The similarity index S is then computed and if S is less than 0.1 it is ignored. If less than 20 similarity indices have been computed, the current result, along with the name and key are stored. If 20 indices are already stored, the smallest stored value is found and compared with the current index. If the current value is smaller, it is ignored. If not, the lowest stored value is replaced by the current result, as is the name and key.

The final flow chart for the mass spectral matching program, Figure 14, is for the exit routine. The final entry is made in the record file, and this file, along with the presearch and main library files are closed

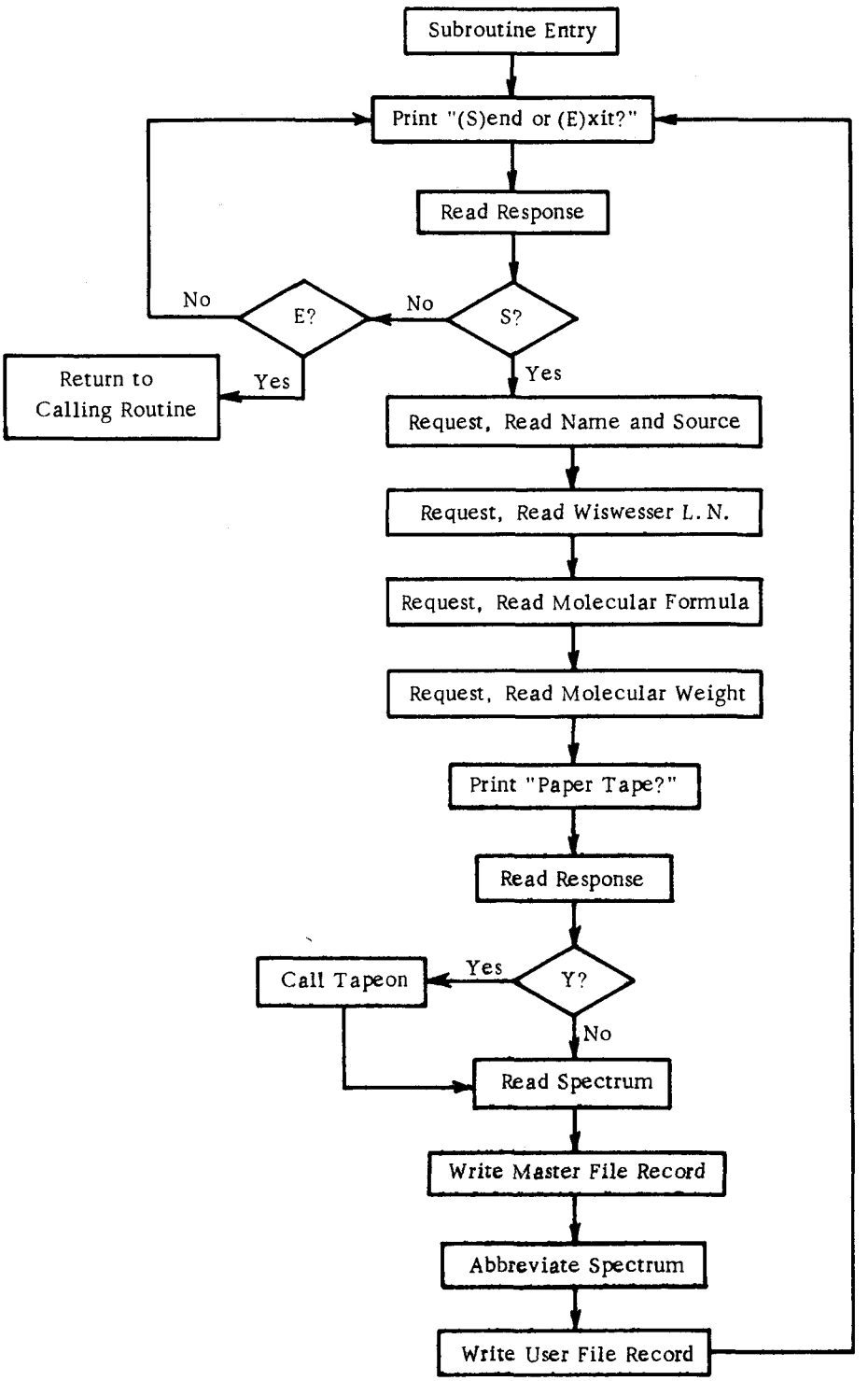

FIGURE 11. FLOW CHART FOR USER OVERLAY ROUTINE NWSPEC

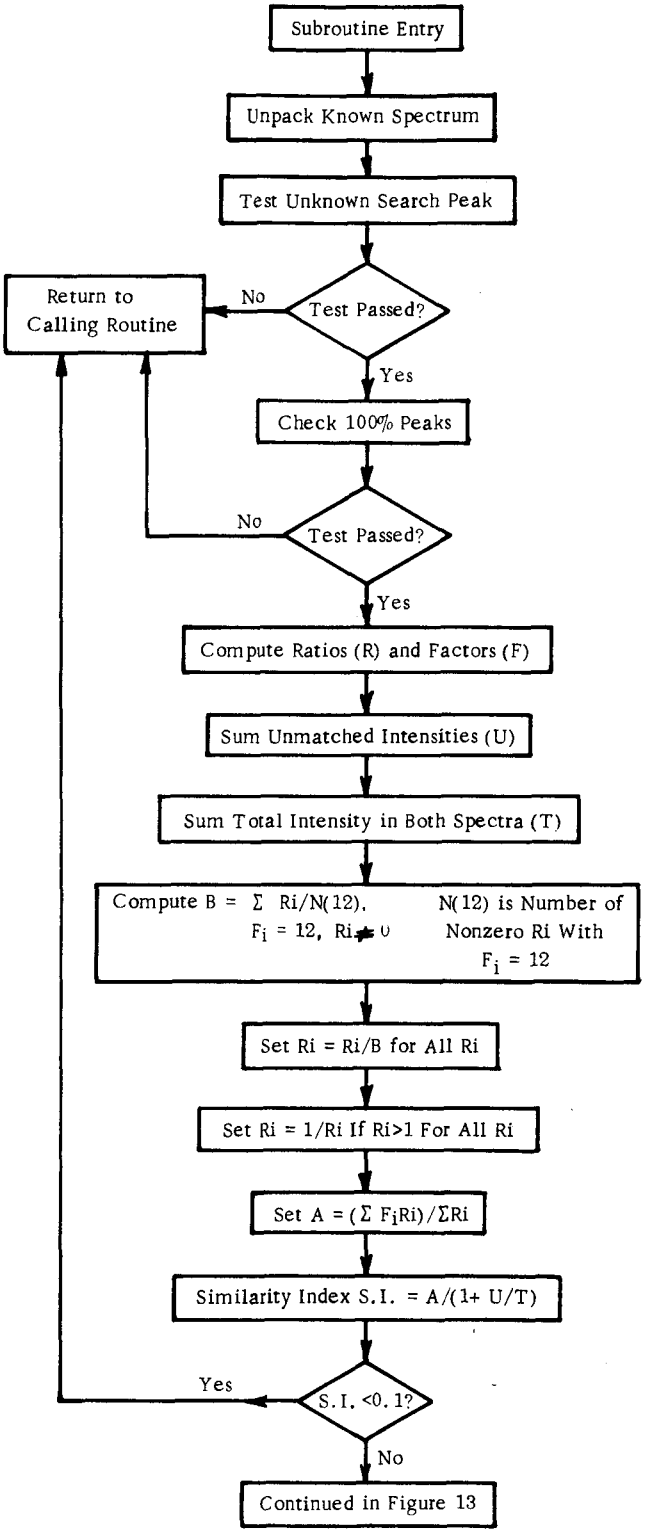

FIGURE 12. FLOW CHART FOR MAIN SEARCH ROUTINE-I

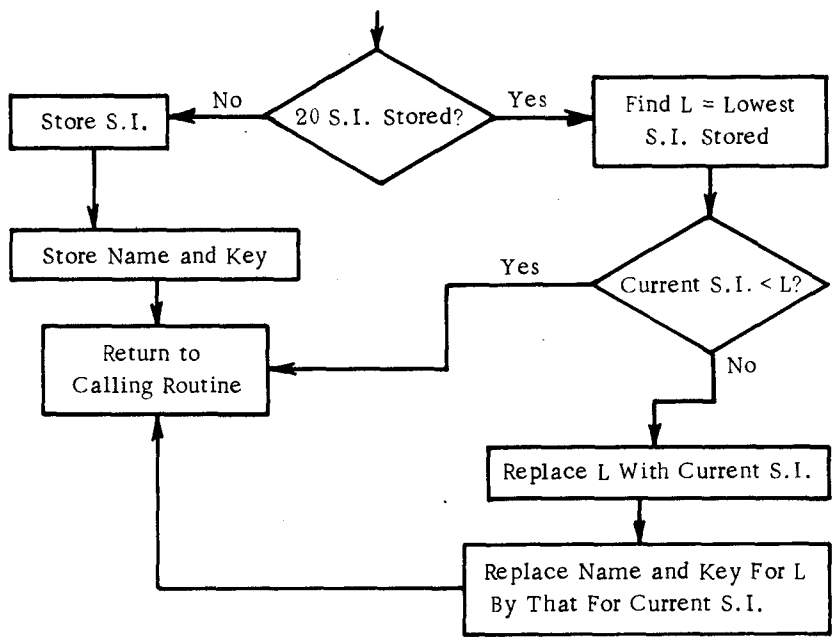

FIGURE 13. FLOW CHART FOR MAIN SEARCH ROUTINE-II

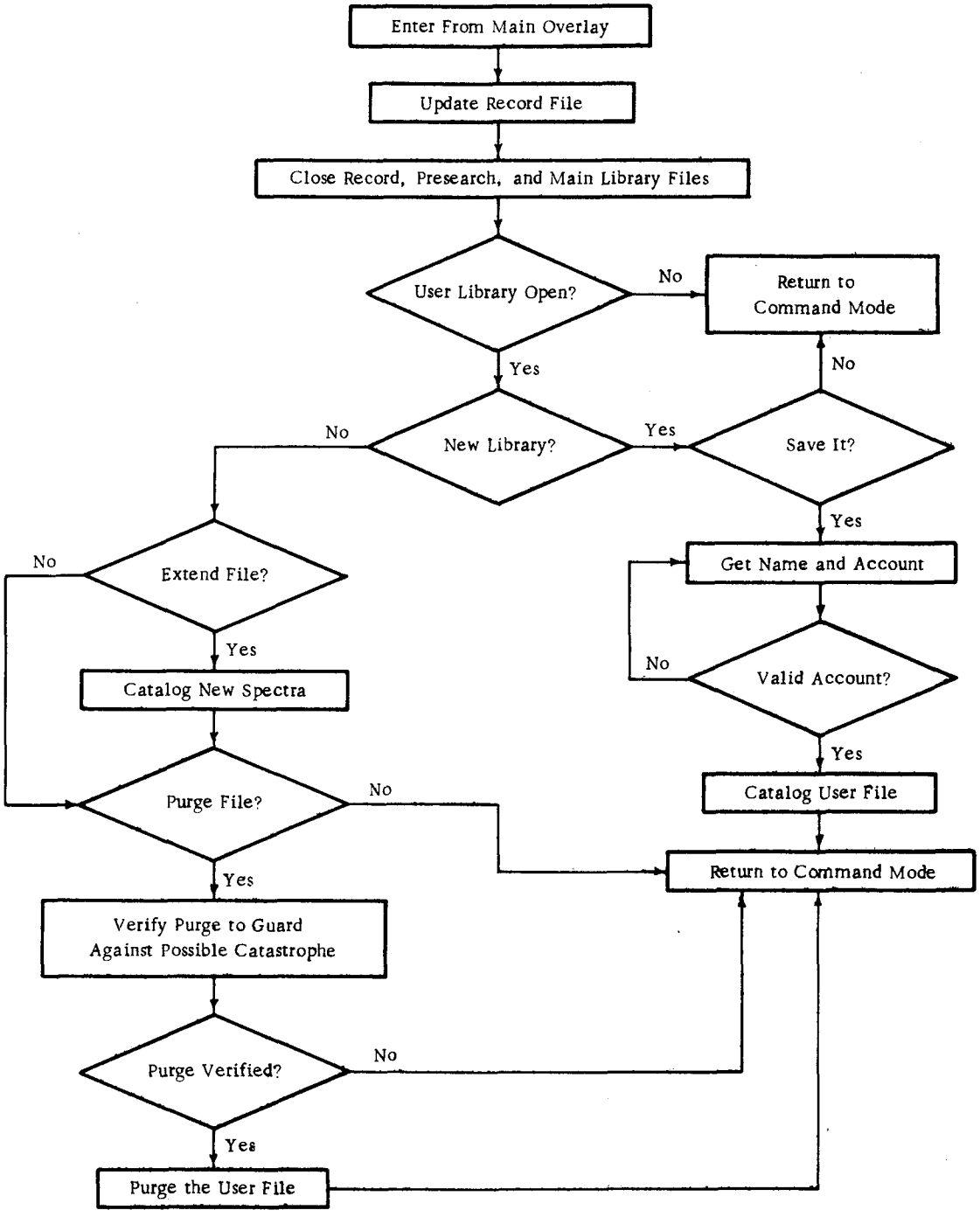

FIGURE 14. FLOW CHART FOR EXIT ROUTINE

and logically disconnected. A test is then made as to whether a user library was connected. If not, a return to the CDC 6400 monitor is made, after which the user may log out of carry out other interactive processing. If a user library has been connected, a flag denoting the creation of a new library is tested. If it is set, the query "SAVE FILE?" is sent. If the library is to be saved, the user enters the name he wishes to give the library and his account number. The latter entry is examined for validity, and if valid, the file is cataloged and control passes to the system monitor.

If an old library is attached, the extension flag is tested. If new spectra have been added, the user is sent the query "EXTEND FILE?" If the reply is affirmative, the new entries are made a permanent part of the library. The query "PURGE FILE?" is sent, and if a negative response is given, the exit to system monitor command mode is taken. If the response is positive, the user must verify the purge request before the operation is carried out to prevent accidental loss of valuable information.

SECTION VI

FILE STRUCTURE

The purpose of this section is to give a detailed account of the various files utilized by the mass spectral matching program and a discussion of their structure.

The presearch file, as noted previously, is a binary sequential file containing two words of information for each known spectrum, and structured such that the record length is 512 words. The order of appearance of compounds in the presearch file determines the file key described later in reference to the main library file. The relation is simply that the file key for the nth presearch entry (nth compound) is n+1.

Information in the two presearch words is not packed so tightly as possible, but rather is packed to provide optimal access for masking operations. The first presearch word contains the following information in the designated bit positions, where bit 59 is the highest order bit and bit 0 the lowest:

Bits 50 - 59	Number of peaks
Bits 40 - 49	Pointer to search peak
Bits 30 - 39	Mass of search peak
Bits 20 - 29	Intensity of search peak
Bits 16 - 19	First rectangular array element
Bits 12 - 15	Second rectangular array element
Bits 8 - 11	Third rectangular array element
Bits 4 - 7	Fourth rectangular array element
Bits 0 - 3	Fifth rectangular array element

The second presearch word contains the intensity array and molecular weight packed as follows:

Bits 50 - 59	First intensity array element
Bits 40 - 49	Second intensity array element

Bits 30 - 39	Third intensity array element
Bits 20 - 29	Fourth intensity array element
Bits 10 - 19	Fifth intensity array element
Bits 0 - 9	Molecular weight

The above arrangement provides easy access to all information through shifting and masking operations. Since only 10- or 4- bit masks are required, the unpacking operation is considerably simplified.

The main library is a random access file prepared and accessed using the Control Data Corporation's Scope Index Sequential (SIS) system. This latter software is an extremely sophisticated package providing optimal random access to records via the specification of a file key. No fixed record length is utilized, but records cannot exceed a maximum value specified at the time the file is built. This value is 60 words (600 characters) for the main library file.

The first record (file key = 1) is the number of spectra in the main library. Thereafter the nth spectra is accessed by a file key of n+1, as described above.

The detailed structure of individual records is relatively simply. To enhance computational speed and eliminate a complicated unpacking scheme, it is assumed that each 14-mass unit interval starting at mass 6 contains two peaks. If only one peak is present in a given interval, or if the interval is vacuous, one or two zero mass values are stored with zero intensity. The terminating interval is determined by the highest mass present in the known spectrum.

The first word of a record contains the number of 14-mass unit intervals in the abbreviated spectrum (bits 50 - 59) and the first five mass values (bits 0 - 49, 10 bits/mass). Each word thereafter contains 6 masses. The last word containing the mass values is zero-filled if it does not contain 6 masses.

The intensities are then packed ten per word, using a six-bit intensity scale of 0-63, such that the 100% peak is assigned an intensity of 63.

Very little loss of information is realized since it is very rare that even two spectra of the same compound obtained on the same instrument will agree consistently to within 1% for each peak. The last word of intensity data is also zero-filled if 10 intensity values are not present.

The identification data is next given (name, molecular weight, formula, etc), the termination being signaled by a semi-colon. This data is written in CDC display code, with 10 characters per word.

Scope Indexed Sequential files are written in a form containing data blocks and index blocks. The current main library file utilizes a double-level index scheme in which there is a capacity of 511 words in the main index, each pointing to a 511 word sub-index block containing the disk start address of individual records. This scheme will thus be satisfactory until the file reaches $(511)^2 - 1 = 261,120$ spectra.

The current library contains about 9000 spectra taken from the Aldermaston collection with most reductant spectra removed, and from data taken from the literature or supplied by the Southeast Environmental Research Laboratory.

User's libraries are not written in SIS format, but rather are sequential binary files with variable record length. This format was chosen since these libraries will be small and will not require random access since no presearch is carried out. The data record format is identical to that used for the main library. The master file which contains all spectra submitted by users to individual libraries is also a sequential binary file with variable record length. The records contain identification data and all peaks and intensities transmitted packed so that each word contains three mass-intensity pairs with 10 bits assigned to each mass or intensity.

A record file is maintained for each laboratory, and is attached by the user's laboratory number. Each time a user logs into the system, the record file for his laboratory is opened and positioned at the end of

information. A header block is written containing the user's name and the date. Thereafter records are written for each search of either the user's or the main library. Data included are the title information, a sample identification, number of presearch bits (main library only), number of final bits, and all data for the best five matches. If the best similarity index computed was less than 0.35, or if no bits were found, the input spectrum is also recorded in the record file. The final record for a run is the approximate connect time entered at the time of exit.

SECTION VII

UTILITY PROGRAMS FOR THE CDC 6400

Several utility programs are required to maintain the files. The first of these is essentially identical to the subroutine NWSPEC discussed in Section V. This interactive routine is available for the purpose of allowing users to send new spectra for inclusion in the master file. This routine opens a new file and writes the spectra and identification data onto a disk for later retrieval.

Another utility is used only in batch mode at Battelle-Columbus and retrieves the data written by the previous program, updates the presearch and main library files, and provides a listing of the spectra and hard copy output on punched cards.

The final utility which is frequently used is one which prints the contents of all record files and re-initializes these. The records are dumped whenever there appears to be sufficient data to warrant doing so, and the printout is sent to the Southeast Environmental Research Laboratory.

SECTION VIII

PROGRAMS FOR THE PDP-8E

Three programs have been written for the PDP-8E or 8M dedicated computers which drive the Finnigan mass spectrometers in EPA Laboratories utilizing the mass spectral matching system. These routines are designed to enhance the utility of this latter system by providing two options for rapid transmission of mass spectral data.

The routine used to abbreviate spectra and write the results in a file suitable for punching or for direct transmission is charted in Figure 15. This program, BRVSPC, is called into execution via the System 150 user routine. A storage file name is requested and a new file is opened. Open errors cause a new request for a file name. A spectrum file name is then requested, and this existing file is opened. A spectrum number is then requested, and it is tested to ascertain that it is in range. If so, a minimum value is read or the system default of 1.587% [100/-(2^6-1)] is utilized if a default is returned. The spectrum is then retrieved and abbreviated, all values less than the minimum being ignored. The abbreviated spectrum is written in ASCII characters with suitable terminating characters to mark the end of line (%), end of record ($), and end of file (?) boundaries.

The query "ANOTHER SPECTRUM?" is sent, and a positive response (Y) results in a request for the spectrum number. If any response other than Y is given to the above query, the question "ANOTHER FILE?" is sent. A positive response results in a request for the file name, whereas any other response is taken to mean program termination. The storage file is closed and catalogued and control is returned to the System 150 Executive.

Figure 16 is a short flow chart of the routine used to punch files created by the abbreviated spectrum routine onto paper tape. This routine is extremely simple. It will accept up to 42 files created by

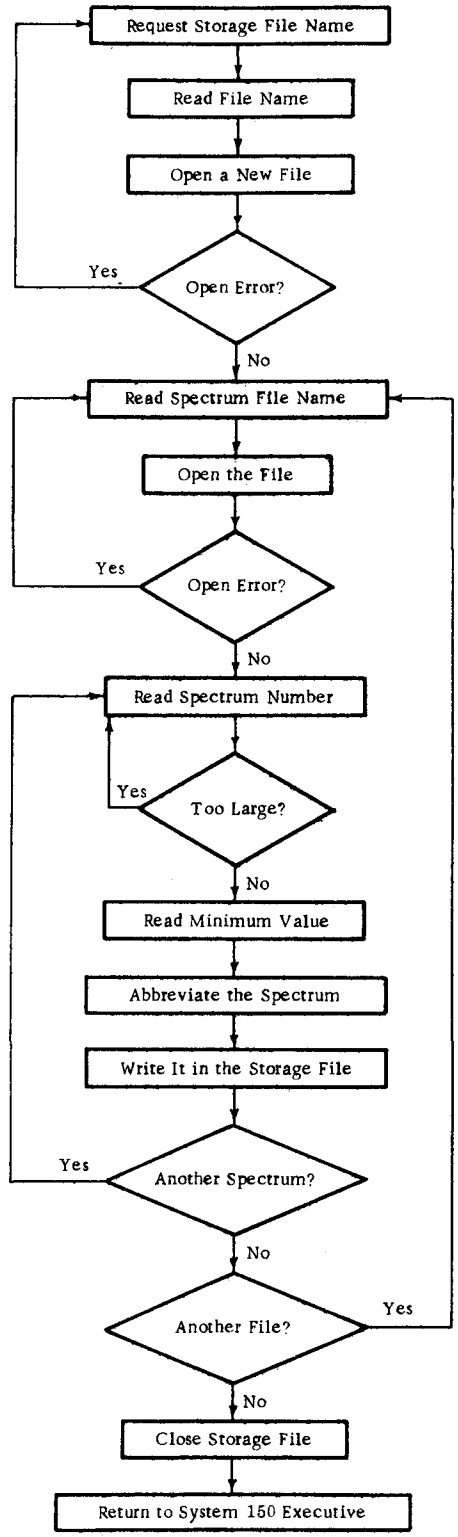

FIGURE 15. FLOW CHART FOR SYSTEM/150 ROUTINE BRVSPC

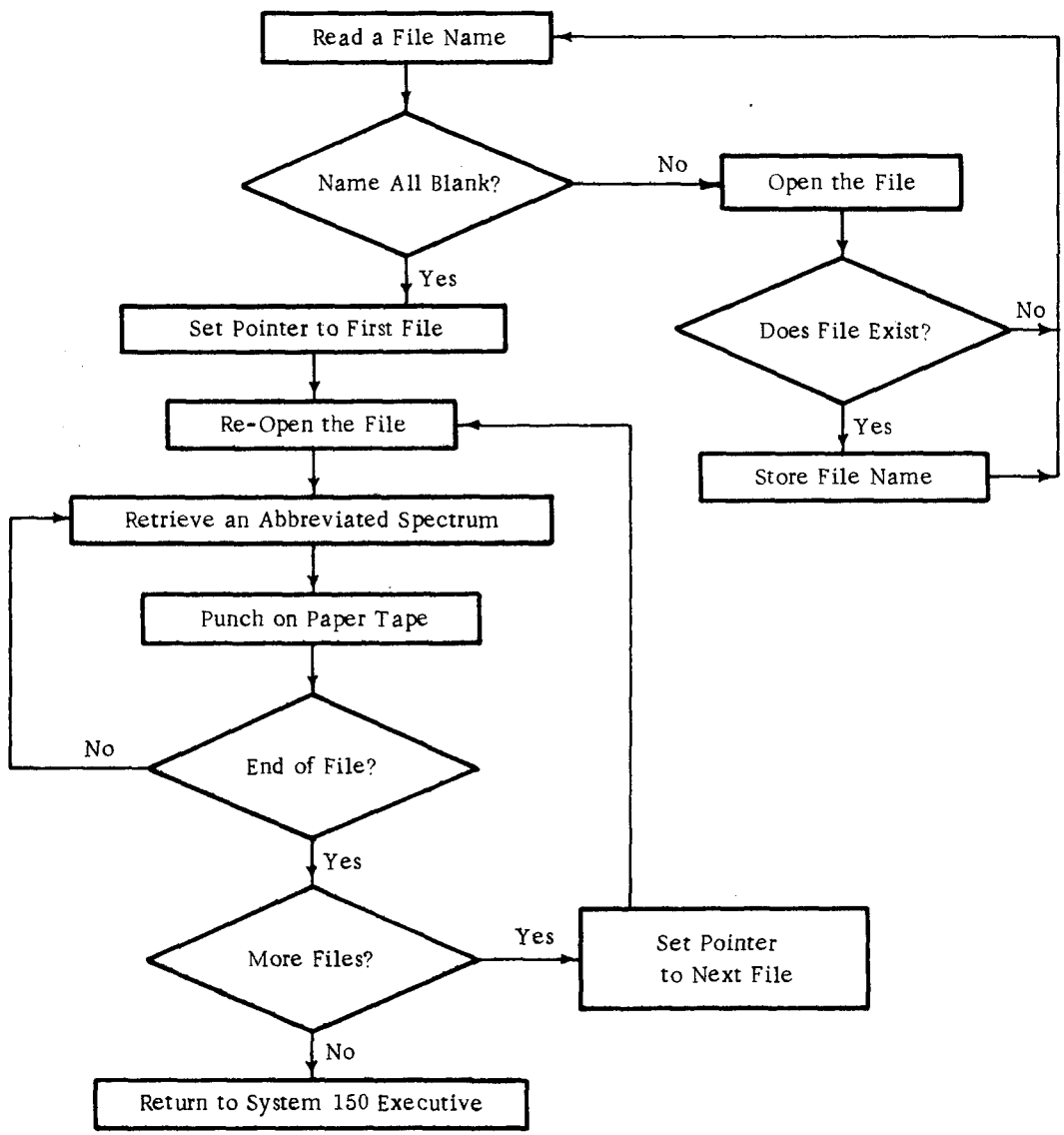

FIGURE 16. FLOW CHART FOR SYSTEM/150 ROUTINE PUNCH

BRVSPC as initial input, testing each file name for the presence of a valid catalogued file. After a blank name is returned, the files are opened, one at a time, and all abbreviated spectra are punched onto tape with proper end of line characters inserted. Spectra and readable headers are separated by strings of rubouts.

An attempt has been made to give a rather complete flow chart of the direct transmission program in Figures 17 and 18, due to the novel nature of this routine. However, the multitudinous branching involved renders this task rather hopeless. Much of the logic has been included and the discussion given below will be helpful in arriving at a better understanding. The present direct transmission program has many unique features. Firstly, the program operates at any combination of user input/output and serial line interface baud rates. In order to achieve this feature, 4000 octal words of core storage have been set aside as a circular buffer. This buffer is filled and emptied at a speed determined by the device baud rates for data transfered from the CDC 6400 to the user for printing. The logic is such that reaching the upper end of the buffer causes the next character to be stored or read from the initial buffer location, thus the term circular buffer. The program is called into core from the System 150 user routine, and requests a file of abbreviated spectra assembled by BRVSPC. Since this is the initial file attached, housekeeping duties are performed, chief among them being alteration of the resident monitor to accomodate interrupts from the serial line interface, initializing circular buffer pointers, and setting a print disabled flag. After completion, the program turns the interrupt on and waits for a flag to be raised. It is at this point that the user should dial and make connection to the 6400.

Interrupt can occur from only four sources, namely, the user's I/O device read or write flags (designated TTY for convenience, although some other equipment such as a 300-baud device or a 9600 baud CRT may be used) or the serial line interface read or write flags (designated KL8). Due to program timing, the KL8 output flag should never be raised under interrupt

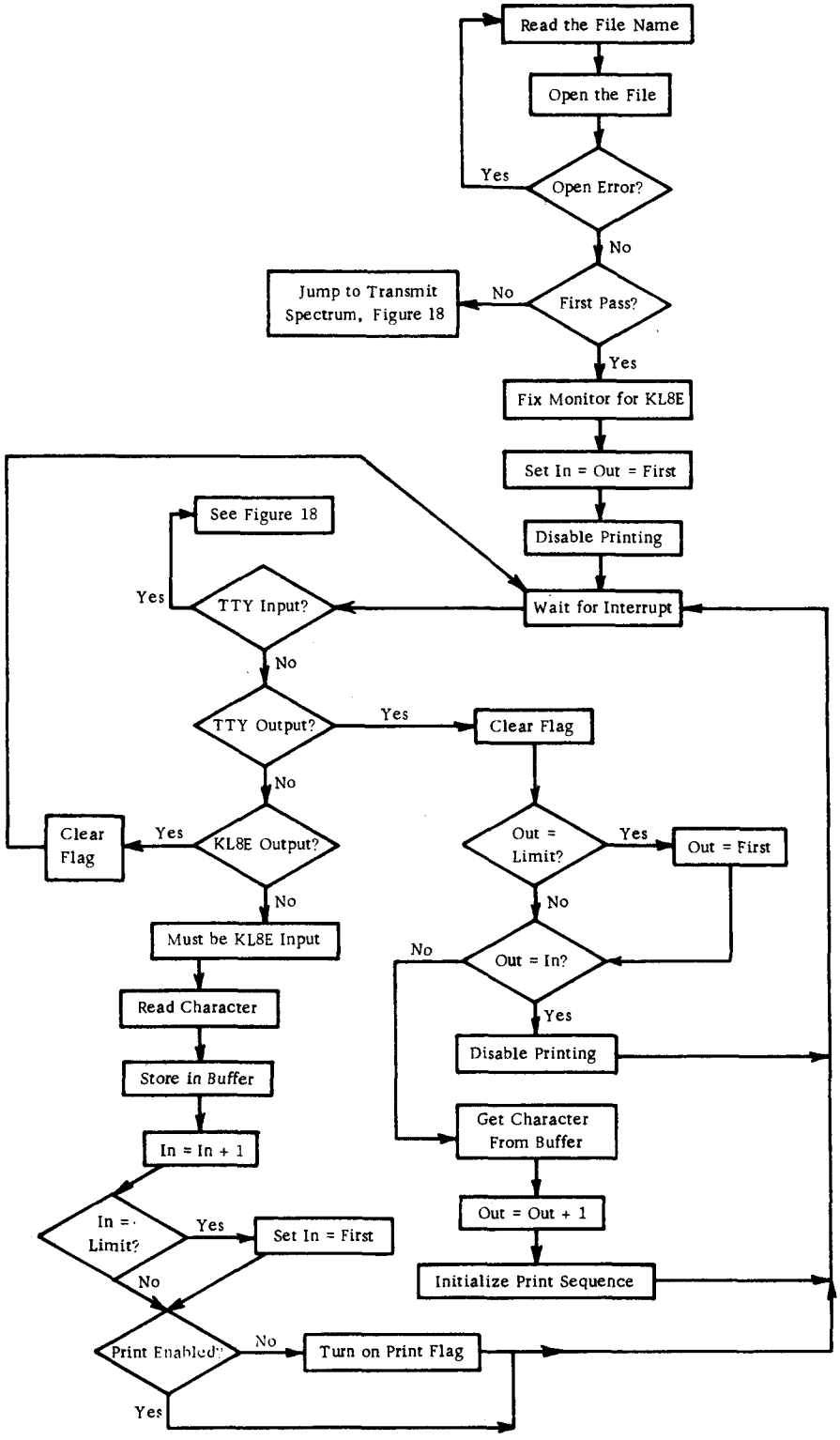

FIGURE 17. FLOW CHART FOR SYSTEM/150 ROUTINE DIRECT-I

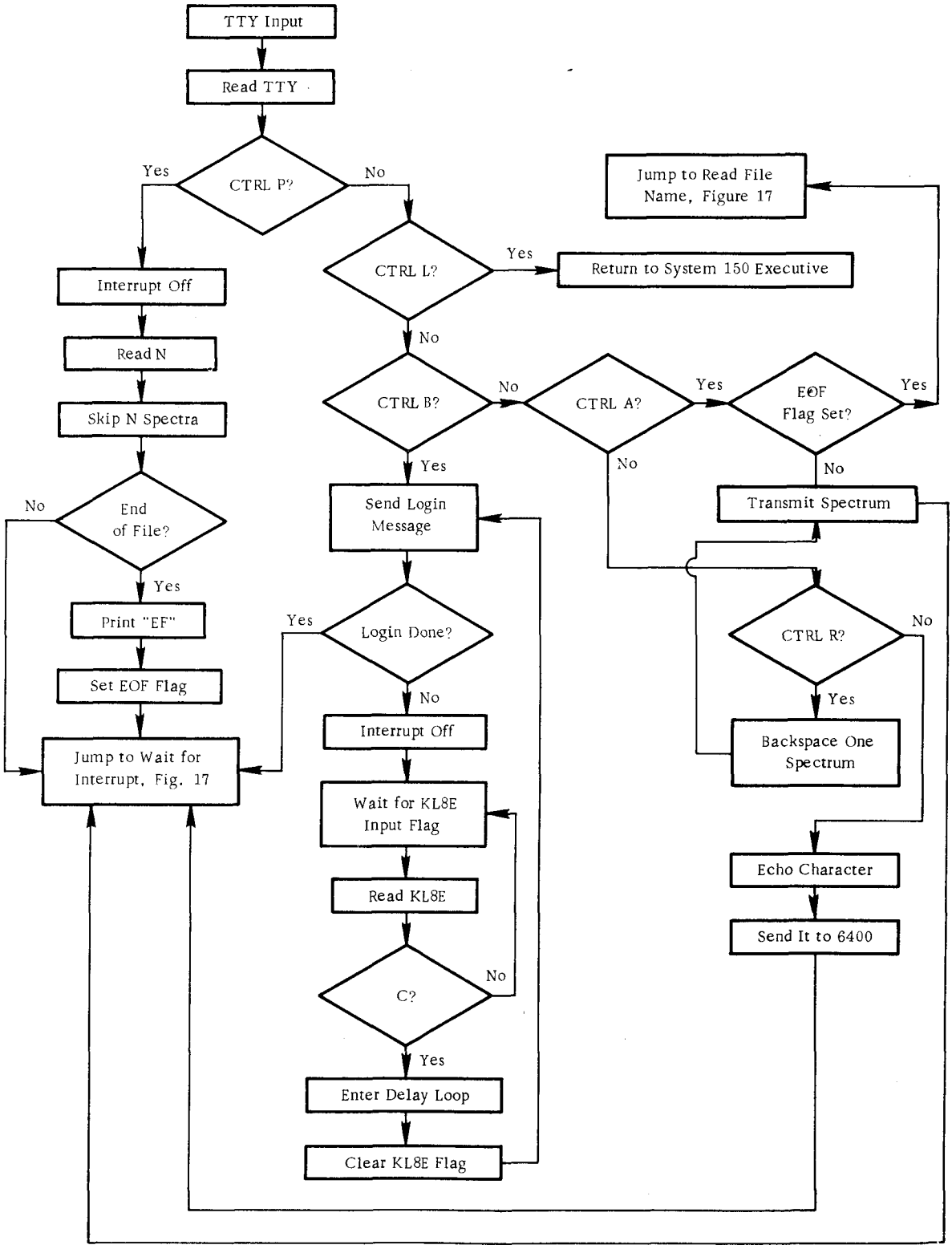

FIGURE 18. FLOW CHART FOR SYSTEM/150 ROUTINE DIRECT-II

control, but a step to clear this flag is included in the event it should happen to be raised under a situation other than local flag check control.

The KL8 input flag raised implies that a character has been sent from the 6400 for printing. The character is read, thereby clearing the flag and stored in core at the location IN. IN is incremented and a test is made as to whether the buffer capacity is exceeded. If so, IN is set to the first buffer location. The print enabled flag is then tested, and the printer flag is set on if a disabled condition is found.

The TTY output section is relatively simple. The flag either from the previous print or from enabling is cleared, and the pointer OUT is checked to see if a circling condition is present. If so, OUT is set to the first buffer location. Next, a check is made as to whether OUT = IN. If so, all characters in the buffer are printed, the print disabled flag is set, and a return to the wait for interrupt step is made. Otherwise, the next character from location OUT is placed in the accumulator, OUT is incremented, and the printer sequence is initialized.

The TTY input section is rather complicated due to special control characters that must be recognized. Any character other than CTRL A, CTRL B, CTRL P, CTRL L, OR CTRL R is taken to be a "normal character", which is echoed on the user's I/O device and sent to the CDC 6400.

The character CTRL P is used to skip one or more abbreviated spectra in the input file. All operations are done under local flag control at a time when the CDC 6400 is waiting for input. The number of records to be skipped is entered and the file is positioned. Occurance of the end of file mark causes the message EF to be printed and the end of file flag set.

Entering CTRL L is done after all communications with the 6400 are complete. A return to the System 150 executive is the result.

Automatic login is accomplished by striking CTRL B. The interrupt is turned off immediately, and the first login message is sent. A test

made to ascertain if all login lines have been sent (five altogether). If not, the system waits for the CDC 6400 to send the character C, the first letter of command. Upon receipt of this character, a 750 millisecond delay loop is entered, the KL8 flag is cleared, and the next message is sent. This continues until the login sequence is completed, at which time the interrupt is turned on, and normal operation resumes.

Spectra are sent to the CDC 6400 by striking CTRL A. A test for the end of file flag set is first made, and if it is set, the user is asked to attach a new file. The next spectrum in the file or the first in a new file is then transmitted under local flag check control. After a carriage return is sent, the program waits for the 6400 to return a line feed, after which the next line is sent. This procedure is repeated until the end of record mark is encountered. A test for end of file is made, and if satisfied, the message EF is printed and the flag is set. The first two lines of data, which is title information, is placed in the output buffer and printed as soon as a return to normal operation occurs and a character is received from the 6400.

One other character documented in Figure 18 is CTRL R. This character causes the spectrum just transmitted to be transmitted once more. This feature is enabled due to possible "garbling" from noisy telephone lines.

SECTION IX

REFERENCES

1. Hertz, H. S., Hites, R. A., and Biemann, K., "Identification of Mass Spectra by Computer-Searching of File of Known Spectra", Analytical Chemistry, 43, No. 6, pp 681-691 (1971).
2. Abrahammsson, S., "The Use of Computers in Low-Resolution Mass Spectrometry", Science Today, 14 No. 3, pp 29-34 (1967).
3. Pettersson, B., and Ryhage, B., "Mass Spectral Data Processing. I. Computer Used for Identification of Organic Compounds", Arkiv for Kemi, 26, No. 25, pp 293-303 (1967).
4. Crawford, L. R., and Morrison, J. D., "Computer Methods in Analytical Mass Spectrometry", Analytical Chemistry, 40, No. 10, pp 1464-1469 (1968).
5. Knock, B. A., Smith, I. C., Wright, D. C. and Ridley, R. G., "Compound Identification by Computer Matching of Low-Resolution Mass Spectra", Analytical Chemistry, 42, No. 13, pp 1516-1520 (1970).
6. Grotch, S. L., "Matching of Mass Spectra When Peak Height is Encoded to One Bit", Analytical Chemistry, 42, No. 11, pp 1214-1222 (1970).
7. Heller, S. R., "Conversational Mass Spectral Retrieval System and Its Use as an Aid in Structure Determination", Analytical Chemistry, 44, pp 1951-1961 (1972).
8. Heller, R. S., Fales, H. M., and Milne, G.W.A., "A Conversational Mass Spectral Search and Retrieval System-II: Combined Search Options", Organic Mass Spectrometry, 7, pp 107-115 (1973).

SELECTED WATER RESOURCES ABSTRACTS

INPUT TRANSACTION FORM

1. Report No.

2.

W

4. Title: IMPLEMENTATION OF A COMPUTER-BASED INFORMATION SYSTEM FOR MASS SPECTRAL IDENTIFICATION,

5. Report Date

6.

8. Performing Organization Report No.

7. Author(s)

Hoyland, J. R. and Neher, M. B.

10. Project No.

16ADN 29

9. Organization

Battelle-Columbus Laboratories
505 King Avenue
Columbus, Ohio

11. Contract/Grant No.

R-800921

13. Type of Report and Period Covered

12. Sponsoring Organization

15. Supplementary Notes

Environmental Protection Agency report number, EPA-660/2-74-048, June 1974

16. Abstract

A computer program has been developed for remote identification of mass spectra. Careful software design has led to a powerful and efficient system with minimum dialog and a highly flexible data input routine. Users may either access the main spectral library or create special libraries of their own. In addition, programs have been developed for the PDP-8/e or PDP-8/m computer of the System/150 to abbreviate spectra, punch spectra on tape, or to send spectra directly from the PDP-8 via a serial line interface.

This report was submitted in fulfillment of Project Number 16ADN 29, Grant Number R-800921, by Battelle Memorial Institute under the Sponsorship of the Environmental Protection Agency. Work was completed as of September 30, 1973.

17a. Descriptors

Mass spectrometry*, Pollutant identification*, Organic compounds*, Computers*, Data processing*

17b. Identifiers

Mass spectra, Computer identification

17c. COWRR Field & Group 05A

18. Availability

19. Security Class. (Report)

20. Security Class. (Page)

21. No. of Pages

22. Price

Send To:

WATER RESOURCES SCIENTIFIC INFORMATION CENTER
U.S. DEPARTMENT OF THE INTERIOR
WASHINGTON, D. C. 20240

Abstractor J. R. Hoyland

Institution Battelle Memorial Institute