# Final Report on Contract Nonr N00014-67-A-0370-0002, and on earlier contracts with the Information Systems Branch, Office of Naval Research, in which John O'Connor was principal investigator

John O'Connor
Center for Information Science
Lehigh University

The work reported on here was conducted from April 1, 1959 through March 31, 1970, at the following institutions, under the specified contract numbers:

April 1, 1959- March 31, 1963 Institute for Cooperative Research, University of Pennsylvania, Phila., Pa., Contract No. Nonr 551 (35).

April 1, 1963- March 31, 1966 Institute for Scientific Information, Philadelphia, Pa., Contract No. Nonr 4183(00).

April 1, 1966- March 31, 1967 Institute for Advancement of Medical Communication, Philadelphia, Pa., Contract No. Nonr N00014-66-C0096.

April 1, 1967- March 31, 1970 Center for Information Science, Lehigh University, Bethlehem, Pa. Contract No. Nonr N00014-67-A-0370-0002

The work was also jointly supported during April 1, 1961 - June 30, 1965 by the Information Research Division, Air Force Office of Scientific Research.

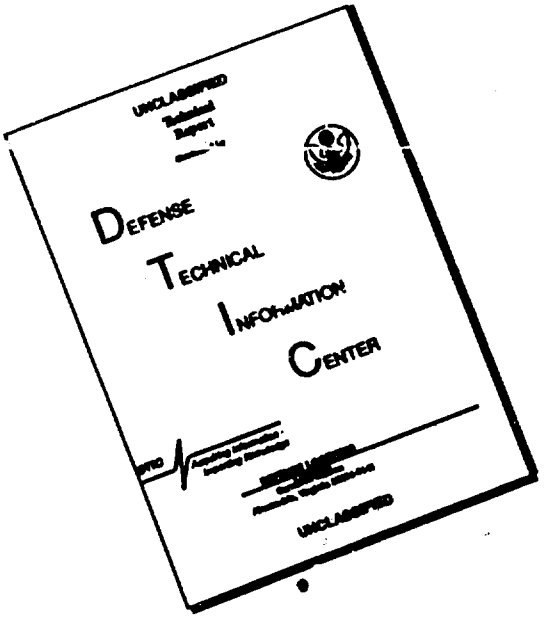
The work has primarily concerned subject document retrieval systems (systems that retrieve documents in response to subject requests). It can be roughly classified into six categories: storage organization, automatic indexing, retrieval evaluation, retrieval relevance of documents, retrieval of "answer-providing" documents, and a miscellaneous few preliminary studies of retrieval. The brief summaries and annotated bibliography below are classified accordingly.

Storage Organization. Suppose each document in a collection has been subject indexed with an unordered set of index terms. How can these index term sets and the corresponding document identifications (e.g. serial numbers or references) be arranged in storage compactly and/or in ways which will reduce search time for men and/or machines? Several new methods are suggested.



10

# DISCLAIMER NOTICE




THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

(1) "The Possibilities of Document Grouping for Reducing Retrieval Storage Size and Search Time. In Advances in Documentation and Library Science, Vol. III, Pt. 1 (Ed. by A. Kent), Interscience, N.Y., 1961, pp. 237-79. The union of index sets assigned to documents in a group is treated as a unit for storage and searching, at a cost in false retrieval. A computer heuristic for forming groups to keep false retrieval small is outlined. Some situations in which document grouping might be useful without a complex group-forming procedure are also described.

(2) A Possibly Inexpensive Attachment for a Microfilm Reader to Permit Synchronized Coordinate Search. J. of Chemical Documentation, Jan., 1963 (Vol. 3, No. 1), pp. 29-32. The attachment is a motorized scroll on which document index sets are recorded in scan columns. (Scan column storage, for rapid human searching, was developed by the author in earlier work at Remington-Rand Univac under an ONR contract in which he was not principal investigator. See The Scan Column Index, Amer. Documentations, April, 1962 (Vol. 13, No. 2), pp. 204-9). Various possible problems and their solutions, and design variations (including use of document grouping) are described.

(3) A Note on the Possibility of a Divided Structure File Permitting Arbitrary Substructure Searches. Issued as a report, June, 1960; CFSTI No. AD 243 354. This is not concerned with searching a file of unordered sets of index terms. It is a brief preliminary discussion of possibly finding a reasonable file organization of the following kind: the file is based on a set of structures $B_i$ each of which is a substructure of at least one but no more than M file structures, and such that any other substructure which is included in at most M file structures also includes at least one $B_i$. Such $B_i$'s could be used as headings for lists of all reasonable search-specified substructures.

Automatic Indexing. How successfully can computers assign subject index terms to documents? There are some general discussions of this question and ways of getting answers to it, and a variety of empirical results.

(4) Some Suggested Mechanized Indexing Investigations which Require No Machines. Amer. Documentation, July, 1961 (Vol. 12, No. 3), pp. 198-203. (Issued as a report in 1960; CFSTI No. AD 240 040.) Procedures are described for systematic small-sample tests of various forms of automatic indexing techniques involving word frequencies and/or a thesaurus. Possible kinds of indexing failure are systematically described.

(5) Some Remarks on Mechanized Indexing and some Small-Scale Empirical Results. In Machine Indexing, American Univ., Wash., D.C., 1961, pp. 266-79. (This is primarily an abridged form of the report, Mechanized Indexing: Some General Remarks and some Small-Scale Empirical Results, issued in December, 1960; CFSTI No. AD 250 209.) The primary general remarks are superseded by pp. 443-8 of (7), and the report of empirical results by (10).

(6) Correlation of Indexing Headings and Title Words in Three Medical Indexing Systems. Amer. Documentation, April, 1964 (Vol. 15, No. 2), pp. 96-104. A small-sample investigation of how frequently index terms assigned to documents by subject specialists in each of the three systems could be assigned automatically by relating words and phrases from a standard indexing vocabulary to words and phrases in document titles by means of a thesaurus. The percentages of humanly assigned index terms automatically assignable in this way for the three systems were 19-45%, 40-68% and 13-39% (ranges estimated with 0.95 confidence from samples). These results were in cautionary contrast to the 86% obtained earlier by others in a study of Index Medicus (see C. Montgomery and D. Swanson. Machinelike Indexing by People, Amer. Documentation, October, 1962 (Vol. 13, No. 4), pp. 359-66.

(7) Mechanized Indexing Methods and Their Testing. J. of the Assoc. for Computing Machinery, October, 1964 (Vol. 11, No. 4), pp. 437-49. (Issued as a report, in slightly different form, in 1963; CFSTI number not known.) Pp. 437-43 contain a systematic review of proposed automatic indexing procedures, which is sufficiently general in formulation to be still essentially not obsolete. Pp. 443-4 discuss various points, including the question of how much less human effort is required to judge which incoming documents are index-worthy than to actually index them (the job handled by automatic indexing). The remainder of the paper summarizes problems of retrieval evaluation (for instance in testing how good is the retrieval permitted by automatic indexing), and also discusses briefly testing automatic indexing techniques by determining how well they can duplicate the results of presumably competent human indexing.

(8) Mechanized Indexing Studies of MSD Toxicity, Part I. Issued as a report in December, 1963; CFSTI No. AD 436 523. This preliminary report is superseded by (10).

(9) Mechanized Indexing Studies of MSD Toxicity, Part II. Issued as a report in March, 1964; CFSTI No. AD 437 868. This preliminary report is superseded by (10).

(10) Automatic Subject Recognition in Scientific Papers: an Empirical Study. J. of the Assoc. for Computing Machinery, October, 1965 (Vol. 12, No. 4), pp. 490-515. Two subject index terms from an operating retrieval system were studied intensively to determine how well a computer could assign them. The humanly produced indexing for the system was used as a standard, with some checking for indexer errors. Thesaurus rules failed to identify one fourth of the toxicity papers, because of nonrecurring expressions, such as "complication due to isoniazid". A new kind of automatic indexing rule was developed for such expressions, which successfully identified almost all the non-thesauric toxicity papers. A variety of other results are given for various automatic indexing techniques previously proposed and several other new techniques described in the paper. In rough summary, the combined techniques tested on new samples assigned the terms to about 90% of the documents that should have them (comparable to the operational indexing), but also produced one false term-document assignment per correct term document assignment. Many examples of various kinds of complications which scientific papers present for automatic indexing are given in the paper.

Retrieval Evaluation. Given an operating subject document retrieval system, or a proposed subject document retrieval technique to be tested experimentally, what are reasonable evaluation procedures? Various problems involved in trying to answer this question are discussed and (hopefully) clarified, and a few positive suggestions are made.

(7) (repeat entry). Pp. 445-7 systematically summarize problems involved in retrieval evaluation, such as, how much reflection and discussion should precede a user's final judgment about the value of a retrieved document, and how determine which unretrieved documents should have been retrieved. It is suggested that the latter problem might be dealt with satisfactorily, using real retrieval requests, by having a group of subject specialists cooperate to cover a collection better than does any usual retrieval system. Such coverage might be a relatively slight extension of their usual work (if the latter involves an interest in the collection).

(11) Review of Cranfield Indexing Tests. J. of Documentation, December, 1961 (Vol. 17, No. 4), pp. 252-61. Besides various specific critical comments on the testing reviewed (which tested different indexings by evaluating the various retrievals they permitted), some general problems involved in experimental testing of retrieval "devices" (such as kinds of indexing) are described, notably that it is hard to generalize from such experiments because of the large

number of variables which might affect the results, and because many of the variables do not have a clear set of significantly different values (e.g. what are the important different forms of requester-searcher consultation?). More positively, it is suggested that close study of particular successes and failures of various alternative retrieval "devices" in such an experimental situation as that of the testing reviewed, especially of successes for one device which are failures for another, might produce further understanding of how and why the various devices work and fail to work, and provide some bases for hypotheses about the kinds of relations which do and can exist among requester's knowledge, searcher's knowledge, indexer's knowledge, requester's language, indexing language, document language, document subjects, etc.

(12) Review of "Searching Legal Literature Electronically: Results of a Test Program" (by J. Melton and R. Bensing, 45 Minnesota Law Review, pp. 229-48, 1960) in M.U.L.L. (Modern Uses of Logic in the Law), March, 1962, pp. 18-21. Various specific critical remarks are made about the retrieval testing reviewed.

Retrieval Relevance of Documents. It is commonly said that a subject document retrieval system should retrieve "relevant" documents. But what does it mean for a document to be "relevant"? Many say that a relevant document is one which "matches" a retrieval request, and many others say that it is rather one which "satisfies the information need" of the requester. However each of these answers involves difficulties which are the concern of the following two papers respectively.

(13) Relevance Disagreements and Unclear Request Forms. Amer. Documentation, July, 1967 (Vol. 18, No. 3, pp. 165-77. Disagreements about the relevance [matching] of documents to retrieval requests occur because relevance judges differently interpret requests or documents. Requests may be differently interpreted because they are unclear. Well-known types of request obscurity are reviewed. Less well known is that a request may be unclear because its form -- "documents about subject S", "documents answering question Q", etc. -- is unclear. Explications are developed of the meanings of the request forms just given and several others. A request of any of the forms discussed is interpreted to be for documents which support statements of a particular kind. (Examples are given which suggest that some, perhaps all, "about S" requests are unclear.) Various ways in which documents may support statements are distinguished; these depend on such factors as parts of a document used, inference strength and background knowledge permitted.

(14) Some Question Concerning "Information Need". Amer. Documentation, April, 1968 (Vol. 19, No. 2), pp. 200-3. The expression "satisfying a requester's information need" is often used but its meaning is obscure. The literature on "information need" in relation to retrieval suggests three different (though not inconsistent) possible interpretations. However each of these interpretations is itself fundamentally unclear. For example, if "satisfy requester's information need" means provide information that will "help his work", it must be asked, "Help his work according to whom?", since scientists often disagree about what problems to work on and how to work on them; this latter point is illustrated by examples from the history of science and current scientific practice. The various obscurities involved in the different interpretations are indicated by critical questions, which those who write of information need are invited to answer. (So far, only F. W. Lancaster has replied, in a letter to the editor (Amer. Documentation, April, 1968, p. 206) which is helpful but still seriously unclear -- see the author's reply (Amer. Documentation, Oct. 1968, pp. 416-7), which has not yet received a further answer.)

Retrieval of "Answer-Providing" Documents. Better understanding of subject document retrieval might result if different functions of subject document retrieval systems are studied separately. The next three papers (and, hopefully, much future research of the author) are concerned with retrieval of documents, in response to a question, from each of which an answer to that question can be inferred ("answer-providing" documents).

(15) Retrieval of Answer-Providing Documents. Amer. Documentation, October, 1968 (Vol. 19, No. 4), pp. 381-6. "Answer can be inferred from document" has many possible meanings (see the last sentence of the annotation for (13)), one of which must be selected (an "inference specification"). Inasmuch as scientists in a field sometimes disagree about the correctness of inferences, have somewhat different background knowledge, etc., any inference specification can only approximate scientific inference practices. Two sources of systematic knowledge of document-statement inference practices in a scientific field are described. The second part of the paper describes a general approach to indexing documents for answer-providing document retrieval.

(16) Some Independent Agreements and Resolved Disagreements About Answer-Providing Documents. Amer. Documentation, October, 1969 (Vol. 20, No. 4), pp. 311-9. Two subject specialist judges independently compared a set of questions and a set of documents to find answer-providing documents. They then discussed their disagreements, attempting to resolve them.

In each case the positive judge (who had independently judged a document answer-providing) was first asked to indicate what answer he inferred, and from what passage(s). The further discussion depended on the details of each case. There were 32 independent agreements on positive judgments. There were 48 disagreements between independent judgments, all resolved by discussion. Thirty-four resolutions were agreements on positive judgments, accomplished by pointing out overlooked passages, unnoticed connections, or alternative meanings. Fourteen resolutions were agreements on negative judgments, accomplished by pointing out document misinterpretations, the challenged positive judge being unable to describe an inference and joint work not finding one, or agreement that both judges lacked sufficient background knowledge. In general, the resolution procedures used will resolve a disagreement about whether a document is answer-providing or reduce it to a familiar kind of scientific disagreement (about a passage's meaning, a statement's correctness, or an inference's correctness). This seems better than the common procedure of treating relevance judgments as subjective and not open to rational discussion.

(17) Answer-Providing Documents: Some Inference Descriptions and Text-Searching Retrieval Results. Submitted to J. of the Amer. Soc. for Information Science (formerly Amer. Documentation). Document-answer inference descriptions are given for twenty question-(answer-providing) document pairs randomly selected from the sixty six found in the study reported in (16). The first premise of each inference consisted of one or more quoted document passages, sometimes accompanied by bracketed additions (justified by other passages in the document) to make meanings explicit. Each inference included one or more background knowledge premises. Three inferences also required "author intent" premises, because (e.g.) a document gave a reason for negotiating requests but did not say it was giving a reason, and the question asked, "What reasons have been given. . .?".

The second part of the paper reports a small-scale study, using the twenty question-document pairs from the first part and the set of documents used in the study reported in (16), of how effectively answer-providing documents can be retrieved by text searching. Only rather simple procedures were used. Answer-providing document passages, their near contexts, titles, and section headings were examined for stem and thesaurus matches to question words. Proximity of matching document words (counting titles and section headings as adjacent to document passages) was used as a measure of relation. Matches to question words probably occurring infrequently in the document collection were weighted more heavily. A procedure was

formulated for ranking question-document pairs for strength of match on the basis of these factors. As a test, the procedure was applied to a random thirty question-document passage pairs in which the documents were not answer-providing. One of these false pairs ranked at least as high as nine of the twenty correct question-document pairs, and seven other false pairs each tied or outranked one or two of the correct pairs. Thus the text searching procedures used are probably inadequate for high-recall retrieval, but might be satisfactory for high-precision retrieval.

Miscellaneous Preliminary Studies of Retrieval

(18) On Retrieval in Aid of Scientific Discovery. Issued as a report in August, 1960; CFSTI No. AD 245 919. This report has been superseded by parts of (7), (13) and (14).

(19) Potential of Alternative Ways for Recording and Reporting Biological Research Reports. In Symposium on Biological Communications, Biological Abstracts, Phila., Pa., 1960, pp. 6-7 (Also published in Biological Abstracts, April 1, 1961.) Some alternative possible ways are sketched of writing or annotating biological research papers to permit them to be more quickly usefully read. This is related to retrieval only as a possible aid to winnowing retrieval system outputs.

(20) What Should a Retrieval System for Scientific Information Do? IRE Trans. on Engineering Writing and Speech, Dec., 1962 (Vol. EWS-5, No. 2), pp. 75-7. Some difficulties in getting answers to the question posed in the title are sketched -- for example, the lack of knowledge about the relations between scientific literature and scientific discovery, and the differences in form and language used in various specialized fields. Some areas of needed research on retrieval systems for scientific information are suggested.

(21) On the Roles of Specialist Labor in Retrieval, Part A. Issued as a report in May, 1963; CFSTI No. AD 409 283. A brief abstract description of a user's retrieval situation, attempting the beginning of understanding of the actual and possible roles of subject specialist labor (as opposed to human clerical or machine labor) in retrieval.

Unclassified

Security Classification

**DOCUMENT CONTROL DATA - R & D**

*(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)*

1. ORIGINATING ACTIVITY (Corporate author)
Center for Information Science
Lehigh University
Bethlehem, Pa. 18015

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP
---

3. REPORT TITLE
FINAL REPORT ON CONTRACT NONR N00014-67-A-0370-0002, AND ON EARLIER CONTRACTS WITH THE INFORMATION SYSTEMS BRANCH, OFFICE OF NAVAL RESEARCH, IN WHICH JOHN O'CONNOR WAS PRINCIPAL INVESTIGATOR

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Final report, April 1, 1959 - March 31, 1970

5. AUTHOR(S) (First name, middle initial, last name)
John J. O'Connor

6. REPORT DATE
July, 1970

7a. TOTAL NO. OF PAGES
8

7b. NO. OF REFS
29

8a. CONTRACT OR GRANT NO.
NONR N00014-67-A-0370-0002

b. PROJECT NO.

c. ---

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
---

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
---

10. DISTRIBUTION STATEMENT
Distribution of this document is unlimited

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Information Systems Program
Office of Naval Research
Washington, D.C. 2-360

13. ABSTRACT

Work done by the author as principal investigator under Office of Naval Research Contracts from April 1, 1959 to March 31, 1970 is summarized. The work, all of which concerns subject document retrieval systems, falls into the following six categories: storage organization, automatic indexing, retrieval evaluation, retrieval relevance of documents, retrieval of "answer-providing" documents, and miscellaneous preliminary retrieval studies. For each of the six categories a brief introductory summary is followed by a list of papers and/or reports, each accompanied by an annotative summary. Sixteen published papers and five reports are listed. Eight other publications (in three cases report forms of later published papers) are cited in the summaries.

DD FORM 1 NOV 65 1473

Unclassified (Over)

Security Classification

| 14. KEY WORDS | LINK A | | LINK B | | LINK C | |
|---|---|---|---|---|---|---|
| | ROLE | WT | ROLE | WT | ROLE | WT |
| abstracts | | | | | | |
| document storage | | | | | | |
| automatic indexing | | | | | | |
| subject indexing | | | | | | |
| information retrieval effectiveness | | | | | | |
| information retrieval | | | | | | |