retrieval alone, no matter how good. The idea of a compilation system is not profound, or new, or revolutionary, but at least the data compilation gets inside the document and works with the information directly. Any improvement that can be made in the availability, quality, and completeness of data compilations will be a positive step in the direction of true information retrieval.

BIBLIOGRAPHY

(1) A Directory of Continuing Numerical Data Projects, Office of Critical Tables, National Academy of Sciences-National Research Council, Washington, D. C., August, 1960.
(2) Directory to Nuclear Data Tabulations, R. C. Gibbs and K. Way, 185 pp., 1958; Supplement, 1959 Nuclear Data Tables, pp. 1-38; both U. S. Government Printing Office, Washington, D. C.
(3) *Journal of Chemical and Engineering Data.*
(4) See *Physics Today*, **14**, No. 10, p. 70 (August, 1961).
(5) K. Way, N. B. Gove, and R. van Lieshout, "Waiting for Mr. Know-It-All (or Scientific Information Tools We Could Have Now)," *Physics Today*, **15**, No. 2, p. 22 (February, 1962).

Library Applications of Permutation Indexing

By R. A. KENNEDY
Bell Telephone Laboratories, Incorporated, Murray Hill, N. J.
Received May 24, 1962

Within recent years interest has been growing in the possibilities of delegating to computers not only functionally simple information storage and retrieval operations, but also certain complex intellectual tasks such as determining the content of a document and labelling it in some systematic maner for future recall. In addition to automatic indexing, machine preparation of abstracts also has aroused close attention. The technical feasibility of performing at least some part of these human tasks with the aid of data processing equipment has been demonstrated. However, certain conceptual and economic problems remain to be worked out before the library administrator can count on his cataloging section becoming a unit of a computation center.

While automatic indexing in any interpretative and analytical sense is therefore not yet a practical matter, a simpler mode of machine indexing is coming into wide use. This is the technique known variously as permutation indexing, permuted title indexing, subject-in-context indexing or, as H. P. Luhn calls it, keyword-in-context indexing (KWIC). Whatever the name for the process, current practice was primarily stimulated by the publication in 1958 and 1959 of reports by Ohlman, Hart and Citron of the System Development Corporation[1] and Luhn of IBM.[2] Differing in method and product, both of these approaches pointed up the practicality of compiling subject indexes automatically, or largely so, from key-punched titles or other parts of documents. To oversimplify, merely by mechanically shifting a significant word to a fixed filing position, surrounding the word with some of its original context, and listing the whole in an alphabetical array, a highly useful index could be got.

Since that time well over thirty applications of the fundamental technique appear to have been made. In a number of cases it has been taken up as a very satisfactory all-round way of combining mechanized listing and swift indexing in a current announcement package. While certain observers have fastened upon the obvious limitations of the approach and dismissed it out of hand, the method is being increasingly adopted either for single uses or regularly scheduled purposes. Among the uses made in the last several years are: *Chemical Titles*, the American Chemical Society's semi-monthly publication covering some 600 journals of pure and applied chemistry; *Bibliography of Chemical Reviews*, a selection of abstracts from *Chemical Abstracts; BASIC*—a title index published in each semi-monthly issue of *Biological Abstracts*; the *KWIC Index* to the *Science Abstracts of China*, issued December 1960 by the MIT Libraries and covering some 3300 Communist Chinese scientific papers; the *KWIC Index to Neurochemistry*, prepared in 1961 by the Mimosa Frenk Foundation for Applied Neurochemistry in cooperation with IBM and covering some 2100 papers; *Dissertations in Physics*, an indexed bibliography of the 8418 doctoral theses accepted by American universities from 1861 to 1959, compiled by the IBM San Jose Laboratory and published by Stanford University Press, 1961; and at the Bell Telephone Laboratories, a number of applications which are the particular concern of this paper.

It may be of interest to note some of the considerations entering into the decision, in 1959, to try permutation indexing. First, the Libraries had been directed to provide Laboratories personnel with more adequate means of access to internal reports than available through the existing non-library facilities. Secondly, we wanted to get a first-hand, working appreciation of the methods and problems of producing relatively simple indexes by punched card and computer processes. Permutation indexing was considered worth experiment under the circumstances because, among other reasons:

1. The report titles to be processed were well endowed with words of indexing and retrieval significance to their author-users. An average of 5 to 6 subject tags designating physical and chemical processes, elements, materials, systems, equipments and so on would be provided for each report indexed.

2. The additional essential search tools—author index, case or project number index, and a listing of reports by the issuing department—would be provided automatically, as adjuncts to the permuted title index.

3. User requirements for both a current awareness bulletin and a retrospective, multi-aspect search facility, could be handled by the one system.

4. An index in book form would be produced. Apart

from the unique attractions of a printed catalog, the report catalog could now be delivered to the desks of the users who, at twenty laboratory locations in ten states, were hitherto removed from the existing central card file by anywhere from several hundred feet to 7,000 miles. Further, for access to personal collections of the technical reports, a ready-made catalog would be at hand.

5. Production of the four-part index could be done on a routine, essentially clerical basis. Continuous or periodic cumulations would be largely mechanical. Extension of the process from the initial handling of several hundred company reports per month to a possible 1000-2000 other items per month could be readily accomplished, it appeared, and would require little addition to what would be, in all, a very small human indexing effort.

The Program.—To do the job a machine program had to be prepared; all the features held to be desirable or essential in the index were not furnished by any other program at the time. A processing routine was accordingly written for the IBM 7090 computer. The program, known as BE-PIP and available to members of the SHARE

and thereby aid discriminating search, a permuted line length of 120 characters is used. After allowance for the item number (up to eleven digits) and dedicated blanks, the keyword-in-context line totals 105–106 characters as compared with the common maximum of 60 characters.

A second device used to provide maximum context for each index word is recirculation, snap-back or wrap-around. Progressive shifting of a title to bring each significant word to the indexing column frequently causes portions of the title to exceed the line space available. To avoid or minimize this loss of context, the computer is instructed to fit into what would otherwise be unused line space as much as possible of that portion of the title which would be shifted past line boundaries. Although this device is less necessary in the 120-character line than in the shorter permuted line, it has been incorporated in the program from the beginning and found to be very useful. The portion of the title which is wrapped around, it may be of interest to note, is always contiguous to the index word; no disjointed segments from other areas of the title are used (Fig. 1).

```
                                  HYDROTHERMAL GROWTH OF OXIDE CRYSTALS.                                  T2-116-82
 THIN TANTALUM FILMS.           THE EFFECTS OF NITROGEN AND OXYGEN ON STRUCTURE AND ELECTRICAL PROPERTIES OF T2-262-6
CLEAN NICKEL.                                            OXYGEN- NICKEL STRUCTURE ON THE ( 110 ) FACE OF  P2-111-70
                         LIGHT EMISSION FROM FORWARD-BIASED P-N JUNCTIONS IN GALLIUM PHOSPHIDE.           P2-115-39
R(15).                        THE CRYSTAL STRUCTURES OF PALLADIUM(17) SELENIUM(15) AND RHENIUM(17) SULFU T2-111-67
ILUTE PARAMAGNETIC IMPURITIES ON THE EPR OF GADOLINIUM IN PALLADIUM.                    EFFECT OF D T2-111-61
 IN PALLADIUM.                          EFFECT OF DILUTE PARAMAGNETIC IMPURITIES ON THE EPR OF GADOLINIUM T2-111-61
HLORIDE AND CHROMIUM BROMIDE.           FERROMAGNETIC AND PARAMAGNETIC RESONANCE LINE WIDTHS IN CHROMIUM C P2-115-42
D Z CUT QUARTZ AND THEIR RELATION TO THERMALLY DETERMINED PARAMETERS. /DINAL WAVES IN SILICON GERMANIUM AN T2-121-30
                                  A LOW-NOISE M-TYPE PARAMETRIC AMPLIFIER.                                T2-124-22
                                     SEMICONDUCTOR PARTICLE SPECTROMETERS.                                P2-113-26
                        PROBLEMS IN NON-ORIENTED DIRECTIONAL PASSIVE SATELLITE REPEATERS.                  P2-124-17
TION.                      THE PROPAGATION OF DISCRETE FLUX PATTERNS IN A MULTI-APERTURED MAGNETIC CONFIGURA P2-632-2
ISE - SHORT-TIME SPECTRAL ANALYSIS BY THE EAR.             PERCEPTION OF COLORATION IN FILTERED GAUSSIAN NO T2-123-49
ECTION.  EFFECT OF THERMAL VIBRATIONS ON DIFFRACTION FROM PERFECT CRYSTALS, PART-2. THE BRAGG CASE OF REFL P2-116-68
OF SYSTEMS OF LINEAR ORDINARY DIFFERENTIAL EQUATIONS WITH PERIODIC COEFFICIENTS.                SOLUTION  P2-217-4
```

Fig. 1.—Permuted Title Index

organization as SDA No. 1239, contains some 2500 orders. A separate cumulative merge program of about 500 instructions also has been prepared. This program permits the merging of up to eight magnetic tapes retained from previous index runs to produce semi-annual, annual or other cumulations.

The basic elements of a permuted title index are now widely known. Only a few of the features of the Bell Laboratories program need therefore be noted.

To display each index word in as much of the full title context as permitted by the available equipment,

Among its other features, the program includes means to: generate project or secondary number indexes; insert predetermined cross-references in the permuted index; put headings, comments, or annotations in the bibliography section; and make temporary alterations to the non-significant word table stored in machine memory. (A sample of this list of 585 words which are not indexed is shown in Fig. 2.) If desired, the author and case or project number indexes may be output in punched card form for off-line sorting and listing; alternatively, these indexes may be sorted entirely by computer and printed

```
EVALUATING      INCLUDED        METHOD
EVALUATION      INCLUDING       METHODS
EVEN            INCOMPLETE      MIGHT
EXAMPLE         INCOMPLETELY    MINIMUM
EXAMPLES        INCORPORATING   MINOR
EXCEPT          INCREASE        MODIFICATION
EXCEPTING       INCREASED       MODIFICATIONS
EXPERIMENT      INCREASES       MODIFIED
EXPERIMENTAL    INCREASING      MODIFY
EXPERIMENTS     INCURRED        MORE
EXPLANATION     INDICATE        NARROW
EXPLORATORY     INDICATING      NEAR
FACILITY        INDUCED         NECESSARY
FACILITIES      INFLUENCED      NEW
```

Fig. 2.—Part of Non-Significant Word List

Fig. 3.—Author Index

out from tape on a high-speed printer in three automatically balanced columns per page (Fig. 3).

Applications.—Permutation indexing by computer has been put to use at the Bell Telephone Laboratories in the following ways:

1. Internal Technical Reports.—The need which stimulated trial of the whole process already has been mentioned. To this end, a monthly current awareness bulletin is produced covering about 200 new internal reports per issue. Each issue provides a departmental number (*i.e.*, broad subject) listing of the items, a permuted title index, an author index and a case number index. Semiannual and annual cumulations are provided for retrospective search. Three annual volumes have been produced and about 6500 technical reports indexed since November 1960.

2. Outside Talks and Papers.—Laboratories people contribute well over two hundred scientific and technical talks and papers to the public domain every month. To meet a need for improvements in records control, announcement and indexing of this information, an integrated, machine-based system has been set up by the Libraries. The punched cards are made at the time a talk or paper is cleared for release; at successive stages these are used to provide:

a. A monthly bulletin and permuted index of the cleared items, for the purpose of alerting staff members at the earliest practicable moment to talks and papers scheduled by co-workers so that items of interest might be followed-up immediately with the authors concerned.
b. A monthly list of *presented* talks and papers, for publication in the *Bell Laboratories Record*.
c. Routine follow-ups to authors' departments to get the details on any unreported talks and papers.
d. Several different annual indexes incorporating permuted indexes.

3. Bibliographies.—Major bibliographies compiled by the Libraries' information scientist and reference staffs are now being processed by the permutation program. When a bibliography is sufficiently long, the standard operations of typing final copy on 5 × 3 cards and shingling these for duplication are replaced by keypunching and machine handling. This approach results not only in mechanical assembly and automatic subject and author indexes; by virtue of these indexes, it also permits the option of chronological or subject-classed arrangements; additionally, it furnishes an efficient means for updating and cumulating previous editions. For example: a continuing survey of new literature in the field of magnetism is now being maintained; references are keypunched as selected; at convenient intervals portions of the expected 2000 papers per year are computer processed and issued as up-to-date, subject classed, subject and author indexed, bibliographies.

4. Specifications.—Following the success of a recent trial, Laboratories materials specifications are being recorded on punched cards for listing and indexing by the permutation program. The initial batch will cover about 3000 specifications. Precise and substantive titles, frequency of amendments and reissues, and the need for maintaining numerical cross-references lists and indexes of equivalent items—these are some of the characteristics of specifications which make the technique highly appropriate.

5. Special Applications:—To date these include indexes to Laboratories computer programs, the literature of binocular depth perception, and a military project file of memoranda, progress reports, work statements, drawings, and other documents. A 450-item bibliography which combines permutation indexing and citation indexing has also been done. Each paper listed has appended to it a chronological list of the other authors in the bibliography who cite the paper. A combined author-citation index is also included. (The significance of a given paper may to some degree, therefore, be judged by the number of citations it has. Possibly more important, the development of an idea or a subject may be traced forward in time, showing the researcher the avenues explored and suggesting others which might be taken[3].) Further exploration of this approach is planned. Additional applications of permutation indexing are being considered, including the suggestion from library users that titles of the 2600 journals in the system be permuted to provide a location key whatever the shelving rule or form of the searcher's reference.

Processing.—The steps followed in handling internal reports are typical of the routines in most applications.

1. Editorial scan of the text to be indexed. Since the punching and printing devices lack lower case characters, superscripts, subscripts, *etc.*, chemical symbols are spelled out and other conventions used for valencies, exponents, *etc.* Other editorial markings include the occasional addition of a supplementary word to the text or the alteration of punctuation as a way of causing or preventing the indexing of a word. The whole editorial operation takes only a very few seconds per item.

2. Keypunching of title, author and reference cards, as applicable. Each report takes an average of three cards

(not all 72 columns of each being used) punched in about two minutes per set.

3. Verifying, or in some cases, proofreading from a tabulator listing.

4. Card sequence numbering on the Reproducer.

5. Punching consistency checking on the IBM 101. Eleven different accuracy tests are made in the one pass.

6. Sorting of the card sets into numerical order on the 101, at a rate, counting handling time, of about 200 cards per minute.

7. Conversion of the data and program decks to magnetic tape on the 1402 Card Read Punch at close to 800 cards per minute.

8. Computer Processing in one recent run the 7090 computer generated a four-part index (bibliography, permuted subject index, author index and case number index in final order) for 2000 reports in 20 minutes.

9. Tape print-out on the 1403 printer at 500-600 lines per minute.

It may seem from this that there is a good deal of work in getting the machines to do work. What do the separate operations add up to in time spent per report listed and indexed? From initial editorial scan through the printing of copy for duplication, the sum of the times for each step has been found to average just about five minutes per item. This is as experienced, not calculated from the theoretical output rates specified for the machines involved.

Pro and Con.—Let us now look briefly at the merits and disadvantages of this form of mechanized indexing.

The two chief limitations have to do with the subject index produced. The first is this: Insofar as the title of a document falls short of conveying exactly and completely what the document is about, any index constructed solely from title words will be inadequate. The second disadvantage is related: Use of each author's own words, of all the verbal variants from paper to paper, makes for subject scatter throughout the alphabetical array.

These imperfections exist in varying degree. How critical they may be will depend upon the nature of the information indexed, the purpose of the index, the way in which it is used, its size and structure, and various other factors. These are matters which the library administrator must determine for the case in hand.

There are, of course, good and bad uses to be made of the technique. A permuted title index to much trade journal material might well be of little value in view of the relatively high percentage of articles with provocatively phrased or catchword titles. On the other hand, specification titles, as noted earlier, appear especially well suited. And many research papers have titles equivalent to one sentence abstracts, packed with 6 or 8 explicit and meaningful retrieval terms.

There are also, it is worth noting, no rules in the game which say that human contributions to mechanized indexing are illegal; that the data to be keypunched for indexing must come from the title alone. Several steps for adding to the coverage or convenience of a permuted index can be taken without compromising the essential merits of mechanization. At the editorial scan stage titles which appear on the basis of this quick inspection to be weak or unclear might be supplemented, say by marking a word or words in the document abstract for keypunching. In this connection a Bell Laboratories library practice with bibliographies may be pertinent. Now that the permutation process is routinely used in bibliography assembly, annotations are placed in parentheses, immediately after the title. The result is additional useful descriptor-like entries for the permuted index.

The second limitation—concept scatter from word variants—can be mitigated by cross-referencing. On the simplest level this might involve a linking of synonyms, or of acronyms with word entries, in the manner, SEE BOTH LASER AND OPTICAL MASER. In this example, insertion of a single punched card would cause the cross-reference to appear in the index at all three possible search points. On a more complex level, a network of stored cross-references might be called in by the program when the specified conditions, for example, some entries under both terms of a related pair, were met.

The best permuted title indexes cannot be expected to match the quality of expert human effort. However, several particular properties of the permuted index should not be overlooked. The use of the author's own terms—the alive currency of new ideas—rather than the considered reshapings to the indexing system may often be of advantage. The automatic generation as index entries of all the separate words in multi-term concepts is definitely so. Access is direct, under any one of the component words, in the unrestricted manner of uniterm indexing. And context minimizes false drops; the author has supplied the term coordination.

It also should not be overlooked that the time and talent invested in hand-crafted excellence are of a different, rarer order than the machine system needs; that skilled indexers are frequently not available for all the purposes where indexes would be highly appropriate; that when available, the normal exercise of human talent may be, for the purpose, unacceptably slow and unduly expensive. Perhaps the question should be asked seriously: In view of the evidence from several quarters that at least two-thirds of the inquiries for information might be handled by an author, title or relatively simple subject index, should not machine indexing be substituted, where at all practicable, for human indexing—freeing human talent for direct and personal search assistance when and where required?

Where the need exists to produce promptly ordered announcement lists of substantial numbers of new reports or papers; to index such items by author, subject, report number, *etc.*; to extend to information users at remote points in an organization the privilege of direct and up-to-date catalog access to information; to make periodic cumulations and revisions for an expanding and changing body of printed information; and where these purposes require to be accomplished quickly and economically, then the utility of permutation indexing merits careful consideration.

REFERENCES

(1) Joan Citron, Lewis Hart, and Herbert Ohlman. A Permutation Index to the "Preprints of the International Conference on Scientific Information," System Development Corporation paper SP-44, Rev. ed., December, 1959; originally published November, 1958.

(2) H. P. Luhn, Keyword-in-Context Index for Technical Literature (KWIC Index). IBM Advanced Systems Development Division report RC-127, August 1959. (The first published use of the Luhn KWIC system appears to be in the IBM bibliography, Literature on Information Retrieval and Machine Translation, September 1958.)
(3) J. W. Tukey, "Keeping Research in Contact with the Literature: Citation Indices and Beyond," *J. Chem. Doc.*, 2, 34–37 (1962).

A Correlative Indexing and Retrieval System for the Screening of Biological Data*

By ARTHUR W. ELIAS** and MARSHALL R. WARREN
Warner–Lambert Research Institute, Morris Plains, New Jersey
Received July 13, 1961

This system for retrieval of biological data was developed in response to the need of providing access to a large body of test results on about 10,000 chemical compounds. A further need was the *arrangement* of such results in an order that would enable correlative searching.

We began by analyzing the characteristics of our particular problem. The subject matter was diverse, and covered several biological disciplines. Within each discipline were many testing techniques. Finally, some of the test data were expressed as numbers; others, as judgments (for example, "this compound has *fair* activity"); and still others as descriptions or observations of response ("this compound produced withdrawal and depression in rhesus monkeys").

When the various needs were related to the characteristics of the data, these various requirements were formulated: (1) The system must be flexible, and able to accept data from many disciplines in several forms. (2) These data must be arranged or indexed so that correlations are possible within and between disciplines. (3) The system must be economical as to the time required for the coding of data, and with regard to the physical means employed.

The first step was to devise a general classification of the total body of information so that it might be indexed. The disciplines involved were studied, and seven general test areas were established (Table I). Four of these were pharmacological. The remaining areas reflected the discipline directly (numbers 5 and 7), or the intent of the study (number 6).

Table I. Classification of Subject Matter

No.	General Test Area
1	Toxicology
2	Central Nervous System
3	Muscle
4	Cardiovascular
5	Microbiological
6	Metabolism
7	Endocrine

Table II. Test Classification

400 Cardiovascular

400A Cardiovascular Screening–(B. P.-Biogenic Amine Response)
400B Cardiovascular Screening–(Induced Hypertension)
401A Contractile Force–*in Vivo*
401B Contractile Force–*in Vitro*

*Presented before the Division of Chemical Literature, American Chemical Society, St. Louis, Mo., March, 1961.

**Scientific Information Section, Wyeth Laboratories, Philadelphia, Pa.