INTRODUCTION

As the state of the art of computer development has advanced, the documentalist has turned more and more to mechanization for solutions to many of his problems. The process of retrieval of information, especially the searching of coded indexes prior to the selection of the document images themselves, has captured the lion’s share of applications effort. The retrieval phase, however, often poses relatively minor difficulties; the most crucial point in the documentation process—the one which may contribute more than any other to its over-all success or failure—is that of indexing the documents prior to their entry into the search files. The time-honored computer principle of “GIGO,” garbage in—garbage out, could apply nowhere better than in the indexing process. Because of the crucial nature of this function, it usually proves to be the most expensive operation of the information center, the presence of a goodly assortment of data processing equipment notwithstanding.

The Indexing Problem

Unfortunately, efforts to mechanize the indexing process have usually met with limited success. Many factors contribute to the inherent difficulty the computer undergoes in attempting to “understand” the documents being processed. Statistical indexing techniques ameliorate the situation only slightly. Such approaches are largely ineffectual primarily because the author of a given document is likely to be dealing in concepts, and is using various words in various combinations throughout the paper only in order to express those concepts effectively. Maron errs in saying that words and sentences are but one step removed from the things and events they describe. Rather, they are two steps removed: the factor, so elusive to the indexer, which exists between the real word and the written representation thereof, is the ideational concept which exists in the mind of the writer and which is conveyed to the reader through the medium of words on paper. It is thus the task of the indexer, whether man or machine, to assign terms to a given document.
on a concept basis rather than merely lifting words from the sentences without regard to content.

The indexing problem, then, is one of conveying to the retrieval system the various concepts which have been expressed by the author, in as accurate and complete a manner as possible. One method of accomplishing this goal is that of utilizing a staff of very highly trained indexers, each so competent in his own technical field that he is able to comprehend the precise ideational content of the document and thereby index it adequately...an obviously expensive proposition, albeit an effective one.

An alternative, one that presents far less expense, and yet insures that the intellectual content of the documents is thoroughly understood, is that of relegating the indexing process to the authors themselves. Certainly a highly structured relationship must exist between the documentation center and the author group for such a system to operate effectively; however, in controlled situations, such as that in which various research laboratories are responsible to a central office, author-indexing could be utilized to excellent advantage. More and more the call is heard for a sharpening of the author's responsibility in coping with the information problem.4

The present paper deals with such an effort, in which the authors contributing papers to a large scientific conference submitted both specific indexing terms and conceptual statements, which were supplied to the computer for analysis, cross-referencing, and final production in the form of a printed index. Analysis of the specific index terms presented little problem; however, processing the natural language concepts demanded novel techniques. In this respect the use by the computer of a specialized thesaurus showed itself to be quite successful and may prove to be an effective operational tool in the future in coping with the indexing problem.

The Federation Project

The Federation of American Societies for Experimental Biology is an organization made up of six separate professional groups, all vitally concerned with the biological sciences, which bonded together primarily because of a closely overlapping interest. The Federation holds its annual conference in the spring, at which time between two and three thousand papers are presented. Prior to the meeting, Federation Proceedings is distributed to all members; this publication contains both the abstracts of presented papers and a subject index. The Federation first approached UNIVAC in the fall of 1959 to investigate the feasibility of generating the index by means of a computer as well as of accomplishing several other tasks; the results of this pilot effort have been previously reported.5

For the 1963 meeting, however, an important experimental innovation was conceived. As before, the authors of the papers selected one or more terms which they felt applied to their reports from a list compiled by specialists in the fields represented by the Federation members. As before, they were given the opportunity of adding any concepts which they felt were pertinent, but which did not appear on the standard list. Blanks were furnished for this purpose at the bottom of the form provided (see Figure 1). Processing of the authors' standard choices was accomplished by using the corresponding numerical codes as input and then converting them to their alphanumeric representation by ordinary table look-up methods, also in a manner similar to previous years.

Many minor departures from prior processing methods were incorporated. The most important difference, however, between the processing of the 1963 index and indexes of former years was in the manner of handling the natural language concepts provided by the authors themselves. It was in this area that the computer thesaurus was put to work; whereas previously these statements had been analyzed, ultimately cross-referenced and entered into the index by trained personnel, now they were processed by machine with the use of the thesaurus.
From the collection of the Computer History Museum (www.computerhistory.org)

The Machine Thesaurus

As a purely experimental device, the thesaurus in its initial form was of necessity but an approximation. Its content was structured in terms of the collection of statements written in by authors in previous years. A projection of the subject matter was made, hoping that the field had not changed too radically, and recognizing that even the most accurate appraisal would fail to anticipate all areas of subject material.

Since the theoretical base of the experiment was the use of the thesaurus in concept processing, the computer was programmed to accept, wherever possible, the entire statements written in. By examination of previous years' forms, it was found that few statements were likely to exceed three valid words—that is, those remaining after the words comprising the statements were processed against a "throwaway" or delete list, and words not contributing to the conceptual meaning of the statements were eliminated. For this reason, the thesaurus was designed to contain a maximum of three valid words in the leftmost, or "look-up," portion.

Further, each of the words comprising the look-up portion was limited by practical considerations to twelve characters, the length of one computer word on the UNIVAC® I Computer. This restriction caused no difficulty, since it was found in previous years that twelve characters were sufficient in every case to insure unique recognition. Thus, each of the words or combinations of words anticipated in the authors' statements were entered in the left three fields of the thesaurus (see Figure 2).

Field 4 contained, where applicable, the valid equivalent of the concept as it was to appear in the index, with Field 5 being allocated for an additional term.

Field 6 was coded according to the action the computer was to take upon encountering a match between the author's statement and the thesaurus entry. Code 5 indicated that the word or words of the statement were to be replaced by the contents of Field 4 and, if applicable, Field 5. In many cases this meant merely contracting the separate words of the statement into a readable phrase acceptable as an index entry. The third entry in Figure 2 is an example.

More importantly, this code was used to cause uniform indexing of various forms of a given concept, including cases in which the
phrase used was on too specific a level to become an entry of the index. Thus, “adrenal cortex” and “adrenal medulla” would both be transformed into “adrenal glands.” Code 7 indicated that the concept appearing in the look-up portion was in itself a valid indexing term, needing no modification. In addition, a further option was incorporated into the system which represented an extension of Code 7, although it was not utilized in the present experiment. Provision was made to recognize Code 6 as indicating that the term in the look-up portion is valid, but the document represented by that term should also be indexed under the terms appearing in Fields 4 and 5. Use of this provision would allow more complete indexing in a collection of larger size or where greater flexibility is desired.

Since each phrase was to be processed as a conceptual entity where possible, the “A,” or “association,” code in Field 6 indicated to the computer that a given term found in the thesaurus was immediately followed by more specific terms using the first word or the first and second words of the phrase in question. Since each concept processed against the thesaurus contained a maximum of three words, a match could possibly occur on all three words. If this was not the case, a match was made at the highest level possible, and the remainder of the phrase was retained for further processing. For example, if the concept “adrenergic blocking agents” were encountered, a match would first be made on “adrenergic,” the association code would be found, a test and match would occur on the second word, and the association code found at the second level. A third level match would occur, and the appropriate action would be taken—the substitution with the two more precise concepts. In the case of “adrenal disorders,” however, although a match would be encountered at the first level, none would be found at the second level. “Adrenal” would become “adrenal glands,” and “disorders” would be retained for reprocessing, at which time it would be found in the thesaurus as an acceptable term. Thus, one of the most flexible features of the thesaurus was its ability to recognize and process the cases in which two concepts were in fact present in a single statement.

Certainly not every phrase to be written in by the authors could be anticipated. Provision was therefore made for “educating” the thesaurus by examining printer listings indicating cases in which a match was found at no level. The thesaurus was continually undergoing a process of amendment, new concepts being added as processing progressed, and the terms replaced as input to be accepted or modified. A flow diagram of the thesaurus process is shown in Figure 3.

![Figure 3. Flow Diagram of Thesaurus Processor.](image)

**Processing for Index Production**

The thesaurus phase, although the newest and most interesting of the processes, was only one of many necessary steps in the transformation of the punched input data into a complete, unified index. Figure 4 illustrates the various stages in the over-all process. One of the first tasks was to eliminate the nonsignificant words from the author statements. This was done by compiling a

---

*The author gratefully acknowledges the valuable participation of Mr. Paul Montgomery, of the UNIVAC Data Processing Center, New York, whose effort both in the areas of programming and supervision of processing was indispensable.*
special table of such words, as complete as was possible based on prior years’ information, against which each word of input was compared. Even so, a few were missed, but came to light as mismatches in the thesaurus phase and were eliminated.

Following the thesaurus phase, the index information, now completely expanded and conceptualized, was cross-referenced. Another special table had been devised for this purpose; each concept which might require cross-referencing (here, again, it was impossible to predict precisely which would and which would not be needed) was listed, together with a second concept to which the first would refer. Thus, two elements existed for each dummy entry: one, a reference “from”; the other, a reference “to.” A typical entry in the table was “vasodilators—(see) (see also) cardiovascular pharmacology.” The first task was to determine, in the case of each entry, whether any documents were indexed under the “to,” or right-hand term; if not, a cross-reference to that term in the index would naturally be useless. Otherwise, the dummy entry was accepted as applicable and was retained. The second task was to insert the dummy entries into the list of index information and, in doing so, to determine whether the particular cross-reference should read “see” or “see also.” Such a choice was dictated by the presence or absence of papers indexed under the “from,” or left-hand term. Thus, in the above example, if first it were determined that references were in fact listed under “cardiovascular pharmacology,” we must learn if entries exist under “vasodilators.” If this is the case, “see also” is to be inserted in the cross-reference, since the word “see” alone would imply that no documents are to be found dealing with vasodilators.

Finally, the completely expanded and cross-referenced index was edited and printed out for photo-offset reproduction (although plans had been tentatively made for production of paper tape coded to drive Photon printers).

A sample of the completed index is shown in Figure 5. Examples of cross-references may be noted. Each term under which one or more papers were indexed appears in alphabetic order, with the designations for all pertinent papers listed thereunder. A specific paper might be listed as “Radiations 2508.” The number refers to the number of the article—abstracts are arranged serially in the Proceedings. The term preceding the number informs the user of the main subject of the paper, so that he may judge the context within which his subject of interest is discussed.

SUMMARY

An experiment in lightening the indexing problem in document processing has been discussed. The use of the mechanized thesaurus appears to be promising, especially combined with preliminary indexing operations such as was accomplished by the authors themselves. It offers the distinct advantage of enabling the user of the index to retrieve in terms of a complete conceptual representation of the document collection, without the burden of

Figure 4. Flow Diagram of System.
# Subject Index

| PHOSPHORESCENCE | PROTEIN NUTRITION 26781 VITAMINS FAT SOLU BLE 4661 NEUTRALIS 216 |
| PHOSPHORUS | PLASMA PROTEINS 23981 BLOOD COAGULATION 205 6.2399-2402 ANTIBODY/ANTIBODY REACTIONS 234 91 RESTRAINT 564 |
| PHOSPHORUS | PLASMA PROTEINS 23981 BLOOD COAGULATION 205-2402 ANTIBODY/ANTIBODY REACTIONS 234 91 RESTRAINT 564 |
| PHOSPHORUS | PLASMA PROTEINS 23981 BLOOD COAGULATION 205-2402 ANTIBody/ANTIBODY REACTIONS 234 91 RESTRAINT 564 |
| PHOSPHORUS | PLASMA PROTEINS 23981 BLOOD COAGULATION 205-2402 ANTIBODY/ANTIBODY REACTIONS 234 91 RESTRAINT 564 |
| PHOSPHORUS | PLASMA PROTEINS 23981 BLOOD COAGULATION 205-2402 ANTIBody/ANTIBODY REACTIONS 234 91 RESTRAINT 564 |
| PHOSPHORUS | PLASMA PROTEINS 23981 BLOOD COAGULATION 205-2402 ANTIBody/ANTIBODY REACTIONS 234 91 RESTRAINT 564 |
| PHOSPHORUS | PLASMA PROTEINS 23981 BLOOD COAGULATION 205-2402 ANTIBody/ANTIBODY REACTIONS 234 91 RESTRAINT 564 |
| PHOSPHORUS | PLASMA PROTEINS 23981 BLOOD COAGULATION 205-2402 ANTIBody/ANTIBODY REACTIONS 234 91 RESTRAINT 564 |
| PHOSPHORUS | PLASMA PROTEINS 23981 BLOOD COAGULATION 205-2402 ANTIBody/ANTIBODY REACTIONS 234 91 RESTRAINT 564 |
| PHOSPHORUS | PLASMA PROTEINS 23981 BLOOD COAGULATION 205-2402 ANTIBody/ANTIBODY REACTIONS 234 91 RESTRAINT 564 |
| PHOSPHORUS | PLASMA PROTEINS 23981 BLOOD COAGULATION 205-2402 ANTIBody/ANTIBODY REACTIONS 234 91 RESTRAINT 564 |
| PHOSPHORUS | PLASMA PROTEINS 23981 BLOOD COAGULATION 205-2402 ANTIBody/ANTIBODY REACTIONS 234 91 RESTRAINT 564 |
| PHOSPHORUS | PLASMA PROTEINS 23981 BLOOD COAGULATION 205-2402 ANTIBody/ANTIBODY REACTIONS 234 91 RESTRAINT 564 |
| PHOSPHORUS | PLASMA PROTEINS 23981 BLOOD COAGULATION 205-2402 ANTIBody/ANTIBODY REACTIONS 234 91 RESTRAINT 564 |
| PHOSPHORUS | PLASMA PROTEINS 23981 BLOOD COAGULATION 205-2402 ANTIBody/ANTIBODY REACTIONS 234 91 RESTRAINT 564 |
| PHOSPHORUS | PLASMA PROTEINS 23981 BLOOD COAGULATION 205-2402 ANTIBody/ANTIBODY REACTIONS 234 91 RESTRAINT 564 |
| PHOSPHORUS | PLASMA PROTEINS 23981 BLOOD COAGULATION 205-2402 ANTIBody/ANTIBODY REACTIONS 234 91 RESTRAINT 564 |
| PHOSPHORUS | PLASMA PROTEINS 23981 BLOOD COAGULATION 205-2402 ANTIBody/ANTIBODY REACTIONS 234 91 RESTRAINT 564 |

---

Figure 5. Sample Page from Final Index.

From the collection of the Computer History Museum (www.computerhistory.org)
performing deep indexing by highly trained personnel.

The need for subject specialists, however, has not been entirely obviated. Luhn\textsuperscript{6} correctly states that the preparation of any special dictionary or thesaurus demands complete familiarity with the field in question. This was certainly found to be true; construction of the thesaurus used in the present experiment depended not only on study of previous years' materials, but on their being completely understood. Once again, the essence of successfully representing the subject matter was the concept, which, when incorporated into the thesaurus, enabled it to fulfill its purpose.

The study, although oriented toward the production of a printed index, has demonstrated a method by which several types of retrieval files may be generated. Certainly, preparation of an index to be searched not by humans but by the computer could be accomplished by such a process. The ability to provide standardized indexing terms automatically, together with the advantage of complete cross-referencing, would result in a high degree of retrieval efficiency.

BIBLIOGRAPHY


