REPORT INDEX SEARCHING ON THE BENDIX G-15D COMPUTER*

By JOSEPH D. GRANDINE, 2nd,[1] EVA M. STARR,[2] and RICHARD E. PUTSCHER[3]

Contribution from the Textile Fibers Department, E. I. du Pont de Nemours & Co., Wilmington, Delaware

ABSTRACT

An experimental system has been developed for the mechanized searching of a subject index to company technical reports. The form and content of the index (which covers 14,000 reports) is the same as that used previously on edge-notched cards. Serial searching of the index on magnetic tape by the Bendix G-15D digital computer has been in operation for several months. Reduced time and increased efficiency are leading to wider and more effective use of the technical report file by research personnel. The design, operation, and economics of the system are presented.

INTRODUCTION

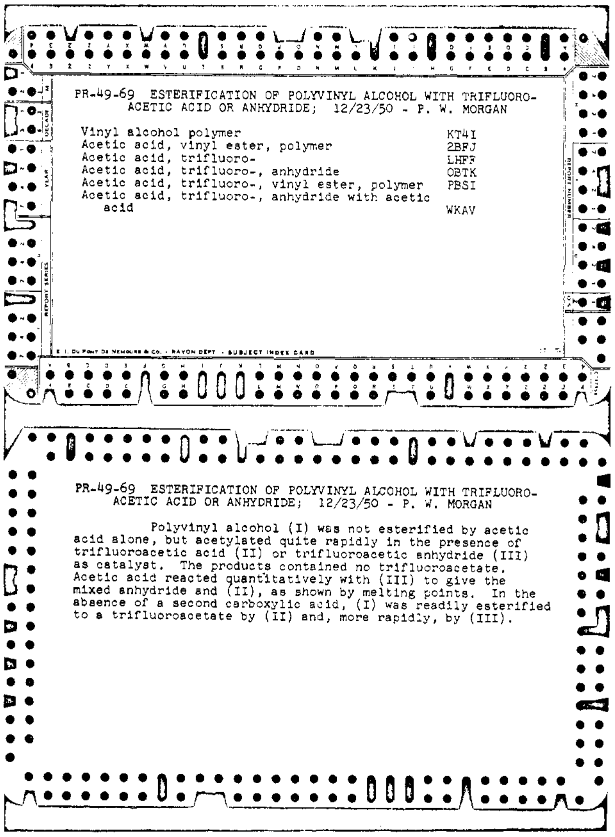
The Patent Division of the du Pont Textile Fibers Department includes in its activities the preparation and distribution of a subject index to internal research reports of interest to the department. The index, which has been maintained[4] on marginal punched (Royal-McBee Keysort) cards, includes more than 14,000 individual reports indexed on 15,000 cards with a total of more than 140,000 index entries, and is growing in size at a rate of 15-20% per year.

The searching of this index to identify reports containing information relevant to day-by-day research and management problems is an important function. Hand searching of the complete file has become increasingly time-consuming, resulting in less than optimum use of the index and, hence, of the corresponding indexed material. In practice, also, the tendency of extensively edge-notched cards to hang together leads occasionally to a failure to retrieve all pertinent documents.

This paper describes an experimental system for serial searching of this file on a general-purpose digital computer (the Bendix G-15D) using magnetic tape storage of the index. The system to be described began operation in January, 1959.

THE INDEX

The indexing and coding system originally was designed for use with marginal punched (Royal-McBee Keysort) cards. No attempt has been made to recode the index for use on the computer. In the present system, each subject selected as an indexing point is assigned a 4-character alphanumeric code. For each report, the report number (series identification and serial number) and a list of these 4-character codes are typed on the face of a single card (Fig. 1). An abstract of the report appears on the back of the card. A maximum of 14 subject codes is punched in the margins of a single card. Reports indexed by more than 14 subjects require multiple index cards. Since each card is a unit insofar as the search is concerned, the codes for important subjects are repeated from card to card for multiple-card reports. This is necessary because of the desirability of searching the index for combinations of subjects.



PR-49-69 ESTERIFICATION OF POLYVINYL ALCOHOL WITH TRIFLUORO-
ACETIC ACID OR ANHYDRIDE; 12/23/50 - P. W. MORGAN

Vinyl alcohol polymer	KT4I
Acetic acid, vinyl ester, polymer	2BFJ
Acetic acid, trifluoro-	LHFF
Acetic acid, trifluoro-, anhydride	OBTK
Acetic acid, trifluoro-, vinyl ester, polymer	PBSI
Acetic acid, trifluoro-, anhydride with acetic acid	WKAV

PR-49-69 ESTERIFICATION OF POLYVINYL ALCOHOL WITH TRIFLUORO-
ACETIC ACID OR ANHYDRIDE; 12/23/50 - P. W. MORGAN

Polyvinyl alcohol (I) was not esterified by acetic acid alone, but acetylated quite rapidly in the presence of trifluoroacetic acid (II) or trifluoroacetic anhydride (III) as catalyst. The products contained no trifluoroacetate. Acetic acid reacted quantitatively with (III) to give the mixed anhydride and (II), as shown by melting points. In the absence of a second carboxylic acid, (I) was readily esterified to a trifluoroacetate by (II) and, more rapidly, by (III).

Fig. 1 – Royal-McBee Card, front and back

A dictionary of subject codes is maintained, arranged alphabetically by subject, with cross-references. This allows the searcher to select

*Paper presented before the Division of Chemical Literature, American Chemical Society Meeting, Atlantic City, N. J., September, 1959. Correspondence regarding this paper may be addressed to Dr. C. C. Conrad, Supervisor, Information Section, Patent Division, Textile Fibers Department, E. I. du Pont de Nemours & Co., Wilmington, Del.

easily the codes to be used in a search. It also allows one to determine rapidly whether a given subject has ever been used to index one or more reports, by determining whether this subject appears in the dictionary.

SEARCHING THE INDEX

In use, the index is employed to provide report numbers and abstracts for all indexed reports pertaining to a specified combination of subjects. For example, a search based on the subjects melt-spinning and polycaproamide would yield a list containing the reports dealing with the melt-spinning of polycaproamide. Some "noise" creeps into the output from the system due to the absence of coding for relationships between subjects. For instance, the foregoing search would also find a report which included data on the melt-spinning of polymer X and also data on the properties of polycaproamide.

It is a simple matter to exclude classes of reports in order to narrow the field of inquiry. For example, a search on melt-spinning and polyamide would lead to a very large number of reports, which could be reduced greatly by discarding those indexed on 66 nylon, if our interest were only in other polyamides.

A hand-needled search of the entire file takes from two to three hours for a person familiar with the file and the Keysort system. Since most searches are run upon request by librarians who have other duties, a time lag of one or two days between question and answer is not unusual.

REQUIREMENTS FOR MACHINE SEARCHING

In setting up a system for searching this report index by machine, several guiding rules were adopted. Some were dictated by necessity, and others were adopted for reasons of expediency or convenience. The rules under which this work has been done are:

1. All input-output shall be in the form already familiar to the creators and users of the punched-card file. Thus, the index itself and all questions and answers will be alphanumeric, and a minimum of modification of the file will be required (none, except that a maximum length of ten characters is imposed on a report identification, for convenience in tabulating search output).

2. A permanent record of each search (questions and answers) shall be produced, preferably on punched tape or magnetic tape. Later studies of the use of the file will be able to use these records.

3. The searching system shall use only commercially available equipment.

4. Whenever possible, machine checks will be included to minimize the human errors which can find their way into the final mechanized system.

GENERAL DESCRIPTION OF THE COMPUTER SYSTEM

The present experimental system for automatic searching of the report index uses the equipment to be described:

1. Bendix G-15D general-purpose electronic digital computer with a magnetic drum internal storage capacity of 2160 "words" and intermediate internal operation speeds. The computer is basically a mathematical tool using binary numbers. Each 29-bit computer "word" is equivalent to an 8-digit decimal number with sign. The time to interpret and execute a basic command (add, subtract, test, and the like) is 0.00054 second, which may be increased up to 0.029 second if the information to be processed is stored in the worst possible configuration on the magnetic drum.

2. MTA-2 magnetic tape accessory to the G-15D. This provides auxiliary storage of up to 300,000 computer "words" which are available under control of the computer program. Access to the information stored on this tape is much slower than that to information on the drum. One such magnetic tape unit is used by the searching program. However, file revision (updating) requires two.

3. AN-1 alphanumeric accessory to the G-15D. This provides a means of reading alphanumeric information into and out of the computer using a 7-channel punched paper tape. A 60-character-per-second punched tape reader is used as input for questions and a 60-character-per-second tape punch is used as output for lists of reports. The output is printed from tape using a Flexowriter.

THE MECHANIZED SEARCHING SYSTEM

The computer programs developed for this project have been submitted to the Bendix Users EXCHANGE Program Library. They are:

Title	Users' Project No.
Ascending 4-Word Merge	270
Insert Shift Codes into Alphanumeric Item	271
Ascending 4-Word Sort	272
Screen Alphanumeric Item	273
Ascending 100-Word Sort	274
Tape Justification	367
File Preparation	368
Question Analyzer	369
File Searching Routine	370
File Correction and Updating Routine	371
Service Routines	372

The name of the current Distribution Secretary for the EXCHANGE Program Library may be obtained from the: Bendix Computer Division, Bendix Aviation Corporation, 5630 Arbor Vitae Street, Los Angeles 45, California.

1. The index file is stored serially on magnetic tape on the MTA-2 unit. All of the subject codes (up to 100) for a single report are stored together with the report identification in a single "block" on the magnetic tape. Each block contains from 12 to 108 computer "words." The report number is stored in alphanumeric form, ready for use in the search output. The subject codes have been converted to compact numerical form, sorted numerically, and stored as differences between adjacent numbers in the sorted sequence. Each code for a given report appears only once in the file.

This arrangement of the file was chosen in order to make possible a rapid comparison between the lists of subject codes in the question and the file, respectively. The basic problem is one of determining which, if any, subjects from one list (the question) are to be found on another list (the file). We may represent the question list by a sequence of upper-case letters, and the file list by a sequence of lower-case letters. If there are Q subjects in the question list, and F subjects in the file list, there are Q x F possible comparisons to be made. It is then possible to devise a program in which one comparison is made per drum cycle, so that Q x F drum cycles will be required for each question-report combination.

If both the question list and file list are each arranged in order of increasing numerical value, the number of comparisons which need to be made is decreased. This is illustrated on Fig. 2. After comparing capital A with small a, we will make only one of the three tests to the right and/or below this test. Likewise, after every succeeding comparison, we select our next test from those to the right and/or below our present position, never going up or to the left. In the example shown, Q = F = 4, so that 16 possible comparisons exist, indicated in the large circles. We start at the upper left and move to the right or down, or both, until we exit from the checkerboard pattern. This may require from 4 to 7 comparisons, depending on the sequence of tests required. In general, the number of comparisons to be made in a Q x F system will be no greater than Q + F - 1. This represents a considerable saving of time.

Figures 3a and 3b show the reason for storing differences between codes instead of the codes themselves, in both the question and the file lists. By using the differential lists, one addition or subtraction operation is saved on every horizontal or vertical move on the checkerboard.

The computer program used makes one pass through a 4 x 4 subset (Fig. 2) and to reload one of the 4-word lists, all within a single drum cycle. If a "hit" is made, the routine drops out and records this fact (using an extra drum cycle to do so), then re-enters the matrix at the point where it left. Final exit is made at any of the "reload" points when it is discovered that either the question or the file list has been exhausted.

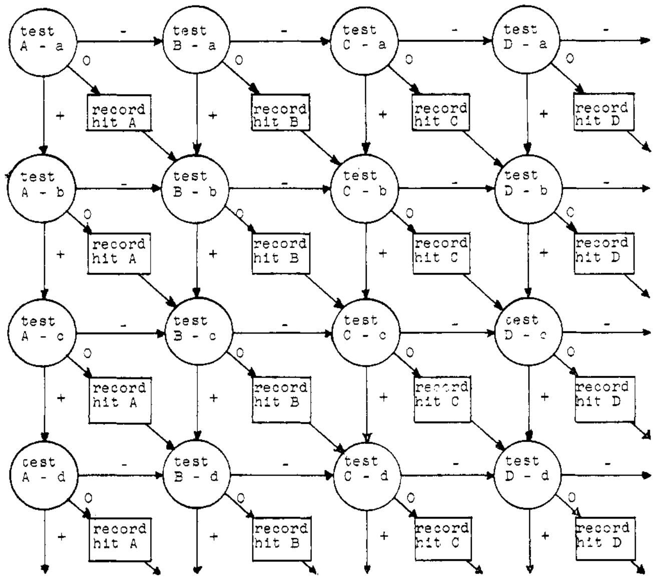
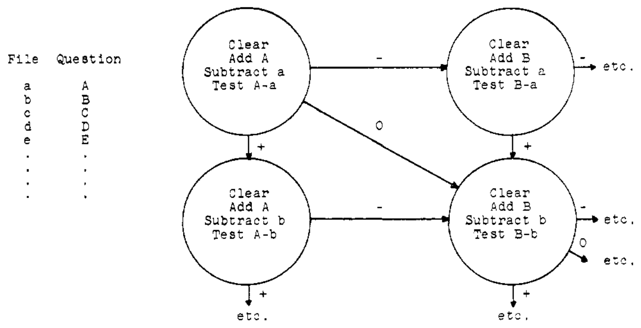


Fig. 2 – Comparisons Between Ordered Lists



File
Question
a
b
c
d
e
A
B
C
D
E
Clear Add A Subtract a Test A-a
Clear Add B Subtract a Test B-a
Clear Add A Subtract b Test A-b
Clear Add B Subtract b Test B-b
+
-
0
etc.

Fig. 3a – Comparisons of Question with File – Ordinary

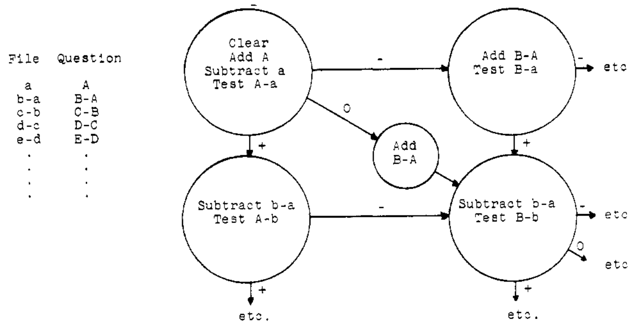


Fig. 3b – Comparisons of Question with File – Differential

2. The search program is stored permanently on 5-channel punched paper tape. During a search it is read into the computer through the standard photoelectric tape reader unit of the computer and is stored on the magnetic drum. This program consists of a system of "commands" which control the operations of the computer and accessories during the searching process. For each question list, the program scans the entire file, from beginning to end, searching for reports having the specified subject combinations.

3. The question list is the variable input to the search program. It consists of up to 16 questions which will be answered by a single search through the index file. Each question may involve up to 16 subject codes, with the limitation that not more than 16 different subject codes may be used by all of the questions in a single search. Each subject in a question may be included (+ sign or no sign) in or excluded (- sign) from the reports which are to be listed as the answer to the question. The question list is read into the computer memory via the high-speed reader of the AN-1 from 7-channel punched paper tape prepared on the Flexowriter.

If more than 16 questions are to be asked, successive unattended searches of the index file may be made. The question list containing all searches, in the order to be searched on the computer, is normally prepared as one continuous punched paper tape and read into the AN-1 at the time the search is started. After one search is completed, the index tape file is automatically rewound and the next search of the file started. This automatic feature permits unattended successive searches of the index file, which can be run, for example, overnight. See paragraph 5, Successive Searches, below.

An example of a question list with six questions using ten different subject codes is shown in Fig. 4. The numbered columns contain the 4-letter subject codes actually appearing on the question tape. In the center column are given the meanings of the subject codes.

Fig. 4 – Question List

Laundering durability of fabrics of "Dacron"* and of blends of "Dacron" with cotton.

01		02
MIGG	Launderability	MIGG
XJ2Q	Polyester fiber	XJ2Q
	Cotton	ZGSD

Thermal stabilizers for polymerization, especially of polymers other than polyamides and polyesters.

03		04
KZND	Stability, thermal	KZND
3BGN	Polymerization	3BGN
	(eliminate) Polyamide	-OARH
	(eliminate) Polyester	-PBSI

Density of 66 Nylon, especially as a function of temperature.

05		06
IXLB	Density, true	IXLB
VGWL	66 Nylon fiber	VGWL
	Temperature	YK3R

*Trademark for du Pont's polyester fiber.

The questions are thus phrased as lists of subject codes (with sign) in a form familiar to the searcher using the marginal punched-card version of the index. The first part of the search program carries out an analysis of the question list, in which duplicate codes are identified, the subject codes are converted to numerical form and sorted, and a minor file is prepared on the magnetic drum, having the same arrangement as the subject code portion of the index file on magnetic tape. This format has been designed for efficiency of comparison on a general-purpose computer.

4. The form of the search output for the sample question list is shown in Fig. 5. Since the search number and questions are included in the output which is punched on 7-channel paper tape, this tape constitutes the desired record of the search. Up to 16 numbered columns are produced, one column for each question on the question list.

Fig. 5 – Answer List: SEARCH NO. 0144-8/5/59.

```
01          02          03          04          05          06

MIGG        MIGG        KZND        KZND        IXLB        IXLB
XJ2Q        ZGSD        3BGN        3BGN        VGWL        VGWL
            XJ2Q                    -OARH                   YK3R
                                    -PBSI

                        AAA-50-345  AAA-50-345
                        AAA-50-385
                        AAA-51-2
                        AAA-51-7
                        AAA-51-91   AAA-51-91
                        AAA-51-268  AAA-51-268
                        AAA-51-372
                        AAA-51-385
                        AAA-52-81   AAA-52-81
                        AAA-52-134
                        AAA-52-146  AAA-52-146
                        AAA-52-291
                        AAA-53-119
                        AAA-53-283
                        AAA-54-20   AAA-54-20
BB-BB21409              AAA-56-293
BB-BB21412
BB-BB21469                                      E-372
                                                E-830
                                                E-1067
CC-54-56                                        E-1077
CC-55-137   CC-55-137                           CC-55-12    CC-55-12
CC-56-31
DD-56-35
DD-56-45    DD-56-45
DD-57-18    DD-57-18
DD-57-50                                        EE-54-17
                                                EE-57-48
FF-57-28                                        FF-53-17    FF-53-17
FF-78-63                GGG-51-132
HHH-56-28   HHH-56-68   GGG-56-109  GGG-56-109
HHH-56-71               II-56-24
                        II-56-28
                        II-56-38    II-56-38
                        II-56-44    II-56-44
                        JJJ-56-7
```

5. Successive searches are carried out automatically until the end of the question list tape is reached. The entire search operation requires no human intervention, once the program tape has been mounted on the computer, the file tape on the MTA-2, and the question list tape on the AN-1 reader. The magnetic tape file is rewound automatically at the end of each search.

6. Search time varies somewhat with the number of lines of output, but is approximately three hours for one complete question list of

up to 16 separate questions. If successive searches are run, a proportionally longer time is required, i.e., six hours for two searches, nine hours for three, etc. Since unattended operation is possible, most searches have been made during the night, usually with two or more question lists being processed in a single run.

OPERATING EXPERIENCE

Experience during the first 30 weeks of operation (January-July, 1959) of the serial searching system is summarized in Table I.

TABLE I. INITIAL OPERATING EXPERIENCE

	30 Weeks (Total)	Per week
No. of search runs	141	4.7
No. of columns of output	1726	57.5
No. of questions searched	501	16.7
Average no. of columns/question	3.5	–
Average no. of columns/search	12.2	–
Total computer hrs. (unattended)	460	15.3
Information specialist time (hrs.)	540	18.0
Flexowriter printing	70	2.3
Abstracts supplied (on request only)	891	29.7
Answers from previous runs	36	–

For this same initial operating period, the average cost per question was about $10 (the average calculated over 501 questions was 3.5 columns of output per question at a cost of $9.96). Of the ten-dollar cost of answering the average question, about one-quarter is for machine rental, and the balance is for the information specialist who accepts and interprets the question, reviews and surveys the search output, and forwards the answer to the originator of the question.

CONCLUSIONS

This initial experience with the system lead us to these conclusions:

(1) Serial searching of an index stored on magnetic tape is practically feasible with an existing small digital computer — the Bendix G-15D. It is economically feasible if the computer utilization is reasonably high. In our case, the computer was shared with a research laboratory which used 80% of the computer capacity. This application should be of interest to organizations having files to be searched of the same order of magnitude as that described here, and having also available for their use a small general-purpose computer such as the G-15D.

(2) Relatively little computer time, all of it unattended, will handle a good volume of questions. Unless the inquiry traffic is very heavy, the computer must be shared with other users.

(3) The chief cost and time-consuming bottleneck is the manual screening of answers to eliminate semantic noise from the results of the machine search.

(4) We believe that the next step in the improvement of our present system is to provide more detail in the indexing of documents and to show relationships between terms.

(5) Searches would be made more rapidly if the index file were inverted to provide parallel entry points; this is the next step in mechanization.

(6) Automatic production as machine output of titles and authors of documents will be programmed; similar production of abstracts or complete documents is being considered.

Subsequent to the presentation of this paper at the September, 1959, A.C.S. Meeting, the use of the Bendix G-15D computer was discontinued; our searches now are made on an IBM-705 computer using an inverted file storage. The change was made because the larger machine, operated by a completely staffed auxiliary organization, was available to us, because we were approaching the capacity of one tape (about 90-95% of a complete 3600 foot roll of magnetic tape on the MTA-2 accessory was used for file storage) and further expansion of the file would require the routine use of two tape units and consequent change in program, and because the Bendix computer was needed for full time research use.

REFERENCES

[1] Kennett Computer Consultants, Inc., 808 Memorial Drive, Cambridge 39, Mass.

[2] 222 Priscilla Lane, Aldan, Pa.

[3] Benger Laboratory, E.I. du Pont de Nemours & Co., Waynesboro, Va.

[4] Dinwiddie, S. W. and Conrad, C. C., "Report Indexing by Hand-Sorted Punched Cards," in "The Technical Report," edited by B. H. Weil, Reinhold Publishing Corporation, New York, N. Y., 1954, pp. 303-16.