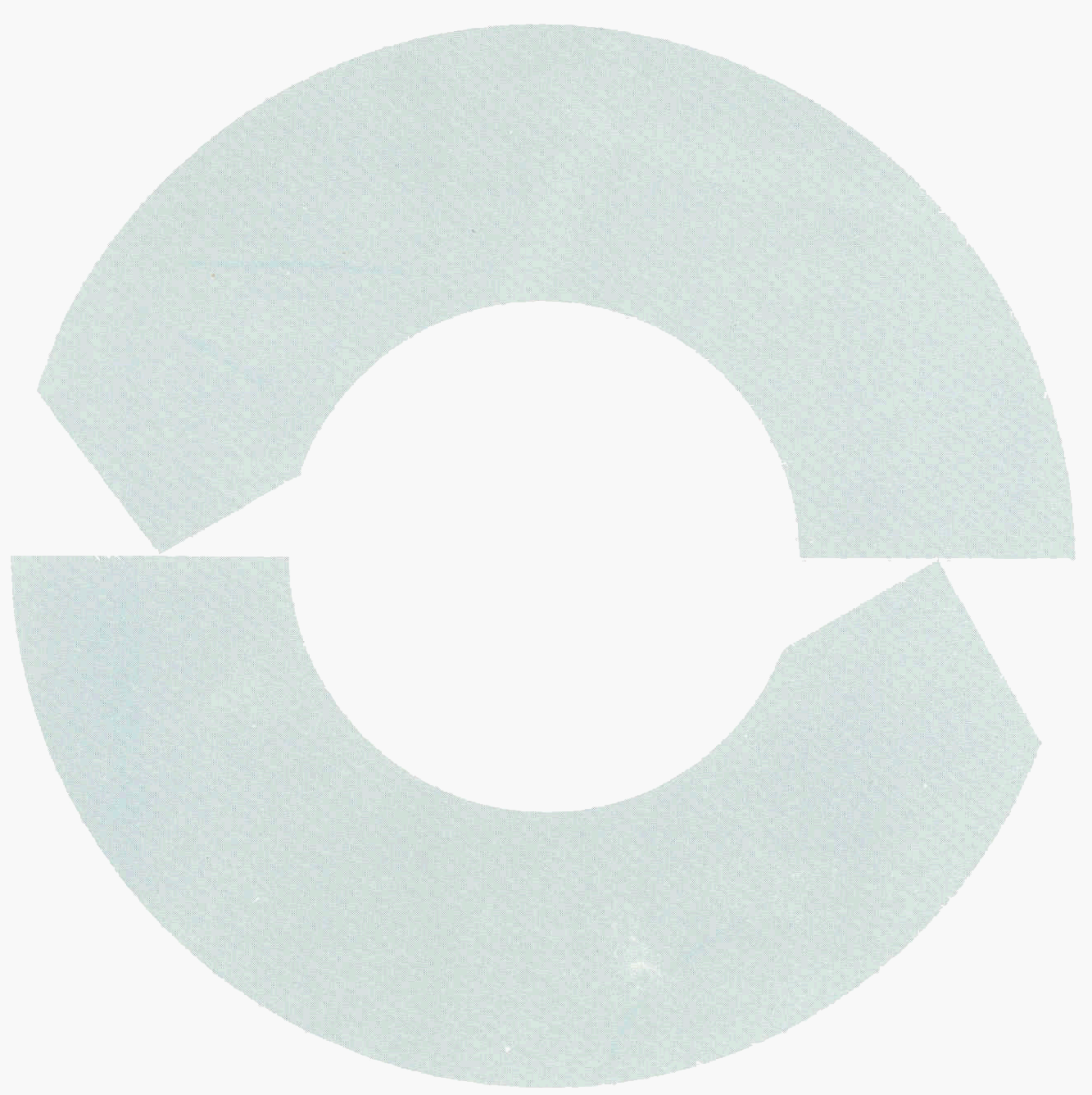
AN INFORMATION RETRIEVAL SYSTEM FOR RESEARCH FILE DATA

GENERAL
TECHNICAL
REPORT
FPL17

1978

FOREST PRODUCTS LABORATORY
FOREST SERVICE
U.S. DEPARTMENT OF AGRICULTURE
MADISON, WISCONSIN



PREFACE

One of the problems with information -- be it raw data, memos, correspondence, or finished reports -- is how to file it so it can be easily found at a later date. This report provides one method of doing this task. It provides for simple, numerical indexing of the material, utilizes a computer program that provides multiple cross- indexing of all information, and eliminates direct interaction of the individual using the system with the complexity of the computer operation. In essence, the system provides for simple filing, computer retrieval, and minimal operator involvement.

AN INFORMATION RETRIEVAL SYSTEM FOR RESEARCH FILE DATA

BY

JOAN E. LENGEL, [1]
COMPUTER SPECIALIST
AND
JOHN W. KONING, JR.,
FOREST PRODUCTS TECHNOLOGIST

Forest Products Laboratory [2]
Forest Service
US. Department of Agriculture

ABSTRACT

Research file data have been successfully retrieved at the Forest Products Laboratory through a high-speed cross-referencing system involving the computer program FAMULUS as modified by the Madison Academic Computing Center at the University of Wisconsin. The method of data input, transfer to computer storage, system utilization, and effectiveness are discussed. Preliminary results indicate that the system is readily usable by either on-line or batch processing, resulting in minimum return of results of about $^1/_2$ hour and a maximum wait of 24 hours. The vocabulary method of single keywords is used with certain modifications to improve the precision of the recalled data.

INTRODUCTION

In recent years the computer has been widely used for the retrieval of information, particularly published material. Large bibliographic programs such as AIDS (Abstract Information Digest Services), NTIS (National Technical Information Service), AGRICOLA (Bibliography of Agriculture), COMPENDEX (Engineering Index), and CA CONDENSATES (Chemical Abstracts) are all readily accessible and provide improved literature searching of published information. Smaller files, such as scientists' personal literature collections, have also been successfully computerized (2) [3] and the implications for libraries of these personal information systems discussed by Burton (1).

With the development of these various information retrieval systems, problems with data handling have been studied, particularly the form of data storage and the strategies used for recovering these data. An example of the pros and cons of two systems of vocabulary control are outlined by Lancaster (4), in which he discusses both the uniterm or single keyword index and the controlled vocabulary index. One problem with vocabulary control is that of consistency in indexing, and Landau (5) points out that most inconsistencies in indexing are due to the indexer rather than the index language; thus one solution is to use a uniterm system.

What is needed is a simple system to retrieve small quantities of specific file data quickly by relatively untrained individuals. This report describes a successful method of handling and retrieving file data on pulp and paper research at the Forest Products Laboratory, and covers the problems and solutions encountered in arriving at the final procedure. The study, begun in 1974, was conducted as much to develop an easy, usable retrieval system for filed data as to gain experience in the input and output handling of data and utilization of the FAMULUS (3) program for information retrieval, The experience gained in this limited study should help in any future approaches to information handling in this and other laboratories.

[1] Now Information Systems Engineer at Mitre Corporation, McLean, Va.

[21] Maintained at Madison, Wis, in cooperation with the University of Wisconsin.

[31] Numbers in parentheses refer to Literature Cited at the end of this report.

PROCEDURE

Input

The data for this investigation were selected, from the information available on approximately 750 research studies conducted at the Forest Products Laboratory between 1948 and 1974. These studies included work conducted by a number of scientists. The information is retained in folders filed numerically, and range from raw data to finished reports.

One unique aspect of this information was that a summary sheet, known as a job closure, was written for each study. This closure contains the project in which the work was carried out, a file number, the date, the title, the objectives of the study, an abstract of the results, the title of any reports prepared, recommendations for further work, and lists the study leader, and project leader. Because of the job closures, it was a simple matter to keyword information about each study. In this investigation keywords are defined as words in the text of the job closure, plus words assigned by the indexer. No attempt was made to go into the raw data of every study to pull out keywords. Though this would have been possible, it would have greatly increased the amount of work.

Input Transfer

The information that was considered data for the system was: (1) file number, (2) author, (3) title, (4) publications, and (5) keywords.

The first problem encountered was how to handle the keywords, regardless of whether they came from the single sheet closure statement or the raw data in the file.

A primary effort was made to expand the number of keyword?, and sufficient storage was obtained by not including an abstract of the results. Anyone identifying the folders that contained information of interest could merely pick up the folder and read the full abstract -- a situation that should be common for most in-house, filed data. This approach of significantly increasing the number of keywords reduces one drawback of uniterm vocabulary indexes -- that of low precision of raw output.

The next problem was to develop a procedure for encoding the information. Initial attempts included having the indexer manually record the data on 80-column computer coding sheets, but completing 25 folders proved this to be economically impossible.

A more successful attempt involved the indexer recording the information on a conventional dictaphone. The dictated information was then typed and this rough copy checked against the study closure. The material was then retyped, using a special format and optical character recognition (OCR) equipment, and transferred to magnetic tape. The tapes were fed into the computer and the data converted for use by the FAMULUS program.

Reviewing these procedures indicates that a trained indexer could review both the job closure statement and the raw data, and dictate the information so it can be typed directly to OCR format for transfer to tape. The transferred input is in tape form; the tape will be mounted at the time a search is requested if it will only be used intermittently.

Computer Program

The program used for the information and retrieval systems was the latest version of the FAMULUS program modified at Madison Academic Computing Center (MACC) at the University of Wisconsin. FAMULUS wasoriginally developed and made available to MACC by the Pacific Southwest Forest and Range Experiment Station, Forest Service, USDA, but has been extensively updated by MACC. The system was not designed to be interactive, though it can be used on-line. The most recent enhancement is the addition of an interactive preprocessor to verify syntaxic correctness of search formulas. (FAMULUS makes no judgment as to the appropriateness or accuracy of the keywords themselves.) A sample search is listed in Appendix A.

After data transfer, all the keywords were printed alphabetically and corrected for spelling errors and elimination of plurals and “ing” forms. Corrections were made through commands of the FAMULUS program specifically designed for major corrections. The FAMULUS program also provides a new alphabetic listing of keywords, and an indication of the citations containing the keyword. This listing is called the Vocabulary (Ap pendix B).

System Utilization

To use the information retrieval system two methods are possible -- on-line or batch. The advantage of on-line, of course, is immediate response, whereas the batch approach greatly reduces the cost. Regardless of whether on-line or batch method is used, the search operator chooses the keywords that will cover the subject and then formulates a search request. The Boolean operators AND, OR, and NOT may be used in conjunction with the keywords so a specific question can be answered. Obviously the search operator must put the keywords together to state specifically what he desires, but the description must be broad enough so no citations are missed.

The uniterm system provides maximum flexibility for the operator, but it is important to put the appropriate words together. To help in this selection an alphabetical listing of keywords, preceeded by the number of citations in which the keyword is used, was generated using the FAMULUS computer program. This helps the searcher formulate the search statement more precisely. The number of citations gives a clue to the searcher as to how definitive to be so that an overwhelming number of citations are not selected.

For example, if one is interested in work done on “aspen bark,” he can search the file for “aspen.” However, the vocabulary (Appendix B) indicates that 135 studies involve “aspen” whereas only 43 involve “bark,” so “bark” would be a better selection. However, the searcher could narrow it down even further by requesting only those studies that involve both “aspen” and “bark,” and this would cut the number of studies to less than 43.

The searcher can specify a number of output formats. One format simply provides a numerical count of the number of references that contain the keywords requested. Another format provides all the information stored about any reference that satisfies the search request. A third format generates a list of the file number, author, title, publications, or keywords or any combination of these fields.

System Effectiveness

Several considerations determine the effectiveness of an information retrieval system. The completeness and accuracy of the extracted data are prime concerns. The convenience, ease, and conciseness with which the user employs the system is also important, as well as the timeliness of system response.

The user must be aware of the scope of the data base being used. He must also evaluate the appropriateness of the data for his use separately from evaluating the effectiveness of the retrieval system. In the research studies providing data for retrieval purposes, the scope of the data base is complete and clear.

Measuring the completeness and accuracy of the extracted data depends somewhat on the user. The goal is to extract all the data, and only the data, that the searcher would have selected by direct inspection. The use of the FAMULUS - generated Vocabulary list of keywords in the search strategy decreases the number of false drops, yet unless there is a logic error in formulating the search strategy, relevant citations will not be eliminated.

The man-&stem interface problem is not so severe in batch processing systems. The cost of computer time is no longer high relative to the cost of user’s time, and on-line systems are more common. We see two alternatives--on-line or batch. As long as the user does not have to sit through the entire on-line session and has some one familiar with the system to guide him--waiting times do not seem long. Interactive procedures normally require that random-access devices be used. On-line storage is not usually warranted for the small volume of requests typical of a document retrieval application. It is certainly not warranted in our situation. Instead, cheaper bulk storage devices (tapes) are used and a batch of requests is processed against the whole file.

APPLICATION

Preliminary application of this research indicates that this system for storing and retrieving both published and unpublished file data will work, but the material must be in an organized file system such as numerically indexed folders. Precision of the search was improved by using keywords from both the title and text. Precision was also enhanced by using numerical indications with each keyword to show the number of citations for a given keyword.

Advantages of this uniterm system are: (1) Less highly trained keyword indexers are required; (2) indexes are not revised, only the vocabulary updated; (3) productivity of the indexer was high; and (4) costs of input were low.

Disadvantages are that precision is based primarily on the ability of the searcher to properly select keyword sequences, and additional screening of output may be necessary.

Initial use of the system by staff members has been favorable. Two types of searches have been performed successfully: One for answers to a specific question, and one dealing with a comprehensive area of interest. Turnaround time using an interactive terminal for the specific question was less than 30 minutes and cost less than $7.00 (in 1975). Turnaround time using batch processing and off-prime time rates for the comprehensive search was less than 24 hours and cost less than $4.00. Precision of these two searches and others conducted so far by the staff appear satisfactory; additional screening of the output by the search recipient has been simple and not very extensive.

The file has been updated by placing the information from the job closure directly on punch cards and then transfering the card data to tape. The results indicate that direct card input is possible and desirable with small amounts of input data. Further refinement of this technique is under way. We hope all data can be taken directly from the source document and punched on cards for transfer to tape file.

CONCLUSIONS AND IMPLICATIONS

Adaptation of the FAMULUS Information Retrieval program has resulted in a system that provides better access to file data covering 25 years of research. The usability of the system to date has been judged satisfactory and it has been possible to update the file through direct punch card input. Thus, a simple method of information retrieval for small file systems has been developed that could be extended to retrieval of reports and data from large file systems and even small libraries.

LITERATURE CITED

1. Burton, H.D.
   1973. Personal information systems implications for libraries. Special Libraries, Vol. 64 (1) Jan., p. 7-11.

2. Burton, H.D., R.M. Russell, and T.B. Yerke.
   1969. FAMULUS: A computer-based system for augmenting personal documentation efforts. U.S. Dep. Agric., For. Serv. Res. Note PSW-193. Pacific Southwest Forest and Range Exp. Stn., Berkeley, Calif.

3. FAMULUS - Reference Manual for the 1110.
   1975. Madison Academic Computing Center, University of Wisconsin, Madison. 89p.

4. Lancaster, F.W.
   1972. Vocabulary control for information retrieval. Information Resources Press, Washington, D.C. 233 p.

5. Landau, H.B.
   1969. A cost analysis of document surrogation: A literature review. American Documentation, Vol. 20, No. 4.

☆ U.S. GOVERNMENT PRINTING OFFICE. 1978-750126/34

APPENDIX A--SAMPLE SEARCH

```
@asg, a 10.kf
READY
@macc*fam.famulus2
FAMULUS2-16
*Search 10

xxxxxxxxxxxxxxxxxxxxxx
SEARCH
 FILE PARAMETERS:
input      10

/ID/  Koning
INPUT  CARD  ...  /ID/     KONING
/fields/8.   (keys)


IDENTIFICATION OF INPUT FILE
KONING
EDITION: 19E
FIELDS   ARE   NUMB   AUTH   TITL   PUBL   DATE   KEYS   D1   D2   D3   D4

     THE DESCRIPTOR FIELDS ARE AUTH KEYS

INPUT CARD.../FIELDS/ (KEYS)
/format/   (edit,numb-titl,v,1)
INPUT   CARD.../FORMAT/   (EDIT,   NUMB-TITL,   V,   1)
/search/  (bark  &(hardboard  $  insulate))
INPUT CARD.../SEARCH/ (BARK & (HARDBOARD $ INSU LATE))
* * *SEARCH PACKET NO.          1  ***

@eof

* * *  SEARCH PACKET NO.          1 ***

77    NUMB    1069
     TITL    EVALUATION OF DENSE BUILDING BOARDS BUILDING PAPERS AND
             PAPERBOARDS FOR HOUSING FROM WOOD WASTE (PULPS)

703    NUMB    1744
        TITL    KRAFT AND MECHANICAL PULPING OF YOUNG, RAPID- GROWTH SYCAMORE
                TREES



2 RECORDS SATISFY THIS SEARCH.

*end

* * END FAMULUS * * *
```

APPENDIX B--VOCABULARY

1	AQUARIUS
2	AQUEOUS
2	ARBOREA
1	ARC
13	AREA
4	ARGENTINA
1	ARID
1	ARIES
9	ARIZONA
2	ARKANSAS
1	ARMOUR
8	ARMY
1	ARRHENIOUS
1	ARROW
4	ARSENAL
1	ARTHUR
1	ARTICLE
1	ARTIFICIAL
4	ASBESTOS
32	ASH
1	ASHLAND
1	ASHTON
135	ASPEN
5	ASPHALT
21	ASPLUND
1	ASSESS
5	ASSOCIATES
18	ASSOCIATION
1	ATLAS
15	ATMOSPHERIC
1	ATOMIC
2	ATTACK
1	ATTAINABLE
21	ATTRITION
2	AUSTRALIA
1	AUTO- OXIDATION
1	AUTOCATALYZE
1	AUTOCLAVE
1	AVIATION
2	AXIS
1	AYACAHUITE
1	B
2	B.
1	B-F-D COMPANY
4	B-FLUTE
1	BACK
22	BAG
7	BAGASSE
1	BAKELITE
3	BALL
1	BALLOON
19	BALSAM
1	BALSIMEA
2	BAMBOO
1	BANGOR
1	BANKERS
1	BANKSIANA
43	BARK
1	BARRANQUILLA
1	BARREL
4	BARRIER
60	BASE
2	BASSWOOD
1	BATAAN
6	BATCH
4	BATTERY
33	BAUER
13	BAY
1	BAYONNE
1	BC
1	BEAN
2	BEAR
1	BEARINGS
82	BEAT
1	BED
20	BEECH
4	BEETLE
3	BEHAVIOR
1	BEKK
1	BELIZE
1	BELOIT
1	BELT
2	BELVOIR
9	BEND
1	BENDIX
1	BENGUET
1	BENTONITE
2	BENZENE
1	BERGSTRUM
1	BERLIN
2	BETA
2	BETULA
1	BETULOIDES
1	BIAXIS
1	BIBLIOGRAPHY
8	BICARBONATE
3	BIDWELL
2	BIGLEAF
16	BIND
1	BIOCONVERSION