Computer Documentation System for Small- and Medium-Sized Information Collections

P. J. LEWI and W. W. BRAET
Scientific Data Processing Department, Janssen Pharmaceutica,
Research Laboratoria, Beerse, Belgium

Received November 10, 1969

A computer service was designed for use in scientific and engineering communities of a pharmaceutical company. Alphabetic information is divided into main subjects, entered in free-format on punched cards, and transferred to magnetic tape. The master tape can be searched on any subject by strings of word fragments and Boolean operators. The program was written in Fortran IV for an IBM–1800 computer installation.

This paper describes how information processing problems in a Research and Development division of a pharmaceutical company have been solved by computer in a simple and straightforward manner.

The system designer, exploring the needs for information and documentation in a scientific and technical community, is usually confronted with a great variety of requirements.

As Altmann[1] has recognized, this is mainly due to R & D people working in small teams on highly specialized subjects for varying periods of time. Some groups operating separately, will sometimes be served best by an individual and manually operated tool, such as a catalog file, peak-a-boo, or edge-notched cards, the usefulness of which is discussed in detail by Kent[4].

In other cases, a personalized computer system, using numerically coded descriptors and a few levels of subordination, such as described by Gillis[2], can be very helpful.

We selected English sentences as input, using the free-format recording method of Korein[5]. The retrieval language consists of English words and Boolean operators, as in the method employed by Heaps[3].

Our system however can be used equally well by those who are more familiar with hierarchical classification, tagged descriptors, or subject headings. Although a natural language system with random ordering has been rated low by Meadow[6], it can be used advantageously in a diversified R & D community, provided that the size of the collections can be kept between 5000 and 10,000 records. The method described here stands halfway between a keyword extraction system and the processing of languages with semantic, syntactic, and statistical analyses, discussed by Salton[7] and Simmons[8].

Our principal objectives were to encourage user responsibility for the performance of the system and to ensure easy formulation of requests and output specifications. Furthermore, we were required to prepare indexes by various subjects, such as those reported by Teal[9].

In a generalized information system, performance is often measured in terms of precision and recall. In a personalized service, these measures are not meaningful, as the precision will generally be very high. Also the recall ratios can be anything between 0 and 1 depending on the effort and endurance of the individual who assumed responsibility for the system. Therefore, we propose a measure that could be termed the utility ratio. If a bibliography for a scientific paper were compiled, the utility would be the number of references derived from the computer, divided by the total number of references (including those obtained from colleagues, personal library, or other services). The utility reflects the user's satisfaction or dissatisfaction and also his willingness to continue or discontinue the documentation service.

DESIGN OF THE SYSTEM

The system assumes that information derived from a document can be divided into a limited number of independent subjects. In a chemical literature retrieval system, we would have the subjects: author names, title, source, abstract, and chemical names.

The number of alphanumerical entries—e.g., names, keywords, phrases—covered by a subject is only limited by the field length of the input record.

For practical reasons (computer memory), we limited the number of subjects to five and assigned a maximum of 10 fields with 74 characters to each subject.

Since 80-column tab cards are used as primary input medium, each field of 74 characters was made a separate transaction. Information is entered on the cards in free format, and a transaction number (columns 79–80) defines card sequence and implicitly identifies the subject (e.g., 01–10 for subject 1, 11–20 for subject 2, etc...).

Since a field length of 74 characters is also acceptable for printouts, there was no need for text editing.

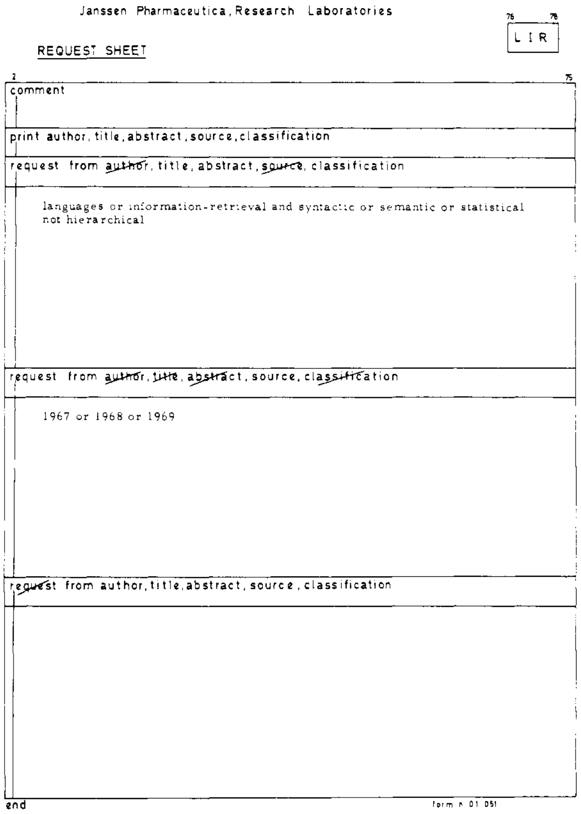
However, since a blank character determines the end of a word, special attention must be paid to word segmentation at the end of a card field. In processing of alphanumerical data, some difficulty is bound to arise with punctuation marks and interspersed blanks (the word A.B-C. might not be different from ABC).

This problem was solved by programmed elimination, during word-matching operations, of all symbols that are different from numerals and alphabetics.

An accession number is punched in a header card. Header card and subsequent transactions are loaded to tape and sequenced in alphabetic order by the nine first characters appearing in an assigned subject—in literature retrieval systems this subject will normally be the author names section (the Literature Cited section of this article was compiled by computer).

When the assignment of a key subject is omitted, a file sequenced by accession number will be obtained.

All reports and results from searches are produced from this master tape, and this limitation required an allowance for some flexibility on the output. The request specification contains a section for printer output (Figure 1). Any combination of the defined subjects can be specified for output. If none is specified, only a count of the retrieved records will be obtained. A request statement is made up from a string of descriptors separated by the Boolean operations AND, OR and NOT. Parentheses are not used, meaning that only a single level of subordination is available. The OR relation is given priority over the AND relation. Negations are by convention AND-ed to the previous expression. The statement A AND B OR C NOT D is in our system to be understood as A AND (B OR C) AND (NOT D).



Janssen Pharmaceutica, Research Laboratories

76 78
L I R

REQUEST SHEET

2 75

comment

print author, title, abstract, source, classification

request from ~~author~~, title, abstract, ~~source~~, classification

languages or information-retrieval and syntactic or semantic or statistical
not hierarchical

request from ~~author~~, ~~title~~, ~~abstract~~, source, ~~classification~~

1967 or 1968 or 1969

~~request~~ from author, title, abstract, source, classification

end

form n 01 051

Figure 1. Request specification

During word-matching of the request terms with the master tape, it may be necessary to neglect prefixes, suffixes, or both.

These options are defined for each subject during initial program setup—e.g., if compound names are searched for the fragment "butyro" then we expect a match on the name "4-fluoro-butyrophenone"). A request statement is directed specifically to any combination of the previously defined subjects. In the present system, it is possible to AND up to five different statements—e.g., request the name "X" from author section, the year "1969" from the reference, and the fragment "butyrophenone" from compound name section.

Requests are also entered in free format with identical layout as used for data entry, except for the first column which contains a control character for the different specifications (comment, output, request statement) (Figure 2).

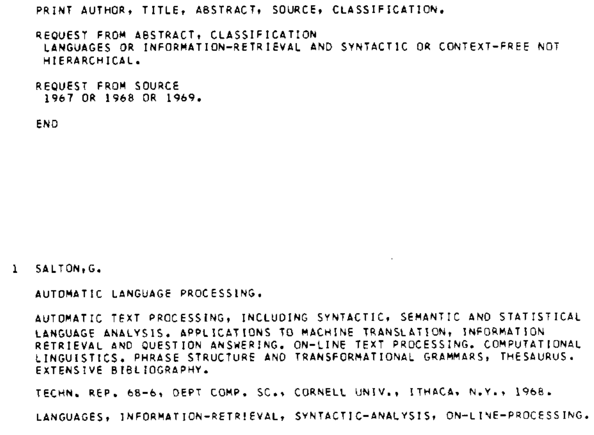
The output layout was designed in such a way that computer-related details (such as transaction and accession numbers, alphabetic sort field) can be easily separated from the relevant information.

PROGRAMMING AND OPERATIONAL CHARACTERISTICS

The minimum computer configuration requires two tapes and a work area on a random access device for sort and merge operations (Figure 3). When sequential arrangement of input records is sufficient, the system might even run on a single tape drive system and no random access storage is required.

Assuming an average of 15 transactions (of 74 characters each) per input record, a single 2400 feet tape reel (800 bpi, 9 track) will hold about 15,000 records.

Programming of the system was simplified by the adoption of similar layouts for card, tape, and printer transactions. In core memory, these transactions are read both



```
   COMMENT  SAMPLE REQUEST, OCT 19,1969

   PRINT AUTHOR, TITLE, ABSTRACT, SOURCE, CLASSIFICATION.

   REQUEST FROM ABSTRACT, CLASSIFICATION
    LANGUAGES OR INFORMATION-RETRIEVAL AND SYNTACTIC OR CONTEXT-FREE NOT
    HIERARCHICAL.

   REQUEST FROM SOURCE
    1967 OR 1968 OR 1969.

   END




1  SALTON,G.

   AUTOMATIC LANGUAGE PROCESSING.

   AUTOMATIC TEXT PROCESSING, INCLUDING SYNTACTIC, SEMANTIC AND STATISTICAL
   LANGUAGE ANALYSIS. APPLICATIONS TO MACHINE TRANSLATION, INFORMATION
   RETRIEVAL AND QUESTION ANSWERING. ON-LINE TEXT PROCESSING. COMPUTATIONAL
   LINGUISTICS. PHRASE STRUCTURE AND TRANSFORMATIONAL GRAMMARS, THESAURUS.
   EXTENSIVE BIBLIOGRAPHY.

   TECHN. REP. 68-6, DEPT COMP. SC., CORNELL UNIV., ITHACA, N.Y., 1968.

   LANGUAGES, INFORMATION-RETRIEVAL, SYNTACTIC-ANALYSIS, ON-LINE-PROCESSING.
```

Figure 2. Sample output from search request

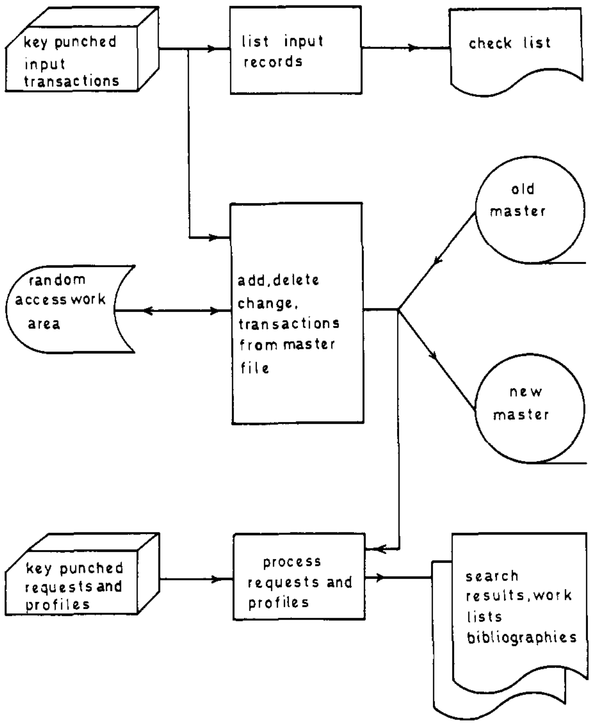

key punched input transactions
list input records
check list
old master
random access work area
add, delete change, transactions from master file
new master
key punched requests and profiles
process requests and profiles
search results, work lists bibliographies

Figure 3. Block diagram of computer procedures

from core or tape into a single field of 25 × 74 characters. Access to a given subject of the record is obtained indirectly via a branch table which contains the transaction numbers. This technique has the advantage that the entire field can be scanned in a continuous way, without being hindered by possible word segmentations at the end of a transaction field.

Once the limits between which the input field is to be scanned are determined, a program segment retrieves the next term (if any) from this subfield. Comparison between the retrieved term and the terms of the request statements is performed after elimination of all special characters.

Depending upon the subject chosen, the matching routine uses suffixing, prefixing, or a combination of both rules and the result is entered into a truth/false table.

Request processing, tape update, and list options are grouped into a single main program. Each option is selected by appropriate control cards. The total length of the program is about 8K words (16 bits), not including the disk and tape utility programs. Grouping of program segments can result in core overlays of less than 4K, so that the system can be made operational on machines of limited core memory size.

The program was initially coded and executed in FORTRAN IV. The most critical segments (word matching and elimination of special characters) were rewritten in assembly language.

DISCUSSION

The main advantages of the system described here were the free format, entirely alphabetic, data input and request formulation. Experienced keypunch operators had little difficulty in preparing and verifying the input cards. The one-to-one correspondence between data records on card, tape, and printer together with the single file concept simplifies maintenance and updating procedures of the master file. Although the present system does not pretend to compete in speed and versatility with larger sophisticated programs, the authors have found that the present system can solve a large number of needs arising in a scientific community. Practical usage of the system is limited, however, to smaller data collections, with an upper limit of 10,000 records of an average size of 1000 characters each.

SUMMARY

A computer system is described, designed for data collections of up to 10,000 records and to be run on a computer with limited available core storage and peripheral equipment.

The design of data input, output, and request formulation was generalized to an extent that the system can be tailored to individual needs. Free format, alphabetic fields, single record layout, and alphabetic word matching combined with Boolean relations are its main characteristics.

The system can be used with advantage as an intermediate stage between manually-operated and large-scale document retrieval systems.

ACKNOWLEDGMENT

Assistance during the preparation of the manuscript by S. Van Gestel is gratefully acknowledged.

LITERATURE CITED

1 ALTMANN,B.
A NATURAL LANGUAGE STORAGE AND RETRIEVAL (ABC) METHOD ITS RATIONALE,
OPERATION AND FURTHER DEVELOPMENT PROGRAM.
J. CHEM. DOC., 6, 154-157 (1966).

2 GILLIS,C.N.
BIOMEDICAL INFORMATION RETRIEVAL.
A COMPUTER-BASED SYSTEM FOR INDIVIDUAL USE.
J. CHEM. DOC., 7, 98-100 (1967).

3 HEAPS,H.S.
BOOLEAN, FRACTIONAL AND ASSOCIATIVE SEARCHES ON TRUNCATED TITLE WORDS.
PROC. AM. SOC. INFORM. SC., 5, 179-184 (1968).

4 KENT,A.
CHAPTER 7, P. 172-182, IN 'SPECIALIZED INFORMATION CENTERS',
SPARTAN BOOKS, WASH. D.C., 1965.

5 KOREIN,J., GOODGOLD,A.L. AND RANDT,C.T.
COMPUTER PROCESSING OF MEDICAL DATA BY VARIABLE-FIELD-LENGTH FORMAT.
J. AM. MED. ASS., 186, 132 (1963).

6 MAEDOW,C.T.
CHAPTER 5, P. 169-173, IN 'THE ANALYSIS OF INFORMATION SYSTEMS', JOHN
WILEY AND SONS, INC., N.Y., 1967.

7 SALTON,G.
AUTOMATIC LANGUAGE PROCESSING.
TECHN. REP. 68-6, DEPT COMP. SC., CORNELL UNIV., ITHACA, N.Y., 1968.

8 SIMMONS,R.
ANSWERING ENGLISH QUESTIONS BY COMPUTER, A SURVEY.
PART III8, IN 'THE GROWTH OF KNOWLEDGE', M. KOCHEN, ED., JOHN WILEY AND
SONS, INC., N.Y., 1967.

9 TEAL,T.W. AND GREENBERG,S.M.
MANAGING LITERATURE IN THE PHARMACEUTICAL INDUSTRY.
DRUG INFORM. ASS. BULL., 2, 136-143 (1968).