From this viewpoint, we have attempted (1) to present the Polish notation as a symbolic language for substructure search, (2) to indicate the nature of computer algorithms employed in search, and thus, (3) to show that no substantial difficulty arises in the treatment of the simplest kind of structural diagrams.

LITERATURE CITED

(1) Eisman, S. H., *J. Chem. Doc.*, 4, 187 (1964).
(2) Hiz, H., *ibid.*, 4, 173 (1964).
(3) Berge, C., "The Theory of Graphs and Its Applications," John Wiley and Sons, New York, N. Y., 1962.

Rapid Structure Searches *via* Permuted Chemical Line Notations. III. A Computer-Produced Index*

CHARLES E. GRANITO,[a] JOHN E. SCHULTZ,[b] GERALD W. GIBSON, ALAN GELBERG,[c] R. J. WILLIAMS,[d] and E. A. METCALF
Industrial Liaison Office, Office of the Technical Director, U. S. Army Chemical Research and Development Laboratories, Edgewood Arsenal, Maryland 21010

Received March 5, 1965

INTRODUCTION

The previous papers in this series (*1*, *2*) have discussed the concept of an index of permuted Wiswesser chemical line notations,[e] the significance of a QUICK-SCAN area, and simple methods for preparing this type of index for a small index file of compounds (up to *ca.* 5000). It has been pointed out that the preparation of an index for a large number of compounds would require the use of a computer. This is the subject of this paper.

The project was started in 1963. At that time, a Univac File Computer (Model II) was readily available and, therefore, used for preparing the index. After the program was written and satisfactorily tested on a trial deck of 1000 cards, approximately 55,000 Wiswesser chemical line notations, on file in this office, were indexed. A program which achieves this same result has been written for an IBM 1401 at the T. R. Evans Research Center of the Diamond Alkali Company. Both programs are available for potential users.

The discussion of a general program to accomplish this permutation will be divided into four categories: (1) computer preparation of the index, (2) cost of preparing an index, (3) the index, and (4) uses of the index.

1. Computer Preparation of the Index. The input is a single punched card per compound, containing an accession number, a two-column screen, and a Wiswesser chemical line notation. The program is designed to effect the permutation of the line notation as each card is read onto magnetic tape; *i.e.*, the operations required to select pertinent symbols, generate the scan area, and permute the notation are accomplished and the results stored prior to acceptance of the next line notation.[f] The path followed by a typical line notation card in these operations will be discussed rather than giving the step-by-step details of the less understandable directions and flow charts of the programmer.

For the purpose of this discussion, it is convenient to think of the input information as occupying one row of a compartment in the core memory (Memory I, Figure 1). For easy visualization, the memory row which can hold 120 characters is further divided into four areas: A, B, C, and D (Figure 2). Each area is separated by a blank space to make the final printout more readable, and the following arbitrary assignments are made: (a) area A (8 spaces) for the accession number, (b) area B (2 spaces) for a prefix to serve as a screen in the index, (c) area C (11 spaces) for the QUICK-SCAN symbols to be generated by the computer, and (d) area D (96 spaces) for the line notation.

Areas A, B, and D contain information read directly from the card. At the start, area C is empty; it will be filled by symbols, selected by the computer in its operations on a line notation, for which an entry will be made in the index. The line notation will be found in the last half of area D (spaces 73 through 120); initially, the first half of area D (spaces 25 through 73) is empty. Space 73, the center of area D, corresponds to the index column of the listing.

After all of the information on one card has been fed into the input section of Memory I, the computer is ready to start generating the information for the QUICK-SCAN area and the permutations of the line notation. The notation is transferred to a reserve memory, Memory II (see Figure 1). In this transfer and all subsequent transfers of the notation, the information in areas A, B, and C is asso-

* Presented before the Division of Chemical Literature, Symposium on Work and Time Studies in Technical Information, 149th National Meeting of the American Chemical Society, Detroit, Mich., April 1965.

[a] To whom all inquiries should be addressed.

[b] Westminster College, Fulton, Mo.

[c] Diamond Alkali Co., Painesville, Ohio

[d] Data Processing Division, Management Science and Data Systems Office, Edgewood Arsenal, Md.

[e] Some notations, used as examples in this paper may not be consistent with the revision of Wiswesser line notation rules currently being prepared for publication by Dr. E. G. Smith, Mills College, Calif.

[f] It should be noted that for the Honeywell 200 400 the first step is a card to tape conversion, with the permutation a subsequent action.

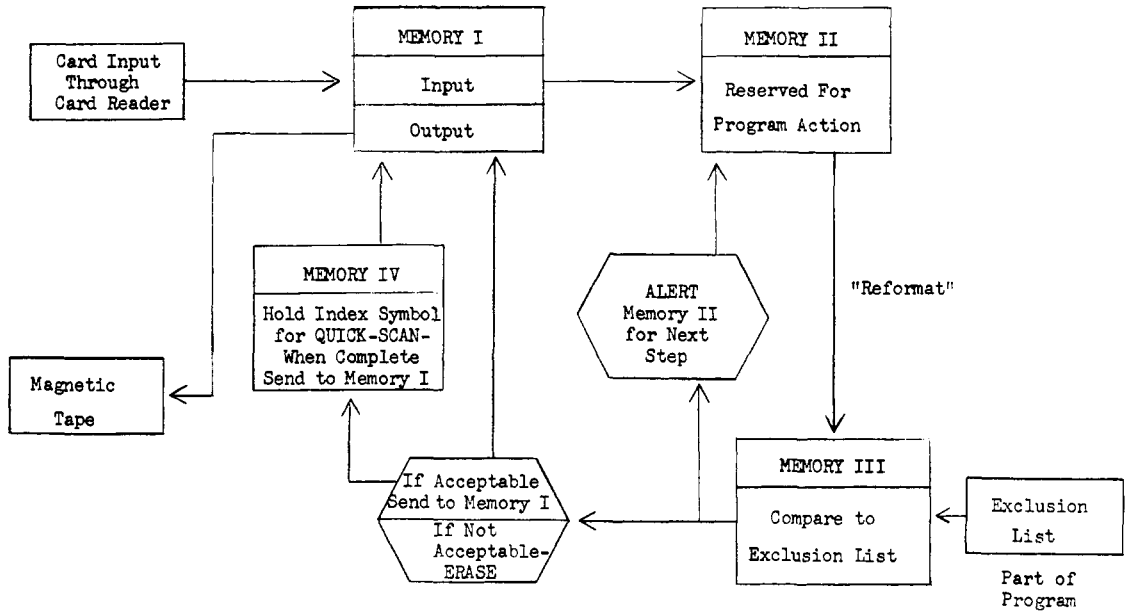

Figure 1. Information path from card to tape.

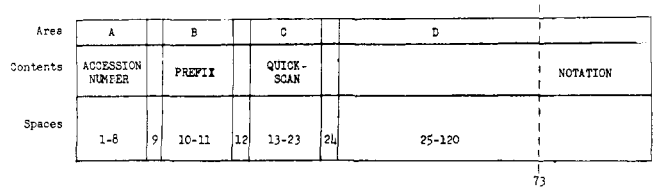

Figure 2. Representation of one section of input–output memory in computer Memory I.

ciated with the notation but is not operated upon. Memory II holds the notation for a series of actions!

(a) The notation is duplicated into Memory III, *i.e.*, the information in spaces 73–120 of area D is duplicated.

(b) In Memory III, the symbol in space 73, the indexing symbol, is compared with an exclusion list (2) (a list of symbols used in the Wiswesser chemical line notations that would not be useful as indexing terms). If the symbol (space 73) is not found on the list, the entire block of information is considered to be an index entry and is returned to Memory I for storage. Simultaneously, the index symbol (space 73) is duplicated and sent to Memory IV where it is held until all entries for this notation have been prepared. *The initial symbol of a notation is always an index entry.*

(c) As the first block is accepted by Memory I a signal is sent to Memory II and again the notation is transferred to Memory III. However, this time the transfer involves a position change (one space left), so that the line notation now occupies the 72–119 block, and the new symbol in space 73 is compared with the exclusion list. If the symbol is acceptable, the course of action is the same as for the first block of information. If the symbol is not a desirable indexing term, the information in Memory III is erased and Memory II is alerted by signal to continue the process.

(d) As soon as Memory III compares two successive spaces containing no symbols (indicates end of line notation) or after 120 (last position), this part of the operation ceases.

After each symbol in a line notation has been considered and the appropriate operations have taken place on each, the symbols stored in Memory IV are returned to I and the SCAN SYMBOLS are entered in area C for each entry that has been returned from the operations in Memory III. Prior to the acceptance of the second card, all information in Memory I is transferred to magnetic tape for storage and future sorting. The magnetic tape contains all the information generated from each line notation; if printed out at this stage, a listing as shown in Figure 3 would be

```
                                  INDEX COLUMN
                                       ↓
00001008 C1 TNQ                      T6NJ BQ
00001008 C1 TNQ                    T6NJ BQ
00001008 C1 TNQ                  T6NJ BQ
00024101  6 QVG                      QV2G
00024101  6 QVG                    QV2G
00024101  6 QVG                  QV2G
```

Figure 3. Example of magnetic tape printout prior to sorting.

obtained. The entire operation from card to storage of the permuted line notations on tape takes less than a second per card. The methodology described can be applied to line notations containing a maximum of 48 columns on a punched card. (This is an arbitrary assignment; and length may be chosen. For the Honeywell 400 we extended it to 60 columns and found that this covered greater than 99.9% of our file.) In our files, approximately 85% of the compounds require twenty columns or less. In some cases, it is desirable to include in the notation field certain nonnotational information, *e.g.*, 95% pure, dimer, etc. To accomplish this the notation proper is followed by two spaces and the remaining area (up to space 120) is used for any desired information. Information preceded by two spaces will be printed out with the various permutations.

In order to create an index, all of the taped permutations must now be sorted alphanumerically. This can be accomplished with an off-line tape sorter which requires no com-

puter time. The tape of sorted notations is used to print out the index. The index column may be indicated either with a line (specially printed paper) or with a printer-produced mark referencing the index column at the top and bottom of each page. The example of a printout (Figure 4) shows a page of permuted notations for some of the sulfur compounds appearing in the "Pesticide Index" (2nd Ed.). The address gives the page where the drawn structure and data may be found.

COMPOUND NUMBER	PREFIX	PERMUTED Symbols	NOTATION (Index Column)
P12 130A	38	HCNRSWR	2-HG-NR&SWR D
P12 152B	37	GSWRG	G1SWR DG
P12 211A	37	GRSWRG	G 3 DR ESWR DG
P12 112C	38	ZSWROPOO	ZSWR DOPO&/01 2
P12 098E	38	ZSWROPOO	ZSWR DOPO&/02 2
P12 112D	38	ZSWROPSO	ZSWR DOPS&/01 2
P12 130F	38	MSWROPSO	2MSWR DOPS&/01 2
P12 1L9C	38	YMSWROPSO	1Y&MSWR DOPS&/01 2
P12 098F	38	ZSWROPSO	ZSWR DOPS&/02 2
P12 052C	37	ZSWRG	ZSWR XG
P12 112E	B5	TNNRSWZNO	T5NNJ AR DSWZ& C DNO E
P12 091A	8	OPSSXVM	10 2PS&SX&&VM1
P12 047B	G7	TVNVUSXGGG	T56 BVNV GUTJ CSXGGG
P12 135A	G7	TVNVSXGGG	T56 BVNVJ CSXGGG
P12 210D	G7	TVNVUSXGGYGG	T56 BVNV GUTJ CSXGGYGG
P12 139A	X5	TOSSXVNVNUUQQ	T66 I65 O 1C ASSXVN GVN KU MU A BT&TTJ E F1Q JQ
P12 093E	8	OPSSSY	20 2PS&S1SY
P12 222B	8	OPOSSYM	10 2PO&S2SY&M1
P12 090B	38	OPSSYRCN	20 2PS&SYR&CN
P12 191C	P3	TNSYSNUS	T C566 BN DSYS HNJ EUS
P12 213C	C3	TNSYSNUS	T C566 BN DSYS HNJ EUS
P12 037B	3	NYUSSYUM	1N1&SSYUS&M 22
P12 128C	3	NYUSSYUSM	1N1&YUS&SYUS&M 22
P12 095B	38	OPSORS	20 2PS&OR C DS1
P12 130D	38	OPSORS	2OPS&1&OR DS1
P12 130E	38	OPSMORS	2OPS&M1&OR C DS1
P12 167C	38	SPORS	SP1&1&OR C DS1
P12 100A	C3	TNNNSMY	T6N CN ENJ BS1 D- F/MY 2
P12 200B	C3	TNNNSM	T6N CN ENJ BS1 D- F/M2 2
P12 167B	C3	TNNNSMYM	T6N CN ENJ BS1 DMY FM1
P12 021A	C3	TNNNSMYM	T6N CN ENJ BS1 DMY FM2
P12 095C	38	OPSORS	20 2PS&OR C DS1 E

Figure 4. Example of printout page of permuted notations for some sulfur compounds appearing in "Pesticide Index" (2nd Ed.). The designations in the Compound Number column refer to the page number appearing in "Pesticide Index."

2. Cost of Preparing an Index. The Univac File Computer (Model II) used for preparing the first large index is a relatively slow machine with a low tape density (92 characters per inch as compared to 556 for the Honeywell 400). It is, therefore, not comparable to most of the computers available in industry today which can perform this operation more economically. In order to make the dollar values more meaningful, the figures given will be based upon our experience with an IBM 1401 at the T. R. Evans Research Center and the Honeywell 400 which has replaced the Univac at the Data Processing Center of Edgewood.

The IBM 1401 was used for permuting 5991 line notations. The input time was approximately 1.25 hr. (80 cards/min.). For the 5991 compounds, a total of 33,080 entries were generated (5.5 entries/compound). These entries were stored on one-third of a reel of magnetic tape. In other words, one standard reel (2400 ft.) would hold all of the entries for the indexing of about 18,000 compounds, or 108,000 entries (assuming 6 entries/compound). The alphanumeric sorting time required for the 33,080 entries was 30 min. The printing time (650 lines/min.) was 55 min. Therefore, the total machine time (input to hard copy) was just under 3 hr. At a rental of $52/hr. and allowing $15 for paper, the cost of preparing an index of the permuted line notations of about 6,000 compounds, excluding labor and programming costs, would be approximately $165, *i.e.*, 2.75 cents/compound.

Experience gained with the Honeywell 400 computer at Edgewood Arsenal leads us to predict that the 350,000 entries generated from 55,000 compounds (6.3 entries/compound) could be stored on 3 to 4 tapes whereas 21 tapes were required for the Univac II. Input requires about 3 hr., sorting time 4 hr., and printing time (900 lines/min.) 5 to 6 hr. Once again using a reasonable rental of $52/hr., the 55,000 compounds (350,000 entries) could be indexed for approximately $850, including $200 for paper. This averages out to about 1.5 cents per compound.

In addition to actual operation costs, the cost of the program for the computer must be considered. This is a one-time job for a given computer. The program written for the Univac took 100 hr. (including debugging time). Since programming usually costs about $10/hr., $1000 would be a good estimate of the cost for writing the first program. However, the program subsequently written for the IBM 1401 required only 24 hr. or $240. Since both programs are available on request, the cost for a program on a different model should be comparable to that required for the 1401. In other words, for about $1100 one could generate an index of permuted line notations for 55,000 compounds.

Generation of 55,000 line notations from chemical structures would cost about $10,000 (or 18 cents/notation). The rates used to arrive at this figure are summarized in Figure 5. It should be remembered that notations, once prepared, will serve as input not only for the first index but subsequent ones as well.

Input (line notations)	Time, days	Cost, $
Writing notations	110 (500/day)	4400 ($40/day)
Proofing notations	110	4400
Preparing punched cards	27.5 (2000/day)	440 ($16/day)
Verifying punched cards	27.5	440
	Total	$9680
Cost of 55,000 cards at $0.00125/card		69
Cost for machine rental (key-punch and verifier)		120
		$9869
Index		$240
Program cost		676
Machine rental (13 hr., $52/hr.)		200
Paper		$1116
	Total	$12,085
Total cost for indexing 55,000 structures		

Figure 5. Time and cost estimates for an index of permuted line notations for 55,000 compounds.

3. The Index. For 55,000 line notations, an index of 7173 pages was obtained. There was a maximum of 49 entries per page; *i.e.*, over 350,000 entries were generated. These pages were divided into 24 volumes. For ease of handling, each volume was cut to 11 × 14 in. and bound with hard covers. The back of each volume was labeled in the same manner as an encyclopedia. After binding, the books occupied 40 in. of shelf space.

4. Use of the Index. In order to test the usefulness of the index some 25 structure searches were carried out. Each of these searches had been made previously using molecular formulas, a fragmentation code (*3*), and tabulated listings of line notations in alphanumeric order. In each case, the index was found to be significantly faster, and a greater number of compounds meeting the search criteria were

found. Some searches which had required a full day were completed in a matter of minutes using the index.

The index is used in the same manner as a dictionary. Both specific and general searches including analogs and homologs may be run at one's desk. There is no further need to use any mechanical equipment in locating desired structures.

a. Specific Look-Up. For a specific structure, one prepares the line notation and looks it up in the index. This process has proven to be faster than writing out a molecular formula and locating it in a molecular formula file. Each notation represents but one compound, whereas each formula may represent many compounds. Since the notation is unique and unambiguous, it can be found in a specific location in the index. An average structure requires about 15 sec. to encode and a notation can be found in the index as fast as a word can be located in a dictionary.

b. General Searches. The procedure for running general searches is nearly as simple as the specific look-up. One decides which symbols meet the requirements of the request and either locates these in the index or assigns a clerical worker to the job. The frequency of occurrence for index symbols would determine the starting place in the index. Table I presents the frequency of index symbols for our first index. This table could be extended to include a more detailed break-down (beyond the first index symbol), if desired. It is not necessary to know what the line notations represent to find entries in the index, just as it is not necessary to know the meaning of a word to find it in a dictionary.

Table I. Frequency[a] of Occurrence for Index Symbols (Permuted Index # 1)

Symbol	Rank	Total	Symbol	Rank	Total
-	27	206	I	22	1,966
1	17	4,381	J	33	11
2	21	2,132	K	20	3,450
3	26	278	L	13	8,421
4	25	647	M	8	20,785
5	32	93	N	1	47,201
6	29	151	O	2	38,320
7	30	114	P	18	4,029
8	28	174	Q	5	27,513
9	31	96	R	4	34,604
A	23	1,731	S	9	19,547
B	24	1,361	T	6	24,401
C	16	4,565	U	10	17,003
			V	3	35,617
E	19	3,624	W	12	8,941
F	15	5,430			
G	7	23,524			
H	14	7,502	Z	11	12,453

[a]These frequencies are for *indexed* symbols, and do not include those excluded by the program (*e.g.*, the T count is for T's which initiate ring systems, not T's indicating ring saturation).

With an index of permuted line notations the individual in charge of chemical structure retrieval could rapidly answer such questions as:

(1) How many quinoline derivatives are on file?

(2) Which quinolines contain a nitro group as well as a hydroxyl group?

(3) How many aliphatic alkynes have been investigated and which ones contain chlorine?

All of these questions could be answered routinely. They would be handled as follows:

(1) Open the T volume of the index to the T66 BNJ section (a matter of seconds).

(2) Visually check the QUICK-SCAN area of the quinoline section, checking off those for which a NW and Q appear (for the nitro and hydroxyl groups, respectively).

(3) Open the index to the UU section, count the compounds not containing an L (carbocyclic), T (heterocyclic), or R (benzenoid) in the QUICK-SCAN area. Check off those which do have a G in the QUICK-SCAN area and list them.

A clerk with no chemical background could find the answers to each of the questions given above if provided with a list of appropriate search symbols. If a large number of searches were to be run every day, this would be the most economical approach. The frequency of searches may well determine the need to perform computer searches. Since the permuted notations are already on magnetic tape, a search program can be written and performed.

When a large number of compounds are found that meet the search requirements, it has been found useful to make a Xerox reproduction of the index pages instead of writing down the numbers for each compound. This avoids transcription errors which mount up quickly for long lists of numbers.

It is realized that the above questions are not complex. However, these are the types of questions most frequently asked. With the index, such questions take only minutes to answer. Any increase in the numbers of parameters or specificity shortens the search time required by reducing the search to a smaller section of the index.

c. Limitations. The only questions the index is not designed to handle efficiently are those involving the relative positions of every atom within a molecule, *e.g.*, which compounds in the file contain a nitrogen atom three atoms from any oxygen atom and two carbons removed from a sulfur atom? In Wiswesser line notations, symbols generally represent groups of atoms rather than an individual atom. This prevents the use of the index for such searches. However, it is not a limitation of the notation, for a computer could be programmed to use the notation for such searches (*4*).

d. Who Needs to Know the Notation? Only one or two chemists within an organization need to know the Wiswesser line notation to make the index a useful means of retrieving chemical structures. The chemist or administrator can request information by structure(s) and receive answers in the same form. After compound numbers are found in the index, structure cards can be pulled, reproduced, and sent to the requester. The notation serves only as a means by which compounds are located.

e. Potential. As demonstrated, the index is a very powerful tool for keeping track of a file of chemical compounds. However, it has an even greater potential. For example, it could expedite procurement procedures. An organization which routinely purchases large numbers of materials from commercial suppliers could encode the compounds listed in available catalogs and include the source and price of each item. Preparation of an index would then permit rapid determination of compound availability, the companies offering an item, and their listed prices. All of this infor-

mation would be found in the same section of the index, under the specific notation.

A second possibility is the generation of a functional group index for all chemical structures appearing in a catalog, textbook, journal, or secondary source publication.

Another possibility that is now being explored is the incorporation of biological data into the index. The inclusion of such data would permit investigation of structure–activity relationships. A glance at a given section would reveal how many compounds contained a given ring structure(s), functional group(s), or combinations thereof, as well as the type and level of activity exhibited. This could be a powerful aid to research. Any type of data could be included; the possibilities are unlimited. The data can be ordered by any desirable parameter with the linearized structures associated with it for comparison.

f. Updating the Index. Updating the index does not present a problem. Since the program is available, supplements, which will include compounds received after the major index was generated, can be prepared at suitable intervals. When the supplements become too numerous for easy searching, the tapes used for each index can be blended and used to create a new master index. The frequency of updating would depend upon the growth rate of the file.

SUMMARY

The preparation of a computer-produced index of permuted Wiswesser chemical line notations is described. The uses and limitations of this powerful and economical retrieval tool are discussed. The utility of such an index may be markedly increased by the inclusion of biological, source, cost, etc., data.

ACKNOWLEDGMENTS

The authors are indebted to Dr. D. H. Frear for permission to use the Second Edition of the "Pesticide Index" as a source for compound to be coded and used as examples (Figure 4); Mr. W. Nugent and Mr. W. Matthews, Diamond Alkali Company Data Center, for their efforts and excellent cooperation in preparing the IBM 1401 program; and the Edgewood Arsenal Data Processing Center for preparing and implementing the Univac program.

LITERATURE CITED

(1) Sorter, P. F., Granito, C. E., Gilmer, J. C., Gelberg, A., Metcalf, E. A., *J. Chem. Doc.*, **4**, 56 (1964).
(2) Granito, C. E., Gelberg, A., Schultz, J. E., Gibson, G. W., Metcalf, E. A., *ibid.*, **5**, 52 (1965).
(3) Gelberg, A., Nelson, W., Yee, G. S., Metcalf, E. A., *ibid.*, **2**, 7 (1962).
(4) Landee, F. A., Abstract of Papers, Division of Chemical Literature, 147th National Meeting of the American Chemical Society, April 6–9, 1964, Philadelphia, Pa., p. 3F.

Computer Searching of Chemical Patents*

P. T. O'LEARY,[a] J. M. CATTLEY,[b] J. E. MOORE,[a] and D. G. BANKS[a]
Gulf Oil Corporation and Gulf Research & Development Company, Pittsburgh, Pennsylvania

Received April 29, 1965

In the fall of 1963, Gulf Research & Development Company bought from Information for Industry, Inc., an index, on magnetic tape, of the more than 100,000 U. S. chemical patents which had issued during and since 1950. The index is of "uniterm" or "coordinate" type. The indexer selects from the text of the patent he is indexing those words he believes characterize the patent. Frequently used words accumulated in this way comprise the "vocabulary" of now more than 9000 terms, or "descriptors," used for the magnetic tape form of the index. Thus, keywords, rather than concepts, are indexed. The index is "inverted"; that is, under each descriptor are listed the accession numbers of patents partially characterized by that descriptor.

The index was available in dual dictionary form for clerical "coordination" and on magnetic tape for search on an IBM 1401 computer with 8k memory.

Better to fit our computer facilities, we adapted the index for search on our IBM 7094. The adaptation of the IFI Index involved re-inversion of the index, during which clerical errors were corrected. A retrieval program was developed for the 7094 to search this re-inverted, or serial, file. This program has greatly increased speed and flexibility in searching, which we feel is essential to our use of the index.

With the index as now set up, we make many searches per week at an average time of about 1.5 minutes per question on an IBM 7094 computer. Most of these are on chemical ideas submitted by our research people. These span many petroleum industry operations: refining, devel-

* Presented before the Division of Chemical Literature, 148th National Meeting of the American Chemical Society, Chicago, Ill., Sept. 1964.
[a] Gulf Oil Corp.
[b] Gulf Research & Development Co.