NBS TECHNICAL NOTE

297

Evaluation of Information Systems: A Selected Bibliography With Informative Abstracts

U.S. DEPARTMENT OF COMMERCE
National Bureau of Standards

THE NATIONAL BUREAU OF STANDARDS

The National Bureau of Standards[1] provides measurement and technical information services essential to the efficiency and effectiveness of the work of the Nation's scientists and engineers. The Bureau serves also as a focal point in the Federal Government for assuring maximum application of the physical and engineering sciences to the advancement of technology in industry and commerce. To accomplish this mission, the Bureau is organized into three institutes covering broad program areas of research and services:

THE INSTITUTE FOR BASIC STANDARDS . . . provides the central basis within the United States for a complete and consistent system of physical measurements, coordinates that system with the measurement systems of other nations, and furnishes essential services leading to accurate and uniform physical measurements throughout the Nation's scientific community, industry, and commerce. This Institute comprises a series of divisions, each serving a classical subject matter area:

—Applied Mathematics—Electricity—Metrology—Mechanics—Heat—Atomic Physics—Physical Chemistry—Radiation Physics—Laboratory Astrophysics[2]—Radio Standards Laboratory,[2] which includes Radio Standards Physics and Radio Standards Engineering—Office of Standard Reference Data.

THE INSTITUTE FOR MATERIALS RESEARCH . . . conducts materials research and provides associated materials services including mainly reference materials and data on the properties of materials. Beyond its direct interest to the Nation's scientists and engineers, this Institute yields services which are essential to the advancement of technology in industry and commerce. This Institute is organized primarily by technical fields:

—Analytical Chemistry—Metallurgy—Reactor Radiations—Polymers—Inorganic Materials—Cryogenics[2]—Office of Standard Reference Materials.

THE INSTITUTE FOR APPLIED TECHNOLOGY . . . provides technical services to promote the use of available technology and to facilitate technological innovation in industry and government. The principal elements of this Institute are:

—Building Research—Electronic Instrumentation—Technical Analysis—Center for Computer Sciences and Technology—Textile and Apparel Technology Center—Office of Weights and Measures—Office of Engineering Standards Services—Office of Invention and Innovation—Office of Vehicle Systems Research—Clearinghouse for Federal Scientific and Technical Information[3]—Materials Evaluation Laboratory—NBS/GSA Testing Laboratory.

---

[1] Headquarters and Laboratories at Gaithersburg, Maryland, unless otherwise noted; mailing address Washington, D. C., 20234.

[2] Located at Boulder, Colorado, 80302.

[3] Located at 5285 Port Royal Road, Springfield, Virginia 22151.

UNITED STATES DEPARTMENT OF COMMERCE
Alexander B. Trowbridge, Secretary
NATIONAL BUREAU OF STANDARDS • A. V. Astin, Director

TECHNICAL NOTE 297

ISSUED DECEMBER 1967

Evaluation of Information Systems: A Selected Bibliography With Informative Abstracts

Madeline M. Henderson

Technical Information Exchange
Center for Computer Sciences and Technology
National Bureau of Standards
Washington, D.C. 20234

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C., 20402 - Price $1

Contents

EVALUATION OF INFORMATION SYSTEMS: A SELECTED BIBLIOGRAPHY WITH INFORMATIVE ABSTRACTS

Madeline M. Henderson

A survey of the literature on evaluation of information systems has been conducted by the Technical Information Exchange, Center for Computer Sciences and Technology, National Bureau of Standards. During the early stages of the survey, the literature was divided among descriptions of programs which compared the performance of two or more information systems, accounts of programs which studied the performance of one system, papers and reports which discussed the problems of evaluation programs, and documents which proposed new techniques for evaluation of systems. From the total literature collected, those references which were judged to be most directly concerned with the subject of evaluation of information systems were selected and abstracted. The abstracts are designed to give a summary of the content of the corresponding paper; the author's own wording was used extensively, in order to avoid misinterpretation. All of the references collected are listed, in alphabetic order of authors' names, in the appendix to the main body of this publication.

Key Words: evaluation, performance, testing of information systems, effectiveness, relevance.

INTRODUCTION

The subject matter covered by this bibliography is the evaluation of information retrieval systems or document reference systems. The scope does not extend to data processing systems, nor does it include the equipment used in such systems --- the hardware, per se. Equipment is considered only as a component of a total system which is being evaluated. Again, the emphasis is on evaluation or testing, rather than design or selection of systems. There is definite overlap in the criteria applied in design and those used in evaluation, but the orientation of the work reported has been the guiding factor in selecting material for inclusion in this bibliography.

The evaluations may be of total systems, or of components or subsystems of total systems. Such systems may be in operation, or may be experimental and set up specifically for a testing program.

The purpose of the abstracts is to give a summary of the corresponding paper, report, or book section. The abstracts were prepared so as to be informative; the author's own wording has been used extensively, in order to avoid misinterpretation.

The citations and abstracts in this bibliography are grouped in four categories: comparative evaluation, descriptive evaluation, discussion of evaluation factors, and proposals. These categories cannot be mutually exclusive, as discussion is included in comparative testing papers and descriptive evaluations precede proposals for further testing methods. But in general the papers are assigned to categories according to their primary purpose or scope.

Preceding each section or category grouping there appears a general summary of the content of that section, its definition, scope and relation to other sections. These summaries are designed to present an overview of the section and, collectively, of the field of information system evaluation. Reference to specific work is facilitated by subject, author and organization indexes.

This bibliography constitutes the completion of one phase in an extensive survey on the subject of evaluation of information systems. The initial effort of the survey resulted in the preparation of an alphabetized bibliography of all references processed (described below) from which this annotated bibliography was then selected and prepared; the finale will be a state-of-the-art report, now in preparation.

This annotated bibliography covers material published or available through June 1965. Some works of more recent date have not been included because they are not yet generally available for distribution. These include papers presented at such conferences as the NATO Advanced Study Institute on Evaluation of Information Systems in July 1965, the FID-ADI Conference in October 1965, and the ADI Annual Meeting in October 1966. Recent progress reports from continuing projects, such as the ASLIB-Cranfield study, and some papers published in journals or as technical reports and memoranda also did not appear in time for processing. Additional material, dealing with assessment or critique of the experimental and theoretical work described among these annotations, was not included either because these assessments were unavailable or because they were judged to be not directly applicable to the annotated bibliography. However, in all these cases, the material will be included in the state-of-the-art summary, and the references are included in the alphabetized bibliography described below.

Prior to this publication a bibliography of references considered for annotating was prepared. The listing is appended to this publication. It contains not only all of the items for which abstracts were prepared but all of the items processed but not abstracted. The latter included papers on system design, as discussed above; those expressing opinions or assessments, rather than reporting tests or evaluations; and papers of borderline interest, potentially related perhaps but not directly concerned with evaluation. The bibliography, including the partial listing of more recent material mentioned above, constitutes an extensive, if not exhaustive, compilation of material on the subject of evaluation of information systems.

The contribution of Mrs. Betty Anderson in the editing and typing of the manuscript has been invaluable. The staff of the Technical Information Exchange, a part of the Center for Computer Sciences and Technology, NBS, provided excellent bibliographic support.

SECTION ONE

COMPARATIVE EVALUATION

Testing and evaluation of information systems and system components can take the form of comparing performance characteristics of several systems or components, or of describing the performance of one system or component. In this section appear the former, those items dealing with comparison of the performance of several systems.

Comparison, again, may be of one system against others or of one system against a standard or ideal. The systems considered include those designed and set up especially for the testing program, or systems actually in operation. In the first case, the primary purpose of the test program is usually the study and development of test techniques and principles. In the other case, it is the performance of the operating system that is under scrutiny.

One of the foremost projects designed to study testing techniques is the ASLIB - Cranfield Study Project. Established in 1957 with a National Science Foundation grant, the project undertook the evaluation of the comparative efficiency of four indexing systems. Careful control was exercised in setting up the test systems and later in comparing the retrieval performance of the systems, resulting in proposed detailed test procedures. The early work on the project was reported thoroughly, and its reports in turn were reviewed widely. Critical appraisals of the techniques and results of the Project were published. All of this material --- original work and subsequent critiques --- are grouped at the beginning of this section, and include items I-1 through I-10, in alphabetic order of author's names.

The current work of the ASLIB team is directed at study of the indexing process itself, and as such appears in the second section, with descriptive evaluation work.

Following the group of abstracts on the ASLIB project, abstracts of other work on comparative evaluation appear, again in alphabetic order of author's names (I-11 through I-33). These reports cover comparison not only of total systems but of components or subsystems. The majority of the reports deal with testing of actual operating systems (15), as contrasted to study of systems specifically designed for test programs (8). Fourteen of the total number of programs compared different indexing methods --- classification, subject headings, coordinate systems, and title indexes, for example. Again, eight of the programs tested performance, or retrieval effectiveness, of the total systems, while the rest tested various other properties of systems or components.

I-1. A Report on a Test of the Index of Metallurgical Literature of Western Reserve University, by Jean Aitchison and Cyril W. Cleverdon, Report to the National Science Foundation on the ASLIB-Cranfield Research Project, Cranfield, England, The College of Aeronautics, October 1963, 270 p.

The method developed at Cranfield for testing the capabilities of indexing systems in retrieving documents relevant to search questions (cf. I-2) was tried out on the Western Reserve University (WRU) index of metallurgical literature. The objective of this test was to evaluate the operating efficiency (but not the economic efficiency) of the WRU index, including evaluation of the index language and of the processes of indexing and search programming by the WRU method. In order to give comparative figures for the performance of a more conventional index at Cranfield on the same test documents, the facet classification for engineering developed at the English Electric Company, Ltd., was used. The test results bear out the preliminary findings that on the average both systems operated at about the same recall level, but that the Cranfield system consistently retrieved a much larger proportion of relevant to irrelevant documents.

From 114 searches of the 1,300-document test store recall and relevance ratios for Relevance 2 documents and for Relevance 2 and 3 documents were computed. (Relevance 2 documents include the source documents on which the questions were based and other documents as relevant to the questions as the source documents; Relevance 3 documents include all those other than Relevance 2 items which have some relevance to the questions.) In the first instance, the figures were 81.4 percent recall and 7.5 percent relevance for WRU and 85.3 percent recall and 16.2 percent relevance for Cranfield. In the second case, the figures were 75.8 percent recall and 17.7 percent relevance for WRU and 69.5 percent recall and 33.7 percent relevance for Cranfield. Analysis of the results indicated that the indexing done by the WRU group was not at fault, but that the main factor in WRU failures to recall relevant documents was the failure to match the concepts used in the search programs with those in the questions. As for the relatively low relevance ratio, analysis suggested that the high level of exhaustive indexing was partly to blame.

I-2. Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems, by Cyril W. Cleverdon, Report to the National Science Foundation on the ASLIB-Cranfield Research Project, Cranfield, England, The College of Aeronautics, October 1962, 305 p.

(This major report was preceded by "Report on the First Stage of an Investigation into the Comparative Efficiency of Indexing Systems", by Cyril W. Cleverdon, September 1960, 166 p., and "Interim Report on the Test Programme of an Investigation into the Comparative Efficiency of Indexing Systems", by Cyril W. Cleverdon, November 1960, 80 p., the essential contents of which are covered by the contents of this report. It was reviewed by L.J.B. Mote in Journal of Documentation 19 (June 1963), and by N.D. Stevens in Library Resources and Technical Services 8 (Winter 1964). In addition, an extensive review of this report was prepared by Richmond (cf. I-7), and reviews of the earlier reports by O'Connor (cf. I-5). An explanation of the total project was given in "The Evaluation of Systems Used in Information Retrieval", by Cyril W. Cleverdon, in Proceedings of the International Conference on Scientific Information, Vol. 1, Washington, D.C., National Academy of Sciences-National Research Council, 1959, p. 687-698. The details of the indexing program and a discussion of the probable test program were included in "An Investigation into the Comparative Efficiency of Information Retrieval Systems", by Cyril W. Cleverdon, in UNESCO Bulletin for Libraries 12, 267-270 (November-December 1958). Another discussion of the total study and results from the first round of testing were published in "The ASLIB-Cranfield Research Project on the Comparative Efficiency of Indexing Systems", by Cyril W. Cleverdon, in ASLIB Proceedings 12, 421-431 (December 1960)).

In 1957 the Association of Special Libraries and Information Bureaux (ASLIB) undertook, under a National Science Foundation grant, an investigation of the comparative efficiency of four indexing systems: (1) an alphabetical subject index, (2) the Universal Decimal Classification, (3) a special facet classification, and (4) the Uniterm system of coordinate indexing. The work was carried out at the College of Aeronautics, Cranfield, England, under the direction of Mr. Cleverdon.

The project was conducted in two stages: In the first, a total of 18,000 documents (equally divided between the fields of high-speed aerodynamics and general aeronautics, between papers from the United States and from other English-speaking countries, and between research reports and journal articles) were indexed by each of the four systems. The indexing staff was selected for their differing experience (technical knowledge but no indexing experience, indexing experience in the subject field, and theoretical knowledge of indexing only), and the time for indexing was controlled (varying from 16 minutes, through 12, 8, and 4 minutes, to 2 minutes). The 18,000 documents were divided into three subprograms of 6,000 documents, within each of which there were 60 groups of 100 documents. Each group was indexed by a different indexer with a different system as the main system or at a different time allowance.

In 1960 the second stage of the project got underway: a test program of searches of the four files for documents in answer to questions submitted by cooperating scientists and technical persons in various organizations. Each question was based on a particular source document, the documents being spread over the whole collection. The major effort was concentrated on the final subprogram of 6,000 documents, with 75 percent of the searches there and 25 percent spread over the earlier subprograms. In the first round of testing, 400 questions were used, making a total of 1,600 searches. Problems involved determination of what constituted a separate search, how many searches could be made, and how general a search was justified. In the second round, rules were devised to solve these problems, and another 400 searches were made in each system. A detailed analysis was made of the searches which ended in failure to retrieve the source document; in many cases the failure was due to the fact that the person making the search had not found the correct program for the particular system. In the third round of searches, again involving 400 questions, the variable of searching was eliminated as far as possible; in the end all systems had been equally treated with regard to search programs.

The combined results of all searches led to the general conclusion that no one of the four systems tested displayed significant superiority over the others. The findings indicated a high degree of recall ranging from 73.8 percent (facet classification), 75.6 percent (UDC), 81.5 percent (alphabetic subject index), to 82 percent (Uniterm). The percentage of retrieval for all systems improved, however insignificantly, with an increase in the indexing time; however, it would appear that no significant improvement in indexing occurred beyond the 4-minute time limit. A rather detailed analysis of the reasons for failure to retrieve relevant documents was made; personal errors in indexing and searching, too little time allowed for indexing, and faulty question formulation or analysis all contributed to these failures. Again no significant difference among the four systems was apparent. The major single cause of retrieval failure was human errors and omissions in indexing or searching.

The performance measure used in the study is the relevance ratio in combination with the recall ratio. Recall ratio equals the number of relevant documents retrieved to the total number of documents in the file; relevance ratio equals the number of relevant documents retrieved to the total number of documents retrieved. In supplementary test programs some 79 searches were made, and the recall ratios for each system were plotted against the relevance ratios. An inverse relationship exists between recall and relevance, so that as the recall figure rises, the relevance ratio falls; conversely, as the recall figure drops, the relevance ratio improves. The supplementary tests together with tests on other existing systems made possible the tentative conclusion that the working level of

information retrieval systems appears to be in the general area of 60-90 percent recall and 10-25 percent relevance.

Since a purpose of the project was to develop techniques for the testing of information retrieval systems, an attempt to check the validity of the test method was made by conducting tests on the facet catalog of the library of the English Electric Company, Ltd., at Whetstone, and on the index to metallurgical literature at Western Reserve University (cf. I-1). At the English Electric Company, a total of 186 questions was collected and typed on to search cards; then searches were carried out in the catalogs at the Whetstone library. A number of these searches was duplicated by members of the English Electric staff. The searches were continued until the source document had been located or until no further search programs could be devised; the searches were considered on this basis as being either successes or failures.

The source document was retrieved altogether in 161 cases of 186 searches, with a success rate of 86.5 percent. Detailed analyses of all the failures show that the reasons for failure were broadly the same as those for the Cranfield project. The efficiency fell within the range of the systems tested in the Cranfield work, but showed an improvement (77.4 percent as against 71.2 percent) when compared with the Cranfield facet catalog. The comparison for failures shows remarkable consistency. The errors due to the system were in both cases mainly concerned with difficulties caused by the chain index. The chain index seemed to hamper unduly the facet system; permutation of the elements at the discretion of the indexer would probably give better results.

In general, the results of applying the test procedure to existing operating systems compared rather consistently with the results obtained in the experimental situation. Despite criticisms that have been directed chiefly at the experimental design, test procedures and methodology, the ASLIB-Cranfield study stands as the first large-scale experiment which tried to measure retrieval effectiveness.

I-3. The CleverdonWRU Experiment: Conclusions, by Cyril W. Cleverdon, in Information Retrieval in Action (Papers presented at 1962 Conference at Center for Documentation and Communication Research, Western Reserve University), Cleveland, Ohio, The Press of Western Reserve University, 1963, p. 101-107.

The testing program, based on Cranfield techniques, attempted to measure recall and relevance for the WRU information system. The quality of concept indexing and the search programs and the power of the descriptor language were under test. The recall figures were surprisingly low, although they were later improved after examining the reasons for failure (cf. I-1). The fault may lie not so much with the search programs as with the indexing, which was done by one highly trained indexer at Cranfield but by several subject specialists at WRU. These tests should give more information on the validity of the test questions as well as on the importance of the indexing step.

I-4. Comments in Discussion of Symposium on Advanced Methods in Information Storage and Retrieval, by Allen Kent, in Information Processing 1962, Ed. by Cicely M. Popplewell, Amsterdam, The Netherlands, North-Holland Publishing Company, 1963, p. 296.

(Included in the discussion following the symposium are the following comments pertinent to this bibliography.)

The ASLIB-Cranfield test questions (cf. I-2) are artificial, made up after examination of a document. Realistically we cannot assume that the questioner has ever seen the needed document. Many of the Cranfield questions were derived from the titles or first few sentences of the document, whereas few real questions are so derived. The work is based on the words of authors, not necessarily corresponding to the way in which questioners make inquiries based on their own backgrounds. More important would be research on relevance, which would be an attempt to determine what correspondences exist between the text of documents retrieved successfully and the text of the question.

I-5. Review of ASLIB-Cranfield Research Project: Report on the First Stage and Interim Report on the Test Programme, by John O'Connor, Journal of Documentation 17, 252-261 (December 1961).

It is worth mentioning, when attempting to draw conclusions from this study, that the results do not apply to indexing by technical people, that a descriptor system was not one of the systems tested, and that time required for developing each index and maintaining its growth was not recorded. However, these omissions do not prevent comparative testing of the factors which were considered.

One question that arises is, are the experimental systems fair samples of their kind? In the experiment, there was the factor of stopwatch strain. There was continual shift of indexing perspective among the four systems. The indexers did not do any searching of the collection during their two years of indexing. There was no checking of the indexer's work by others. All these circumstances might have contributed to making a Cranfield index poorer in quality, or better, than a nonexperimental index. Could the design of the experimental indexing have been made more like real-life conditions? Real-life indexing systems have so many characteristics which might affect indexing quality that an experiment cannot take account of all of them. The systems being compared in the experimental case are more comparable in such factors as documents, indexers, indexing time, etc. The close study of successes and failures of the various systems can produce valuable insight into how the systems work or fail to work. In summary, the experimental indexing is unavoidably imperfect but nonetheless worthwhile.

The basic question on the test program is, do the test searches differ from searching of nonexperimental indexes in ways which might affect the test results? Thus, it is not always possible to determine from a request alone, without consultation between questioner and searcher, what documents would be acceptable as answers to the request and therefore how best to plan the search program. Also, each Cranfield question was based directly on a particular document, and the language of the questions was probably unusually close to the language of the document, and therefore perhaps unusually close to the language of the indexing of the document. This factor might affect the performances of different indexes differently; it might also affect the results for various indexing times differently.

Can a good test of a system's effectiveness be described? We can say roughly that a test of good retrieval is whether it provides everything that the requester is glad to have, weighted by the amount of human skill required to formulate search questions and to examine selected documents. But how do we evaluate different systems if the requester finds no gross differences between their document selections? Perhaps we should judge them as about equally good. But the requester's judgment about them might change, or his colleagues might make somewhat different judgments about the relevance of retrieved documents. Thus it is not completely clear what good retrieval is. But if an acceptable test for good retrieval should be developed, how good is the retrieval for each index when operating under nonexperimental search conditions? Variations in real-life conditions might not affect the performance of the different indexes significantly, but we cannot tell in advance if this is so.

These critical comments are made in hope that they will be of use in the design of better tests of experimental indexes. Small-scale analytic studies of the Cranfield indexes should be conducted in addition to the larger-scale statistical studies described in the reports.

I-6. The Cleverdon-WRU Experiment: Search Results, by Alan M. Rees, in Information Retrieval in Action (Papers presented at 1962 Conference at Center for Documentation and Communication Research, Western Reserve University), Cleveland, Ohio, The Press of Western Reserve University, 1963, p. 93-99.

For 1,300 documents on tape at Western Reserve University, 137 questions, each based on a document in the test file, were compiled by the Cranfield group. The WRU staff then analyzed the questions and formulated search strategy to identify the source documents. Results given here are based on 125 searches made, of which 104 were successful on the original or an alternative search strategy. This represents 83 percent recall. Detailed analysis of failures showed they were due either to judgment errors (indexer judged a concept as of minor importance in the source document, the semantic code did not carry the needed association, the search programmer specified incorrect punctuation levels, or concepts were selected incorrectly) or to nonjudgment errors (keypunching or machine procedure errors).

The synthetic nature of the questions caused some problems in the test: the subjective judgment of the WRU indexers had to be balanced against the subjective interpretation of the document by the Cranfield question compilers; also, the relevance of items other than to the source documents is artificial and its relation to relevance in searches based on real questions requires further investigation; in addition negotiations with the question-asker, sometimes necessary to clarify and define search strategy, was precluded. Another problem is connected with the assessment of relevance of documents other than the source document; there seems to be little uniformity in judging relevance of documents to specific questions. These factors might have an influence on the final results. This project should contribute to the development of an effective test for comparative evaluation of information retrieval systems.

I-7. Review of a Report of the ASLIB-Cranfield Test of the Index of Metallurgical Literature of Western Reserve University, by Alan M. Rees, Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, October 1963, 32 p.

(The material presented in this report has also been published as "The ASLIB-Cranfield Test of the Western Reserve University Indexing System for Metallurgical Literature: A Review of the Final Report", by Alan M. Rees, American Documentation 16, 73-76 (April 1965)).

The principal contribution made by the Cranfield report on a test of the metallurgical literature index of WRU (cf. I-1) lies in the detailed descriptions and analyses of the individual failures to retrieve relevant documents and reasons for retrieving irrelevant documents. Appreciation is expressed for the valuable data amassed in the report and for the diligence and insight with which the results were analyzed. The performance figures from the test raise the question, why did the WRU system, with exhaustive and specific indexing and syntactical relationships, not perform substantially better than the faceted catalog?

Evidence seems to indicate that the trading relationship between recall and relevance does in fact exist; interfixing is preferable to role indicators; the effort put into the semantic code seems hardly justified; a computer is not more effective as a retrieval tool than a card catalog; and the devices used to ensure high recall and relevance have worked against good performance by the system. However, an operational system is more affected by the problem of consistency of indexing than is an experimental file. Also, many of the search failures at WRU can be attributed to some artificiality in the processing of the questions into search programs, since the analysts attempted to predict generic-specific levels and associated concepts, without benefit of discussion with the questioners.

Further research is needed before conclusions about the WRU semantic code and other features of the indexing language are justified. Tests of indexing languages must take into account the structure of the language and the extent to which devices built into the language contribute to the recall and relevance scores achieved in searching; also the consistency and effectiveness with which the indexing language is applied to documents must be considered. Although much work remains to be done, the pioneering efforts at Cranfield have been extremely valuable.

I-8. Review of the Cranfield Project, by Phyllis A. Richmond, American Documentation 14, 307-311 (October 1963).

In the testing described in the report (cf. I-1), three major difficulties were encountered. First, the question had to be answered as asked. In contrast to a real-life situation, no further inquiry for clarification could be made. However, this difficulty applied to all systems and cannot be considered as seriously affecting the comparative results. The other two difficulties involved procedure: What constituted a single search? How long was one justified in searching? As might be expected, the project turned up a human factor in searching success and failure. A standardized procedure was adopted to try to overcome this factor, but the description of it is not clear enough to explain what it was or why it was an improvement.

The report includes a thorough analysis of failures; the largest failure factors were personal errors (36%) and too-short time allowance (22%). The personal errors are analyzed in detail, but the time allowance errors are passed over on the assumption that there would have been none if more time had been allowed. For errors of "insufficient indexing" this is a valid assumption, but what proportion of "careless indexing" errors were due to stopwatch strain? The fact that two-thirds of the failures could be attributed to faulty input indicates how important this aspect is to any kind of retrieval. The failures due to searching were caused largely by lack of understanding --- a matter which would have been settled with the requester in a real-life situation --- and by insufficient searching.

Something that is noticeable in many parts of the report is that the test results could have received better interpretation. A summary of the effect of this report is not easy to make. One is overwhelmed by figures, and it is not always clear what they stand for. In the last analysis, which indexing system was the most efficient? The answer given, that one system is best in one situation and another in another, hardly satisfies the purpose of the experiment or does justice to its results.

These results will probably be worked over for some time to come. Not all of the possibilities have been exhausted yet. A major step has been taken in compiling a wealth of data which will be available as new questions arise. The project deserves to be continued but preferably with some improvement in numerical methods of reporting.

I-9. Review of Aitchison-Cleverdon, Alan M. Rees, and National Academy of Sciences-National Research Council Reports, by John R. Sharp, Journal of Documentation 20, 170-174 (September 1964).

The Cranfield report (cf. I-1) shows the care which was taken to make this test as fair and objective as possible. The conclusions are amply supported by factual findings. Whether the basic philosophy of Cranfield's testing methods is sound and whether the interpretation of the results provides a proper picture of relative performances are discussed in this review. Apart from the findings on this particular retrieval system, there are several issues of general application and interest about which some very valuable information was discovered. One of these is the relative effectiveness of role indicators, interfixing and levels. The effectiveness of interfixing in avoiding false drops is very significant. Of 85 false drops analysed, interfixing was potentially capable of preventing 75 percent, role indicators 59 percent, and levels 15 percent.

The criteria by which the efficiency of retrieval systems must be judged are the two concomitant variables, recall and relevance. It is disturbing that Cranfield should be so committed to the inevitability of a reduction in one if the other is increased. The recall/relevance curve should be regarded as a picture of what happens in a given system with controls of a given standard when search programs are broadened or narrowed. Curves can only show what happens when we use a given system in different ways. The general levels of performance of different kinds of system should appear as points in the recall/relevance square. The salient findings of this investigation are two pairs of figures --- a recall figure and a relevance figure for WRU and a recall figure and a relevance figure for Cranfield. These provide two pairs of ordinates which fix two points in the diagram. We should be quite clear what the Cranfield curve really means, and what we have learned so far should not deter us from trying to step off that curve for the top right-hand corner of the square, for that point should be our ultimate goal.

Mr. Rees attempts to explain away the reasons for the poor showing of the WRU system (cf. I-7). A careful examination of data produced by the tests seems to suggest that he has not fully exploited the evidence which he has. First, his argument that a multiplicity of indexers militates against consistency certainly has some substance. Secondly, with regard to the artificiality of the questions, the use of source documents as relevance criteria is suspect, for rematching concepts in a system with concepts in a question when the source of those concepts is a common one in the form of a single document is a doubtful measure of the efficacy of a system. There is factual evidence regarding the validity of the Cranfield method, and the report itself provides the figures.

The documents were all assessed for relevance, and each document was graded as either as relevant as the source document (Relevance 2) or as of some, though less, relevance (Relevance 3). The overall recall figures for WRU and Cranfield were 82.4 percent and 90.3 percent, based on the recall of source documents only. When the figures are presented for the recall of source documents plus Relevance 2 documents, the gap between the two begins to narrow. When Relevance 3 documents are also taken into account, the tables are turned, for WRU's recall is now 75.8 percent and Cranfield's is 69.5 percent. The significant thing is that the relevance to a question of a document other than the source document was assessed after the question had been asked.

This is clearly a better simulation of natural working conditions than is the use of the source document criterion. It is surely logical to argue that if the figures for one system against another converge and are eventually completely reversed when conditions which are known to be natural are introduced progressively into the test, then the original figures were derived from conditions which were artificial. That part of the test which more nearly simulated natural working conditions is the only evidence in the Cranfield report which has validity and the proper conclusion should be that the WRU system is better than the

Cranfield system at least as far as recall is concerned.

The third report covers three separate activities pursued by an ad hoc committee set up by the Academy (cf. II-37). The first project was the initiation of operations research studies to devise means of evaluating information retrieval systems; the second was the testing of the WRU searching services by comparing the WRU system with a number of other systems serving the field of metallurgy; and the third project was a survey of users' reactions to the service given by WRU. In the test a number of actual questions (including current awareness questions) asked at WRU was used, and each question was put to the WRU system and to one of the other participating organizations. The principal figures reported are the number of retrieved documents sent by WRU to the reviewer, the number of items accepted by the reviewer as relevant, the number of items retrieved by the parallel searcher, the number of items common to both searches, and the number of items retrieved by the parallel searcher which were in WRU's system but not retrieved there in the search.

One set of figures which the committee apparently did not acquire was the number of items which WRU retrieved which were in the parallel searcher's file but which that searcher did not retrieve. It seems that WRU did quite well on recall; its figures for retrospective search often showed a much larger number of documents retrieved, with the number judged relevant exceeding the total found by the parallel searcher. One remarkable aspect is the low figures for common items retrieved by WRU and the parallel searcher. The relevance of WRU is clearly not good, but the committee and the users rate relevance of less significance than recall. The report has a section detailing some studies of searching strategy which hints at what might be done, for experiments seem to show that improvements can be made.

As far as the performance of the WRU system is concerned, the results of the work reported in these three documents are inconclusive. What is clear is that we have a lot to learn about how to test systems. As far as thoroughness and general competence in carrying out the work of testing are concerned Cranfield has no peer, but the source document principle should be dropped and future tests carried out on the basis of taking into account all relevant documents retrieved.

I-10. The Evidence Underlying the Cranfield Results, by Don R. Swanson, The Library Quarterly 35, 1-20 (January 1965).

This review is principally concerned with the first Cranfield project (cf. I-2) and emphasizes the implications of its experimental design. It is contended that the design itself almost guaranteed the particular results that were found.

There has been considerable dispute and discussion on the "artificial" nature of the questions used in testing the four indexing systems. Rather the "artificial" or "biased" nature of the relationship of the question to the source document should have been emphasized. Specific documents may certainly be used as sources of questions, but any meaningful tests of the retrieval system must be performed on new documents which can be assumed free of any unusual or direct influence on the wording or nature of the question.

In a real situation a "source document" generally does not exist, and when it does it may presumably exhibit a variety of subtle relationships to the question that it inspires. It would be quite unlikely that such a source document would be the only document of interest in response to a question. It should be noted that the Cranfield reports make no effort to conceal or distort the facts of their experimental design; it is only the consequences of such design that are explored here.

A close relationship exists not only between the source document and its corresponding question, but between the title of the source document and the question. The consequences and extent of the close correlation in language between titles and questions have been overlooked in most of the Cranfield literature. It might be informative to show what degree of retrieval success could be achieved by a machine match of title to question. Consider a mechanizable matching procedure in which recall of relevant documents is calculated only for those cases in which the statistically expected number of irrelevant documents also retrieved will be small. This could be done if good data were available on word and phrase frequencies for the collection of titles searched.

For a rough approximation, a count was made based on the 100 titles listed in Appendix 4B of the report (cf. I-2). Each of these 100 questions was then matched to its corresponding source title. The recall figure for successful recovery of source documents (plus reasonably low retrieval of irrelevant documents) based on statistical matching of title to question is 85 percent for the 100 questions and documents that were analyzed. If care were taken to incorporate an adequate sample of the language of the subject field in each of the four indexing systems tested, nothing else should be required to obtain 85 percent recall. It might be of greater interest to compare the four systems on the basis of their ability to retrieve information not susceptible to being retrieved by matching title to question. All of the systems may operate at a level well below the recall achievable for those cases in which the question language is closely matched to the title language.

Consider the a priori probable consequences of experimenting with retrieval of source documents with high question-title resemblance. These probable consequences are divided into the four areas which the Cranfield reports consider most important:

1. Recall comparison -- The terms used in each of the systems tested were specifically compiled to reflect the nature of the collection. One would expect a fairly high degree of uniformity in search results. What Cranfield must have tested was their own consistency in equalizing the four vocabularies.

2. Number of indexing terms -- If an indexing system is to be tested by questions based largely on titles, there would be little point in indexing very much more than titles. It is hardly surprising to find Cranfield concluding that little improvement results from "detailed indexing." The various assessments of errors due to over-detailed indexing and indications that documents were missed because of a failure to index the title further illustrate that the Cranfield authors tend to at least imply a more general significance to their findings on detailed indexing than is justified.

3. Indexing time -- The time required to index an article is related to the number of terms assigned. Any indexing time expended beyond the few minutes required to index the title could not result in significant improvement of retrieval success for source documents. Thus "efficiency of 4-minute indexing time" hardly merits the importance with which it is presented.

4. Subject skill of indexers -- It seems unlikely that the recognition of index terms contained in the title of a document would require great knowledgeability in the subject matter of the document. It is difficult to see what significance can be attached to the level of technical training the indexers did or did not have.

The design of the Cranfield project suffers a major flaw in standardization of the searching procedure for the four systems being tested. The final process of human scanning of a batch of catalog cards is an important part of the experiment, and the average number of cards examined per search question is an important property of any indexing system. The comparison of systems by testing their success in retrieving source documents can have no meaning unless one standardizes to some common number of catalog cards to which each system leads or tabulates the total number of cards finally searched

for each question. The variation in the average number of cards scanned per search could have been between 25 percent and 100 percent. The uncertainty in results implied by the lack of search control indicates that just about as much statistical significance could be attached to controlled experiments with a dozen searches and a hundred documents as was realized for a thousand searches and six thousand documents.

Since the same people in the Cranfield project performed searches as did the indexing, it is virtually impossible to assess the effect that human memory might have played in the process. The experiments inextricably mixed together tests of human memory with tests of indexing systems. The appearance of another uncontrolled element illustrates again the numerous pitfalls of experimental design in this field.

It is pertinent to look at the Cranfield project in the light of those experiments which did in some way deal with nonsource documents and which accordingly attempted to deal with the question of relevance. The project proceeded on a modest scale to collect bibliographies, externally compiled, of articles of probable relevance to each of 88 selected questions, and then selected those references (359) which had also been indexed in the Cranfield project. Each of the 359 documents was then assessed in relation to the appropriate question, rated as 1 if it was useful as the source document, 2 if it was of some interest, and 3 if it was of no interest.

There were 53 documents which had a rating of 1 and 67 with a rating of 2. Thus 120 documents out of 359 were retained as being relevant to the original questions. Why the others were judged as not being of any interest is unclear. Was this filtering based largely on titles, for example? This question is obviously crucial to the subsequent interpretation of the results. The reporting of recall ratios without any reference to the "relevance ratio" or to the "number of cards scanned in order to arrive at such recall" prevents comparison between these recall figures and those given earlier for source documents.

An analysis of 120 questions searched by Cranfield and WRU is given in this report (cf. I-2). An attempt was made to analyze all the documents in the collection in relation to each question in order to come to some relevance judgment. If such judgment has reasonable validity and consistency, it is possible to draw some interesting conclusions regarding the recall and relevance ratios of Relevance 3 nonsource documents. Whatever tendency was exhibited in the Relevance 2 documents for relevance to improve as one decreases the number of indexing entries essentially vanished. Furthermore, the recall ratios are much lower than were reported for the source documents.

In another Cranfield report (cf. I-1) the WRU results are compared with the Cranfield results on the basis of both recall and relevance. Whereas the Cranfield recall (85.3 percent) was somewhat higher than that for WRU (81.4 percent) for Relevance 2 documents, this was not so for the combined Relevance 2 and 3 documents. This is a point worth noting that does not seem to have received attention in the Cranfield literature.

Notwithstanding the emphasis here on the shortcomings of experimental design, it is true that more is known now about indexing systems and about the design and control of retrieval experiments than was known before Cranfield.

I-11. Investigation of Systems for the Intellectual Organization of Information, by Susan Artandi, Report to the National Science Foundation. New Brunswick, N. J., Rutgers, The State University, Graduate School of Library Service, June 1964, 40 p.

Systems for the organization of materials for the effective retrieval of information contained in those materials are examined individually, with the aim of preparing reasonably uniform descriptions and discussions of the systems in terms of characteristics, advantages, disadvantages, and limitations. A set of descriptive criteria applicable to information systems in general can thus be created, and such uniform descriptions can serve as a useful basis for comparative studies to show which system is most efficient under a given set of circumstances.

Characteristic system features include input to the system, materials in the system (the store), searching methods, and output of the system. Input characteristics, including methods by which the index language is created, choice of index terms, skills needed for indexing, physical processing of materials, and others, are compiled for the Universal Decimal Classification, SYNTOL, and alphabetic-specific subject indexing. (These three systems were presented in a seminar series at Rutgers; each presentation was followed by a panel discussion with audience participation. The points brought out in discussion were incorporated in the final systems descriptions.) A tentative conclusion is that the method of preparing uniform descriptions is reasonable and useful.

I-12. Three Experiments with Citation Indexing and Bibliographic Coupling of Physics Literature, by Pauline Atherton and J. C. Yovich, New York, American Institute of Physics, April 1962, 39 p.

The three experiments discussed were (1) the production of an index to authors in footnote references in one issue of The Journal of Chemical Physics; (2) the compilation of a citation index for articles published in The Physical Review on the topic of photodisintegration of the deuteron and general field theory, during the period from 1956 to 1960, starting with one paper published in 1961 and covering only references in Physical Review articles to articles in the same journal; and (3) application of M. M. Kessler's bibliographic coupling method to the citation index/bibliography compiled in experiment (2).

The 64-item bibliography was evaluated by a nuclear physicist who found the titles alone insufficient to make any decision with respect to relevancy or pertinency. An incomplete survey of all papers in Physical Review from 1956 to 1960 revealed six additional pertinent papers not turned up by the citation index system. In the bibliographic coupling test, 23 of the 64 items on the test list were dropped because they did not share a reference with at least one other paper. The nuclear physicist who evaluated the subgroups of papers generated by applying coupling criteria felt that such groups formed incomplete bibliographies and that one could not trust the system's ability to retrieve the pertinent literature.

I-13. The Analysis of Medical Documents with a Comparative Evaluation of Three Indexing Procedures, by William C. Buscher (Paper presented before Rochester Conference on Data Acquisition and Processing in Biology and Medicine, July 1963) Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, 13p.

Document analysis procedures can be grouped into four general classes: (1) human analysis for human retrieval, (2) human analysis for machine retrieval, (3) machine analysis for human retrieval, and (4) machine analysis for machine retrieval. At least four criteria must be applied in an evaluation of techniques used for document analysis --- cost,

convenience, completeness, and accuracy. A limited evaluation of indexing procedures used the criterion "completeness" in comparing three procedures representative of the first three classes above: subject-heading index, telegraphic abstracts with semantic codes, and permuted-title index.

From the Center's file of diabetes-related literature published in 1960, 22 articles were selected at random; references to other articles found in them were compiled, resulting in a list of 222 articles. Subject headings from National Library of Medicine's Medical Subject Headings were applied for the first procedure with an average of five subject headings used per article. In the title index, articles appeared under an average of 4.5 keywords. The average telegraphic abstract had 25 words with appropriate role and link indicators. The 22 articles selected in the beginning were studied closely to obtain questions for which the cited articles could be listed as answers, and 172 questions were obtained, of which 30 were selected for the test. Four volunteers answered the questions in the manual indexes, and staff members at the Center programmed the questions for a computer search of the abstracts. For the 30 questions there were 123 acceptable answers.

Articles were considered "right" if they were the ones used by the citing authors and "wrong" if they had not been so used (disregarding their possible bearing on the author's needs). Completeness (ratio of number of right answers found to total number of right answers) and accuracy (ratio of number of right answers found to total number of answers found) were computed: subject index, 62.6 percent and 36.1 percent, respectively; title index, 57.7 percent and 37.1 percent, respectively; and telegraphic abstracts, 56.1 percent and 38.1 percent, respectively. (The figures for the two manual systems include answers guessed to be right by the searchers.) The results raise many questions, including why were the machine system scores low, why were some answers missed, and what sort of information is required for an effective search. Continuing testing and experimentation at the Center are aimed at answering such questions.

I-14. Seven Years of Work on the Organization of Materials in the Special Library, by C. Dake Gull, American Documentation 7, 320-329 (October 1956).

The Armed Services Technical Information Agency (ASTIA) Reference Test is described in detail in this paper. In June 1953 ASTIA approved a comparison to determine the relative performance of the (then new) Uniterm System for information retrieval. A coordinate index for 15,000 documents in the regular flow of material to ASTIA was prepared by Documentation, Inc. (DI), which posted an average of eight terms per document onto terminal digit cards. At the same time the ASTIA Reference Center (ARC) prepared its regular subject headings for the same documents.

Ninety-eight requests for report bibliographies routinely received by ASTIA were searched in the two systems. In addition to about 580 reports listed in common, each group found that the other group had included some other pertinent references, increasing the area of agreement to 1,390 common reports. ARC accepted 492 of DI's reports as relevant, and DI accepted 318 of ARC's reports as relevant. The retrieval of these 1,390 reports would have been the optimum performance for the independent search. ARC produced 64.6 percent (580 + 318) of the optimum; DI produced 77.1 percent (580 + 492) of the optimum.

In addition to testing the character of the retrieval system, another variable recognized as increasingly important had to do with the difficulties of interpreting a request. ARC cited 1,089 reports which DI considered irrelevant, while DI cited 488 reports which ARC considered irrelevant. The difference of opinion could have been resolved by submitting both lists to the requestors and obtaining their opinions as to which method was more

effective. Some account could have been taken of such questions as the following: Was the written request an adequate reflection of the user's question? Was the user's idea of his question changed by the lists he received? However, the suggestion that both lists be submitted to the requestors was not followed, so no independent judgment was obtained on the satisfactory character of the reference supplied.

Still another variable that came to light was the skill with which the systems were used. Users of both systems had difficulty. The 492 references which ARC accepted from DI had been missed because it failed to search all applicable subject headings, because subject heading assignment was inadequate, because the interpretation of the request was inadequate, because there were so many cards under the subject heading as to discourage the searcher, and for other varied reasons. The 318 ARC references accepted by DI had been missed because of clerical errors in posting numbers, errors in coordinating numbers, inconsistencies of indexing, and failure to consult the proper terms.

A fairly thorough search of the literature made at the time indicated that this comparison of two reference systems was the first undertaken, so the clerical errors and incomplete design of the test revealed during the course of the work were not surprising.

I-15. Convertibility of Indexing Vocabularies, by William Hammond, in Proceedings of Conference on the Literature of Nuclear Sciences: Its Management and Use, Oak Ridge, Tenn., U.S. Atomic Energy Commission, Division of Technical Information Extension, December 1962, p. 223-234.

(Some of the work reported here appears in greater detail in "Experimental Study of Convertibility between Large Technical Indexing Vocabularies--with Table of Indexing Equivalents", by William Hammond and Staffan Rosenborg, Technical Report IR-1, Contract NSF C-259, Silver Spring, Md., Datatrol Corporation, August 1962.)

A study was made to determine the difficulties inherent in applying one technical indexing vocabulary to information previously cataloged in another. The Armed Services Technical Information Agency (ASTIA) descriptor vocabulary was compared with the Atomic Energy Commission (AEC) subject headings; the Table of Indexing Equivalents shows for each of ASTIA's 7,145 descriptors the identical, synonymous, or usefully equivalent counterpart among AEC's subject headings. Indexing equivalents were found for all but 777 or 10.9 percent of the ASTIA descriptors, but these 777 descriptors accounted for only 6.1 percent of the total ASTIA assignment of descriptors over the past eight years.

Conversion by means of such a table will serve as a means of eliminating duplication in cataloging and indexing effort. Frequency of use was highly significant; the most frequently used terms were identical or directly convertible. Thus, the 905 most frequently used AEC headings, while representing only 7 percent of the total AEC vocabulary, accounted for some 80 percent of AEC term assignments. Over half of these headings were identical or directly convertible and the others easily convertible to ASTIA descriptors.

In a small related study, the AEC and ASTIA indexing of 277 reports cataloged by both agencies were compared. ASTIA furnished its catalog cards and AEC furnished copies of the original worksheets. ASTIA used 2,571 descriptors and AEC 840 subject headings; of the latter, 392 were either completely identical or nearly so, including matches achieved by coordination of ASTIA terms. In 59 documents all AEC terms were matched by ASTIA terms. Time was not available to make a detailed study of equivalence for the remaining terms, but it was apparent that conversion by means of the Table of Indexing Equivalents would substantially have generated the terms AEC used.

Potential areas of future work include (1) field and group setting and (2) multi-directional equivalence. Comparison of vocabularies was facilitated by using the ASTIA field-and-group subsumption scheme. Generically related terms became more obvious, errors such as in homonym equivalence were less likely to occur, and "best" synonyms were more easily discovered. If more than two vocabularies are to be compared, a composite or new scheme might be required. The Table is unidirectional, but ideally a dictionary of equivalents should lead from any one vocabulary to any or all the others.

I-16. Project Lawsearch: A Statistical Comparison of Coordinate and Conventional Legal Indexing, by Lawrence B. Heilprin and S.S. Crutchfield, in Parameters of Information Science (Proceedings of the American Documentation Institute, 1964 Annual Meeting, Vol. 1, October 5-8, 1964, Philadelphia, Pa.), Washington, D.C., Spartan Books and London, Cleaver-Hume Press, 1964, p. 215-234.

The purpose of the experiment was to index a body of the law for search by the coordinate index method, and to test this method of search against conventional search of the same body of law. The coordination system used was a mechanico-optical system, not entirely manual but less expensive to use than a computer. The particular body of law indexed was that concerned with motor carriers --- in particular, personal injury to passengers and loss of or injury to goods. This index occupied three volumes and an appendix. The present report describes the work which started with delivery of the set of index volumes and is confined to statistical aspects of designing the test, computations of the data and interpretation of the statistics.

There are many potential sources of error or uncertainty in comparing conventional with coordinate legal search. Those which could be forestalled in whole or in part are as follows: (1) The questions used in the test might not fall within the coordinate-indexed area --- removed through prior examination; (2) the searchers might differ in training or ability to search legal material --- group of law review students and professional law librarian selected; (3) searchers might differ in exposure to test material, i.e., a system of rotation of test questions minimized contact among searchers and conflict in use of library materials so that each of 9 searchers investigated each of 8 questions (72 searches), no searcher was assigned the same question twice, and at any time not more than two searchers worked on a given question; (4) samples might not be large enough for clear statistical significance --- sample of 36 searches produced about 200 cases per sample, large enough for reliable results by Chi Square and Variance Ratio tests; and (5) assignments of applicability class of cases might not be uniform --- judges who examined cases retrieved by the searchers were experienced attorneys and professors of law or indexers of legal literature, and each repeated at a later date his appraisal of each case.

Those sources of uncertainty which could not be allowed for are as follows: (6) Selection of descriptive words might not suffice in quantity or quality to insure with high probability that an indexed reference would not be overlooked --- indexing density ran from 15 to 20 terms per case up to as many as 45 to 60 terms, the indexing term vocabulary was about 800 terms, and as a result there was no failure to find cases that had been coordinate-indexed provided they were of the highest applicability; (7) the indexing within the legal area of the test might not be exhaustive --- this could occur if the limits of the field were not well defined (as happened since the coordinate-indexed universe was limited to motor carrier law while the conventional universe might include other transport media if these cases seemed related and applicable), but the effect was not sufficiently great as to render the samples noncomparable or to destroy the essential statistical similarity of the two samples; (8) the searchers might not develop skill in the use of coordinate methods comparable to that which they had in conventional search --- such a possibility could be shown to be either present or absent and some partial correction made for it (demonstration that there was a strong learning curve was one of the highlights of the test); and

(9) the searchers by coordination might not learn to use the specific-to-general approach to search, and their habits would tend to conceal possible speed advantages of coordinate search relative to conventional search --- in law there is a high premium on exhaustiveness, resulting in searching a broader class than needed to locate samples and then eliminating all but applicable answers (the general-to-specific approach) as opposed to starting with the maximum number of descriptors and broadening the search by removing some of them (the specific-to-general approach), such search habit being based on the tendency of the conventional system of hierarchical class indexing to deteriorate with age and to require using broader classes which do not lose their mutual exclusiveness so rapidly. The effect of these factors could not be predicted but a means was found by which the existence of the effect could be demonstrated and to some extent measured independent of the searcher's knowledge.

The parent populations of citations of cases included 1,800 cases referring on the one hand specifically to law on motor carriers and on the other hand to law on motor carriers plus analogous law on other modes of transport. The first consideration was the extent to which the two samples resembled each other, testing the distribution of cases retrieved in each of the four classes defined to characterize the applicability of the cases to the search questions: case in point, applicable on the facts and the law; applicable by analogy, either on the facts or on law, but not on both; dicta, containing no binding decision as to law or facts of future cases; and cases not applicable to search question. The class rating used for the statistics was the rating by the judge who drafted the question and who read all cases found by the searchers in response to it. The distribution of cases in the samples shows the combined result of the labor of nine searchers and the judgment of one person per question as to their success. The samples show no greater differences in distribution than should be expected from random sampling of identical populations. That is, the two samples had enough statistical similarity so that they could be considered as a whole as if drawn by chance from the same populations.

It was found that the probability of retrieving the same case by the two search methods increases with the applicability of the case to the search question. Although all cases of highest applicability were retrieved by both methods, there is a difference between the statistics of the two samples and the statistics of the individuals. In order to be sure that a case in any class was found, on the average it would be necessary to conduct at least three and often four independent searches of the same question. Percentages expressing the probability of retrieval of identical cases were probabilities based on the combined efforts of all searchers in producing the samples. But the probability that an individual would locate all the cases in a class by either method was much lower. Conclusions which could be drawn from the test as to the content of the searches are (1) that the two methods appeared to yield approximately the same quality of cases; (2) that the probability of finding identical cases by both methods was strongly related to the applicability of the cases; (3) that for a case of given applicability, the probability that an individual would locate it was much lower than the probability that a group of several individuals searching independently would locate it; and (4) that on the basis of the samples, the smallest groups required so as to be almost sure to locate cases of highest applicability were two independent searchers and more than two for cases of lower applicability.

When interest changed from what was retrieved to efficiency of retrieval, the data are the times taken to perform the search, or the total search time. The data on total search times were recorded for each individual searcher and for each question; the time by question was summarized, again dividing the time by search method, and combining to give the overall time. The overall mean search time for 36 searches by the conventional method was 137.5 minutes per search. The overall mean for the coordinate method was 97.0 minutes per search. To test the reality of the difference an analysis of variance was made. There is a significant difference in the mean search time by the two methods.

These statistics neglected possible change with time of the search time, i.e., a possible "learning curve" by the searchers in mastering the coordinate search technique. In a graphical plot showing the mean search time as a function of trials, or days, in the order in which they occurred, a sharply decreasing mean search time is evident. Since the curve shows no sign of flattening out it suggests that the five days of trial were not sufficient to master the new search technique and a larger time improvement might be anticipated. It was believed that this was the first time a learning curve had been demonstrated for coordinate search, and curiosity attached to what was being learned.

It was shown that the time advantage could be greatly increased by varying the order in which generality was used in search, and that the order of use appeared connected with the need to relearn this order. It is strongly suggested that the improvement in search time was attributable to gradual appreciation on the part of the searchers that their habits of general-to-specific search, with much subsequent scanning, need no longer be used. Since in the present test situation the obtaining of one or several cases-in-point might well be considered the end of the search, the coordinate search should show some average improvement in search time over the conventional method if the questions were ranked in the order in which cases-in-point were found. It was found that for conventional search there was no great sensitivity of search time to the finding of a case-in-point, but on the other hand for coordinate search there was a marked progressive decrease in average search time as cases-in-point increased.

In answer to the question as to why similar time relations were not discovered in the ASLIB-Cranfield Project (cf. I-2), in testing the Uniterm system the Project did not use optical coincidence, and without the mechanization which gives the coordinate method its speed it did not show significant search time differences compared with other classification systems.

I-17. The Distribution of Term Usage in Manipulative Indexes, by Nona Houston and Eugene Wall, American Documentation 15, 105-114 (April 1964).

A method has been developed for predicting the distribution of use of index terms in manipulative systems which develop their vocabularies empirically. Manipulative indexes are characterized by the coordination of more or less elemental indexing terms at the time of search. In this paper either of two broad types of system is referred to: the "prefiled manipulative" system, in which each term in the vocabulary has posted to it a set of document surrogates, and the "random manipulative" system, in which each of a collection of document surrogates has posted to it a set of index terms. By "empirically" is meant that the vocabulary is not prescribed in advance but is developed on the basis of the terminology employed in the indexed documents. Such vocabularies can be edited, but such editing operations do not detract from the empirical nature of the vocabularies.

Data from 10 indexes were used in the study: a chemical engineering index, a private experimental index, the Air Force's AFOSR Project ECHO, five different samples of the Uniterm Index to Chemical Patents (in order to observe any effects of increase in size of a given collection), the DDC system, and a private research report index. The data were correlated to show a log-normal relationship, up to about the 95th to 98th percentile, between total index entries and distribution of term usage. In nine of the 10 collections the equation expressing that distribution depended principally on the total number of postings in the index. For the remaining index, which employed syntactical controls, the equation contained constraints different from those for the other systems' equations.

The equations can be used to predict the approximate size of a system vocabulary, based upon the number of postings in the index, and to determine whether the admission of new terms to a vocabulary is too tightly or too loosely controlled. Also the rate of vocabulary growth can be estimated. An additional basis for deciding whether or not to employ dedicated-space storage media for indexes is thus provided.

I-18. Comparison of the Results of Bibliographic Coupling and Analytic Subject Indexing, by M. M. Kessler, Report R-7, Cambridge, Massachusetts, Massachusetts Institute of Technology, January 28, 1963, 30 p.

(This report also appeared in American Documentation 16, 223-233 (July 1965).)

Bibliographic coupling is a method for separating a large body of technical literature into small related groups: a single item of reference used by two papers is called one unit of coupling between them, and a number of papers constitute a related group if each member of the group has at least one coupling unit to a given test paper. This report is concerned with a comparison of the results of processes for grouping papers, that is, with the mechanics of group formation but not with the meaning or relevance to retrieval of the groups so formed.

The experimental material consisted of 334 papers in Vol. 112 (1958) of Physical Review. These papers were arranged into groups according to analytic subject indexing (ASI) and bibliographic coupling (BC). In the first case, the papers were placed into one or more of 73 subject categories; there were 73 groups, although some were empty. In the BC process, each paper was compared to each of the remaining; there were 334 groups, although some of them were empty also. In the ASI, the collection was distributed among a smaller number of groups, and each paper was a member of fewer groups than in BC. In the case of BC, the process is completely mechanical; in ASI much is left to the discretion of editors and indexers.

The comparison was performed in four stages: (1) Given the papers in a group according to BC criterion, how are they regarded by ASI; (2) given the papers according to ASI, how are they regarded by BC; (3) and (4) given two papers strongly related according to one of the methods, what is the verdict of the other method. In the first stage, each of the 334 papers was used to form a group according to BC; each of the papers in these groups was then checked for its classification according to ASI. If all the members of G_A (groups formed by BC) could be accounted for by three ASI categories, there was considered to be a good match between the two methods. Data were calculated as A/B and C/D, where A = number of papers in G_A, B = largest number of papers in G_A included in three ASI categories, C = total number of ASI categories used to describe all the papers in G_A, and D = number of ASI categories that would account for all the papers in G_A. The ratio of the two fractions A/B and C/D may be used as a measure of the correlation between the results. The average value of A/B for all BC groups having four papers or more was 0.9; for the same groups the average value of C/D was 0.5.

In the second stage, results could not be compared statistically since there is no measure relatedness for ASI categories. But the qualitative conclusion is drawn that correlation of ASI categories with BC groups is good where the ASI categories seem to be of small enough "logical size" (e.g., refer to a particular state or property rather than to a major discipline, etc.).

In the third stage, pairs of papers formed between a test paper and each member of G_A are considered. There were 876 pairs of papers formed variously coupled. All those with five or more coupling units were placed in one group, those with three and four units in another, then papers with one and two coupling units each formed two separate groups. The pairs were then examined for relatedness by the ASI method: if the pair shared one ASI category, a relation exists. Coupling strength is a measure of relatedness consistent with the judgment of ASI. As the coupling strength increases the probability of sharing an ASI category also increases.

The fourth stage experiment could not be performed because no measure of strength of relatedness between pairs of papers could be found for the ASI method.

I-19. Evaluation of Analog-to-Digital Converter Patent Information Retrieval Systems, by Donald W. King, WRA PO 12, Denver, Colo., Westat Research Analysts, Inc., June 1964, 42 p. plus appendices.

Three index and search systems were involved in this study: S_1, the U.S. Patent Office classification system subclass 340/347 (called "Code converters" and including analog-to-digital, digital-to-analog, and digital-to-digital converters), which is searched manually through file drawers; S_2, the U.S. coordinate index system containing a duplicate set of the U.S. patents identified above and comprising 89 terms grouped into nine categories, which is searched by needle sorting on key-sort cards and manual retrieval from the files; and S_3, the Netherlands Patent Office coordinate term index system comprising all patents dealing with pulse-code modulation systems disclosing converters and using 356 terms arranged in nine main categories, which is searched by passing IBM cards through a tabulator. Of the 750 U.S. patents in S_1 and S_2 and the 630 U.S. patents in S_3, only about 350 were the same patents.

Search results include 155 searches performed under S_2 operationally and a total of 30 experimental searches performed on all three systems (30 were searched under S_1 and S_2, and 21 of these were searched under S_3). In experimental searches for S_1 and S_2 the examiner selected the broadest and narrowest search question to be used in S_2, search of the broad question was made by another person, search of the S_1 was made by the examiner, and the documents from the two searches were merged and judged for relevance as to citable in the case, useful in the case, and prefer not to have. A search of the narrow question in S_2 was made by another person and those documents identified as to relevance. Since assessment of relevance was made on the broad set of documents, it was possible to draw conclusions concerning the effectiveness of the narrow searches by observing whether or not the relevant documents were retrieved. This procedure allowed the examiner to pass judgment on both systems for one search, or to evaluate the two systems on a common ground. For S_3 a number of search questions were formulated, averaging 3.9 search questions for each application, and the documents retrieved were screened for those most relevant to the search.

The distribution of the number of documents retrieved was found to fit a logarithmic-normal distribution with S_2. The observed distribution of the number of documents retrieved using S_3 apparently was not log-normal.

One means of characterizing the search procedure is by a retrieval profile, described by the elements of the relevance-retrieved contingency table. The assessment of relevance was made in the first instance by the examiners in each of their searches. The retrieval profile for classification searches is given. There were four independent searches made on the applications used in the experiment: S_1, S_2, S_3, and S_2 operational.

All four systems produced a set of relevant documents. Retrieval profiles are given where the total number of relevant documents is the set sum of the relevant documents found in all the searches. A set of documents was assessed to be relevant from these four independent searches, and a new set of documents was selected as being minimal for citation in these cases. The retrieval profile for this assessment is also given. Again, a set of documents was selected as being the best references for each application; the retrieval profile for this method of assessing relevance is also given. It was found that the recall ratio (number of relevant documents retrieved/total number of relevant documents) is nearly constant within a particular system for all assessments of relevance. Also, the estimated recall ratios for the individual examiners searching with S_1 and S_2 for the minimal set of documents judged to be relevant and the documents judged to be the best references show that the differences between the examiners using one system were not significant statistically.

It was found that S_1 produced at least one best reference in 71 percent of the searches, S_2 in 86 percent of the searches and S_3 in all of the searches observed. The average number of best references per search was found to be 2.0. This model assumes that the best references are of equal value and provide the same information. However, in some instances several references were best references for different claims. In addition, two or more references occasionally provided pieces of information which when combined formed the best reference for a single claim. These varying situations could be treated as separate cases when estimating this criterion of reliability of the system.

The search time was observed for S_1 and S_2. The S_1 search time is subdivided by the time spent manually searching and the time spent making a final selection. The S_2 time is subdivided by the time spent mechanically sorting cards and selecting documents and the time spent making a final selection. The average total search time is greater for S_1 than for S_2. The estimated average time required for final selection is about four minutes per document.

An inspection of search strategies indicated that examiners normally asked a specific question, then broadened their search by dropping off terms. Two additional search strategies were investigated: ranking of terms in order of importance and then searching by sequentially dropping the least important term, which resulted in a recall ratio of 0.67 and an average number of documents retrieved of 48, and assigning a value of relevance to each document according to the number of terms in the question found in each document and then retrieving the documents in the order of this relevance, which resulted in a recall ratio of 0.70 and an average number of documents of 53. It was felt that the broad coordinate index search was more efficient than these two.

During the encoding operation 21 patents were encoded by three examiners and the results evaluated to yield a "true" selection of codes for these documents. The accuracy of encoding was described by the conditional probability that a term is selected given that it should be and that a term is selected given that it should not be. The consistency of encoding was described as the set intersection of the selections of two coders divided by the set sum of the selections of two coders. The best references not retrieved in the experiment were investigated to determine if they were missed because of encoding errors; no misses were distinctly attributed to indexing errors, but inadequacies of the term list accounted for missed references and would also have resulted in reduced indexing consistency.

The following general conclusions were reached: the U.S. coordinate index system is an improvement over the manual classification system, principally in reducing search time, and its reliability is at least as good and somewhat better if used properly; the art in this particular field is amenable to a well structured term list and the term descriptions are relatively unambiguous; three methods of improving the system are modifications of the term list, improvement of search strategy (which has the most potential), and adoption of a sophisticated computer searching technique; the Netherlands system performed well, missed fewer than 20 percent of the relevant documents on the average, had more breadth and depth in the structure of the terms, and allowed more freedom in selecting alternative search questions; the evaluation proved useful in pinpointing weaknesses in the operational systems and in establishing the effectiveness and demonstrating areas of potential improvement in the new system.

I-20. A Comparison of Keyword-in-Context (KWIC) Indexing of Titles with a Subject Heading Classification System, by Donald H. Kraft, American Documentation 15, 48-52 (January 1964).

Keywords, such as would be listed by the KWIC system of automatic title indexing, from 803 titles in the 1959-1960 Index to Legal Theses and Research Projects and from 2,625 titles in the January-June 1961 Index to Legal Periodicals were compared with the subject headings actually used in the conventional human indexing of the papers. The titles were grouped as (1) those which contained a word the same as its subject heading or in some root form of the heading; (2) those which could have been indexed as well under another subject heading which appears in the title; (3) those which contained a synonym of its subject heading; (4) those which contained keywords that would enable a searcher to find it in a KWIC index; and (5) those which contained no descriptive keywords. It was found that 64.4 percent of the computer-generated index entries were identical with the subject headings used. Only 10.5 percent of the titles examined did not contain any keywords which might help to index them.

I-21. A Comparative Study of Fragment Versus Document Retrieval, by J. L. Kuhns and Christine A. Montgomery, in Parameters of Information Science (Proceedings of the American Documentation Institute Annual Meeting, Vol. 1, October 5-8, 1964, Philadelphia, Pa.), Washington, D.C., Spartan Books and London, Cleaver-Hume Press, 1964, p. 369-377.

This experiment tests whether certain types of requests for information could be satisfied by portions or fragments of documents and, if the assumption proved valid, of what size the fragments should be and by which criteria they should be selected.

Four types of fragments were created from a 60-document sample of an experimental library on advanced propellants: (1) a machine-produced document extract, which represents about 25 percent of the total material contained in the document; (2) a paragraph characterized by a dense cluster of search terms; (3) the sentences of the machine extract, taken individually; and (4) the individual sentences of the entire document. Retrieval questions were formulated by project consultants who were experts in the field. The consultants were permitted to base their questions on the abstract cards for the experimental library. The consultants also expanded content terms of their questions by adding synonymous concepts, thus creating thesaurus-type word groups to aid in the retrieval. Terms were defined for each question and specified for retrieval according to the following rules: (1) documents and extracts must contain all terms; (2) paragraphs must contain all terms or, if not, the two paragraphs containing the highest number of terms would be selected;

and (3) sentences would be taken only from documents selected under rule 1 and must contain the primary search term plus one or more secondary terms, according to the number of secondary terms in the search prescription.

Retrievals were carried out manually using a set of three machine-produced concordances prepared for the 60 documents. These concordances listed document, paragraph and sentence numbers for each occurrence of each word. Those documents, extracts, paragraphs, and sentences which satisfied the retrieval requirements were listed. For nine of the original 41 questions, no material could be found in the sample library. The sets of retrieved fragments were ordered for each question in terms of relative size. Machine extract sentences were generally the smallest, being assigned to level 1 in 34 out of 54 cases; the extracts were found to be the largest fragments in 23 out of 33 cases; the document itself was considered as a fifth level of size. For each of the five size levels, a decision was required as to whether the information contained in that level was irrelevant to the question, constituted a partial answer to the question, formed a base from which the answer could be inferred, or constituted a complete answer to the question. For each question the consultants were also asked to indicate by single and double underlines, respectively, material in the body of the documents which (1) was relevant to the question and (2) constituted a minimal complete answer.

The evaluation of the consultants' responses to the retrieval sets was carried out from three points of view. The first evaluation consisted in a study of the relationship between the size of a fragment and its potential for answering questions. Of the 32 questions presented to the consultants for evaluation, 21 were judged to have received a complete answer, 2 contained only irrelevant material, 7 contained material giving a partial answer, and 2 contained enough relevant material to enable the questioner to infer the answer. Of all the questions receiving complete answers 81 percent were answered within an average of 12 sentences (about 16 percent of the document). These figures seem to present a convincing argument for fragment retrieval as against document retrieval, at least for the types of questions used in these experiments.

The second evaluation stressed content rather than size, considering the probability of a given fragment type containing as much information as the full document, containing less information than the full document but some relevant information, being irrelevant when the full document contains information, or not being retrieved when the full document is retrieved. The probabilities, conditional on retrieval, of a fragment type containing as much information as the document are .48, .82, .61, and .69 for extract sentences, document sentences, document paragraphs, and extracts, respectively. The second type (document sentences) dominates in effectiveness, both in regard to size and amount of information.

The third evaluation examined the effectiveness of the fragment types as carriers of relevant material. The procedure is to use a retrieval effectiveness measure (based on "cost" considerations) which is defined as the ratio of the net gain from the retrieval operation. Let the document have a total of t sentences, m of which are relevant; and let the fragment type have n sentences of which x are relevant. The requester's net gain from the information obtained in the fragment type is $x(t/m)-n$. The maximum possible net gain occurs when $x=n=m$ and is, therefore $t-m$. The final fragment type score is then obtained as the ratio of these quantities, $x/m - (1/t-m)(n-x)$. This is of particular interest in that x/m is the proportion of relevant material retrieved and $(n-x)$ is the amount of irrelevant material retrieved. The quantity $1/(t-m)$ is therefore the penalty coefficient for irrelevant material (cf. II-53). The final effectiveness scores were .17, .34, .15, .15, and 0 for extract sentences, document sentences, document paragraphs, extracts, and full text of documents. Comparable scores for the maximally relevant sentences were .28, .54, .30, .27, and 0, respectively (roughly twice as great). The results of the effectiveness measures confirm the probabilistic results.

It was determined that the document sentences fragments have a high probability of containing as much information as the document. However, all four fragment types carry a large proportion of irrelevant material, and a fragment type which could be characterized as optimal with respect to the proportion of relevant versus irrelevant material contained did not emerge in the course of this study.

I-22. Value of Titles for Indexing Purposes, by Robert E. Maizell, Revue de la Documentation 27, 126-127 (August 1960).

The subject index entries supplied by editors of Physics Abstracts (PA) and Chemical Abstracts (CA) for 25 papers from the February 1, 1957 Physical Review were examined in order to study the importance of titles in indexing. PA provided 52 major subject index entries, of which 36 or 69 percent were found given in the words of the title. Sixteen titles or 63 percent contained all of the entries supplied by the indexer. CA indexed 23 of the 25 articles and provided 92 subject headings for them. Of these, 43 or 47 percent were given in the words of the title. Nine titles or 23 percent contained all the information supplied by the indexer. Most of the additional subject entries supplied by CA and some supplied by PA, which were not in the titles, were names of chemical elements or compounds.

Many of the titles studied could have been modified easily so as to include all their indexing concepts; titles are, therefore, closely related to index entries.

I-23. An Evaluation of Information Retrieval Systems, by Max W. Mueller, Memorandum Report No. 7170, Burbank, Calif., Lockheed Aircraft Corp., 30 September 1959, 114 p.

Basic principles and problems are common to a number of information storage and retrieval systems available for use: various methods of selection from a store, linguistic problems, encoding of concepts, code system design, etc. One type of information system, the concept coordinate indexing technique, may be applied to hand- or machine-manipulated systems of varying degrees of complexity. Costs can be computed for selection, abstract production, and updating and correction of files; these costs can be compared for different types of concept coordinate systems. Analyses and comparisons are made of four Uniterm systems, an IBM punch card system, and two IBM computer systems.

I-24. Correlation of Indexing Headings and Title Words in Three Medical Indexing Systems, by John O'Connor, American Documentation 15, 96-104 (April 1964).

In a study of the Index Medicus wherein the heading-title synonym inclusion relation was analyzed (cf. II-36), results indicated that 86 percent of the heading-title pairs were related and the heading assignments could have been produced automatically by computer, with the aid of a thesaurus. The study raised several questions, one of which was what would be the results of a similar study of other indexing systems, especially in medical literature. Such a study, presented here, analyzed the Index Handbook of Cardiovascular Agents (Index), the Merck Sharp and Dohme punch card retrieval system (MSD), and the National Institutes of Health Research Grants Index (NIH) for the same synonym inclusion relations. A random sample of 50 papers was selected from the Index, and for each paper one of the headings assigned to it was randomly selected. Of these 50 heading-title pairs, 32 percent were related by synonym inclusion. Of the 50 random heading-title pairs from MSD, 54 percent were characterized by synonym inclusion. Of the 50 random title-heading

pairs (almost all selected by an every-tenth-page process) from NIH, 26 percent had synonym inclusion. The general conclusion is that at present any proposal to produce the indexing headings by computer should be viewed with caution.

I-25. Automatic Abstracting Evaluation Support, by Dan Payne and John F. Hale, Report No. RADC-TDR-64-30, Final Report to Rome Air Development Center, Griffis Air Force Base, N. Y., Rome Air Development Center, Information Processing Branch, February 1964, 38 p. plus appendices.

The objectives of this study were (1) to identify jobs and tasks in industry which are dependent on scientific documentation, (2) to compare the utility of abstracts with that of original text on each type of task within several scientific fields, and (3) to determine the feasibility of tailoring abstracts specific to each type of task. The American Institute for Research had earlier developed a technique for producing a good general-purpose informative abstract. This study evaluates the utility of abstracts generated by the technique; both the general-purpose and the task-specific abstracts were so evaluated.

Four general tasks were identified: screening, comprehension and retention, fact retrieval and problem solving. Two consultants each in the fields of chemistry, physics, biology, metallurgy, and psychology prepared test items based on 40 technical papers in trade journals of the respective fields. The papers were abstracted: full abstracts, abstracts tailored to the information requirements of each type of task, and a summary or brief abstract used as a control in the test program.

For the screening task, the test subjects (college students majoring in one of the fields or professionals from local industries) were given a problem in the form of a single test item and "available literature" in the form of the 40 articles or abstracts related to the problem. They were instructed to select nine references on the basis of perceived relevance to the problem. Accuracy was defined as the number of correct references selected (correct articles were those upon which the consultants had based the remaining task-test items), and performance time was recorded from start of search to completion of selection.

The comprehension-retention task required the subjects to read carefully three of the nine selected references and to answer nine questions on the basis of what they remembered from those references. They were later allowed to refer back to the three references in order to verify their answers. Thus there were two measures of accuracy, one for retention alone and one for retention and search ability. Total performance time was the sum of times for reading, answering, and correcting.

In fact retrieval, the subjects were given three other references from the original nine and nine questions answerable by simple concise facts. Accuracy was the number of items correctly answered, and performance time was the time spent looking up the answers.

For the problem-solving task, the subjects were given the three remaining references and three essay-type questions, and instructed to solve the problems using the references as source material. Accuracy was measured on a 0-9 scale, with each question graded on a 0-3 basis. Times recorded were total performance time and also the amount of time spent reading or referring to the three articles.

General conclusions include that the abstracts produced nearly equivalent accuracy, relative to full text, on screening and comprehension-retention across all fields; that full text produced somewhat superior performance on problem solving and markedly superior performance on fact retrieval; that there were no substantial differences in accuracy between full and tailored abstracts on any task; that use of abstracts resulted in highly substantial time savings relative to full text in performance of all tasks except fact retrieval; and that tailored abstracts result in a time savings relative to full abstracts on all tasks except fact retrieval.

I-26. Chemical Documents and Their Titles: Human Concept Indexing vs. KWIC-Machine Indexing, by Mary Jane Ruhl, American Documentation 15, 136-141 (April 1964).

As machine-produced title indexing becomes more widely used, authors and editors should be concerned with the composition of titles. Comparison of permuted title indexing of 84 titles in Chemical Titles and concept indexing of the same documents in the Chemical Abstracts Subject Index shows that 57 percent of the titles included all important concepts or equivalents, 17 percent contained all but one concept, 14 percent all but two, and 12 percent missed three or more concepts. On the whole, titles defined adequately the work reported in the documents; their composition should emphasize descriptiveness and retrievability.

I-27. Experience with Indexing and Retrieving by UDC and Uniterms, by J. A. Schüller, ASLIB Proceedings 12, 372-389 (November 1960).

At the Armed Services Technical Documentation and Information Center of The Netherlands' Ministry of Defense, the Universal Decimal Classification card system and the Uniterm system of coordinate indexing on peek-a-boo cards are used together for indexing and searching technical report collections. A few hundred questions from the register of queries received at the Center in the last few years were put to both systems. In the first test, 100 queries were used, and the number of relevant and irrelevant documents retrieved and the time for searching were recorded for each system. In addition, the number of documents found by both systems was noted. Time for searching was essentially the same. Both systems found 50 percent of the relevant documents, the UDC found 23 percent more, and the Uniterm system found 27 percent more. The close results support the conviction that the two systems should be used together. In the second test, 200 queries for specific reports known by title were put to both systems. Engineers consulted the UDC system at an average of 10 minutes per report and did not find 37 reports. Librarians used the peek-a-boo cards at an average of 1.5 minutes per report and did not find 2 reports. The conclusion drawn for the Center is that the Uniterm system is most useful, especially in its peek-a-boo form, as a complement to the Universal Decimal Classification system.

I-28. A Comparison of Dictionary Use within Two Information Retrieval Systems, by Glaire K. Schultz, Phyllis D. Schwartz, and Leon Steinberg, American Documentation 12, 247-253 (October 1961).

The Merck Sharp and Dohme (MSD) punched card system and the Armed Services Technical Information Agency (ASTIA, now Defense Documentation Center) computer system were compared as to the use of their dictionaries for building their input files. The average number of descriptors per document in the MSD system was 12. The average number in the ASTIA system was 5.4 at the time of the study; later provision for adding automatically certain generic terms should have increased this by three. Statistics were compiled about the frequency of occurrence of descriptors in single, double, and triple combination: in both systems the shape of the curves for the occurrence of single descriptors is one-half of a U. Curves for the use of single descriptors in terms of the average number of uses of descriptors coincide, and the authors hypothesize that the shape of this curve (nearly a straight line for cumulated frequency of use of descriptors) is an intrinsic characteristic of dictionary use.

I-29. A Comparison of Systems for Selectively Disseminating Information, by Ralph H. Sprague, Jr., Indiana Business Report No. 38, Bloomington, Indiana, Indiana University, Graduate School of Business, 1965, 70 p.

The primary purpose of this research project was to determine the best method by which the Aerospace Research Applications Center (ARAC) could select documents from the large store of documents abstracted and indexed by the National Aeronautics and Space Administration (NASA) and selectively distribute them to industrial scientists and engineers. Selective dissemination is not a new endeavor, but recent growing interest in mechanized selective dissemination has been prompted by the accelerated pace of development and proliferation of information, and the need to maximize the productive time of available manpower. As the size of a document store increases and a total search becomes more lengthy and expensive, selective dissemination may become a more important service than retrieval. The advantages of automatic selective dissemination should not blur the fact that it needs to be economically justified. The effective performance of selective dissemination rests on the quality and scope of the body of documents used as input, user acceptance and cooperation, and the actual processing system that performs the selective dissemination. A comparative test of systems operating on a centrally created body of documents should prove helpful to anyone planning to exploit such sources for selective dissemination.

The performance of a selective dissemination system can be measured in terms of two types of errors: documents relevant to a user's interest that are not selected (miss), and irrelevant documents that are selected (trash). The relative "cost" of each depends mainly on the user. A selection method can thus be evaluated by the extent to which it minimizes the cost functions of many users. This experiment provided empirical data that indicated which type of selection method performed the selective dissemination function best. The tested methods are models and the results of the experiment pertain directly to only these models; however they form the core of most methods currently in use. A more general contribution of the research results from the methodology, suggesting a straightforward technique for comparing several types of selection systems in many specific situations.

Four decision systems for document selection were comparatively evaluated in this experiment: (1) the Boolean method, selecting documents that satisfy certain conjunctive and disjunctive relationships between profile terms; (2) the simple term counting method (T-system), selecting documents in which a certain number of terms coexists in the index

and profile; (3) the percentage hit method (P-system), selecting the document when a certain proportion of the document index terms match those in the profile; and (4) the weighted term method (R-system), selecting documents when the sum of weights on matching terms exceeds a given level. All four of these selection logics depend on the keyword concept, the assumption that the intellectual contents of a document and the scope of an interest profile can be adequately characterized by a list of terms.

The research was conducted with the cooperation of ARAC, formed under contract with NASA to explore the feasibility of expanding nonmilitary use of techniques and processes developed in space-oriented research and developments efforts. ARAC acts as intermediary between industrial firms and various sources of information, centered around the indexed information available through NASA. The store of information involved in the research project is contained in two abstract journals, Scientific and Technical Aerospace Reports (STAR) and International Aerospace Abstracts (IAA). An ARAC staff engineer was assigned to each interest center (a point in an organization's structure at which decisions requiring information were made) to design a profile of needs for machine search and edit the machine output for relevance. He was the interpretive liaison between the interest center coordinator and the information store. These staff engineers were the "users" participating in the experiment.

The three major assumptions underlying the research methodology employed are: (1) the meaning and role of relevance, including the simple distinction of relevant or nonrelevant; (2) the independence of relevance decisions, each document being judged on its own merit irrespective of other documents received (these two assumptions combined presume that all nonrelevant documents are equally nonrelevant and all relevant documents are equally relevant, and in terms of error measurement in system evaluation one missed document is as serious or costly as another); and (3) the adequacy of the abstract as a basis for the relevance decision.

In the experiment construction of the profiles was the most critical step and it was necessary to establish a "typically good" or "average operational" profile for each system. The most important element in profile construction was the selection of the index terms that made up the profiles. The set of index terms was the same for all system profiles for a given interest center. Any significant performance of the systems resulted from the manner in which these terms were combined and interrelated in the search logic. Since the set of terms was common, the user first chose the terms that adequately described his interests; the profiles for methods 2 and 3 were then complete. To finish the profiles for methods 1 and 4, the user stated the conjunctive and disjunctive relationships between terms and assigned a weight to each term that indicated the relative importance of that term to his interests. To validate the choice of terms, the analysts performed a manual search of the abstracts in an issue of STAR and IAA and indicated which were relevant. The indexes of these documents were examined to insure that no important words were omitted from the profile. The amount of time necessary for construction of each of the profiles was observed. In general, it took the analysts a little longer to establish the Boolean relationships (from five to ten minutes) between profile terms than it did to weight them (from two to five minutes).

To convert the inverted file supplied with NASA's search system to a linear file to facilitate matching terms in a profile with terms of a document, each term number from the inverted file was paired with the number of each document it indexed. The resultant document-term pairs were blocked and sorted; grouping all the terms associated with a unique document number in a single tape record completed the conversion. This linear file tape was used by the term matching systems while the inverted file and ARAC's search system were used to perform the Boolean search. To facilitate the development of performance data over the full range of selection criteria, the status of each document with

respect to the profile was indicated for each system. It was possible to tabulate the number of documents that would have been selected under each possible selection criterion for each system.

In order to evaluate the effectiveness of the systems, it was necessary to determine which documents in the total score were relevant to each interest center. All relevant documents were assumed to be included in the set of documents indexed by at least one of the terms. The result of a very broad machine search was submitted to the analysts, whose task was to identify the documents relevant to a given interest center. It was then possible to tabulate the hits, misses, and trash for each system at all desired selection criteria. The experiment involved 25 profiles; 19 of these were under test for the first five issues of IAA and STAR, containing 7, 110 documents; six were under test for the seventh issue, containing 2, 015 documents.

A graph relating the relevance ratio (fraction of documents retrieved that are relevant) and the recall ratio (fraction of all relevant documents retrieved) shows the performance of the four systems for each interest center. Twelve hypothetical searches were performed by each of the three term matching systems; each search yielded a single relevance ratio and recall ratio that was plotted on the graph as a single point. The 12 points comprise the curve for the system; each system is represented by a distinctive line. The dimension lacking from the curve is the differing requirements of users regarding miss and trash. The system that performs at the lowest cost for a given function relating these two types of errors is the best system in that instance. The C value, which is called a cost here, is not a cost in the usual sense of the word, but indicates the detriment suffered by an individual user. A cost function k = 1 represents the user who cares little about missing a document but does not want to receive much trash; when k = 3, the user is rather concerned about not missing good documents and is willing to spend some time and effort eliminating the trash; if k = 5, there is strong concern over a missed document and a willingness to cull quite a bit of trash.

For a user whose cost is specified by k = 1, the Boolean system has the lowest cost, the T-system and P-system are tied, and the R-system is fourth. Where k = 3, the rank from first to fourth is T, R, B, P; the Boolean drops to fourth place when misses are costly (k = 5) and the ranking becomes T, R, P, B. Three important facts are revealed by the data analysis: (1) the weighted term system performed best for all cost functions used in the ranking; (2) the P-system performed the worst for all cost functions; and (3) the relative performance level of the Boolean system decreased as the cost of missed documents increased. It is significant that the weighted term system showed the best over-all performance. The main characteristic of the system, not possessed by any of the other systems, is that it allowed the profiler to indicate the importance of the profile terms to his interests. This is a valuable search system characteristic. Another important implication resulting from the findings is that the Boolean system was hindered by its inability to adjust to various user requirements without additional profiling effort; such flexibility is also a valuable system characteristic.

I-30. A Comparative Study of Three Systems of Information Retrieval: A Summary, by Norman D. Stevens, American Documentation 12, 243-246 (October 1961).

(This is a summary of the report on a research project undertaken in the Graduate School of Library Service of Rutgers, The State University of New Jersey, and published by the Rutgers University Press in 1960. The full report was reviewed by Cyril Cleverdon in American Documentation 14, January 1963; by C. W. Hanson in Journal of Documentation 18, September 1962; and by P. A. Zaphyr in Computing Reviews 3, July-August 1962.)

Three systems dealing with the literature on explosives, namely a punched card file, a handbook reproduced from that file, and a more conventional library cataloging-reference approach, were compared as to input costs, degree of use and of user satisfaction, and output performance. The input costs and costs per use of the punched card and handbook systems were much greater ($250 per report and $100 per use respectively) than those of the conventional library catalog ($0.70 per report and $1.42 per use respectively). The card files (11 sets were distributed) have been used a total of only 32 times in seven years. The handbooks (75 copies were distributed) also were used infrequently; the fifteen most often used showed, for example, a total of just eight uses in two months. The reasons suggested for such little use include lack of equipment to handle the cards, difficulties in consulting the handbook, and lack of critical evaluation of the data presented. To examine output costs, time, and comparability of answers, 58 questions which had actually been asked by people using either the punched cards or the handbook were checked through all three systems. Manual search of the handbook was in almost every instance more efficient than machine search of the card file. For simple matters of fact the reference approach was superior. In general, a data-extracting system may be feasible if there is a mass of data available on a subject in a large number of different sources, there is heavy demand for the information, and there are no adequate conventional reference tools in the area.

I-31. Mathematical Analysis of Various Superimposed Coding Methods, by Simon Stiassny, American Documentation 11, 155-169 (April 1960).

The method of coding several key words into one and the same field on a punched or edge-notched card is known as superimposed coding. The main advantage of the method is that it saves considerable space; its main disadvantage is that by superimposing several patterns new ones are created, some of which may belong to code words. These spurious words may cause a false match. It is the purpose of this analysis to estimate the probability of a match being false and to indicate how this probability can be minimized.

In practically all superimposed coding schemes the patterns of the words are chosen in such a fashion that all positions in the field have about an equal probability of being punched. It will be assumed therefore that the punches are placed in the field at random, and that the field is a single unit with no division into subfields.

The chain spelling method was designed for direct coding of alpha-numerical words on punched cards. A rectangular or square field of R rows and C columns is chosen on the card. The assignment of headings to rows and to columns is done in such a fashion that all rows and all columns have as equal a probability of being punched as possible. Suppose now that a number w of words is to be punched into a square of size R x C by abbreviating (or extending) each word to m letters and then applying the chain method (i.e., spelling the word in groups of two letters, each letter connected to the one following it and an end-around punch connecting the last letter to the first). A procedure to estimate the probability of a match being false under these circumstances is derived as follows: Let $Z_e(w)$ denote the total number of patterns created in the field by punching w words. The value of $Z_e(w)$ depends on the number of holes in the field, a random variable. We shall compute the expected number of holes and use this number for the subsequent estimation. A good approximation to G_e (the expected number of holes) is $G_e = RC - RC_e - wm/RC$. If the vocabulary size is V, then the probability that a pattern belongs to the vocabulary is V/L^m (where L = number of characters). Therefore an estimate of the number of vocabulary words created in the field by punching w words is $[Z_e(w) - w]\, V/L^m + w$. In all practical situations V/L^m is a negligible fraction. Therefore $Y_e(w) = Z_e(w)\, V/L^m$ is an estimate of the number of vocabulary words created in the field in addition to the w original ones.

Substituting for $Z_e(w)$ and thereafter for $G_e(w)$ we get the probability of a match being false under these circumstances as $P_e(w) = Y_e(w)/[Y_e(w) + w]$.

To estimate Z(w) for the chain method without the end-around punch, we follow the same arguments as in the case above and find that an estimate for the number of additional vocabulary words created in the field is $Y(w) = Z(w)\ V/L^m$. The probability of a match being false is $P(w) = Y(w)/[w + Y(w)]$.

The random number method of coding differs from the chain method only because of the absence of chaining. Using the same line of reasoning as in the chain method, we find that an estimate of the number of additional words created in the field is $Y^*(w) = G^m(w)\ V/10^{2m}$ (where 10 headings are distributed over the rows R and also over the columns C). The probability of a match being false is, as before, $P^*(w) = Y^*(w)/[w + Y^*(w)] = P_e(w)$.

It can be concluded from the results presented thus far that for a given field, the number of additional words created in the field by punching w words (and the probability of a match being false) depend only on the number of m of punches per word, regardless of which of the coding methods is used. In other words, there is no difference between the chain and random number methods nor between any of the variants described.

Letting the field size RC and the number of words punched w be fixed, consider the problem of finding that m which minimizes P_e or, equivalently, Y_e. The optimum number of punches per word does not depend on the vocabulary size. It is sufficient to minimize the function $g(m) = \log_e \left(1-e^{wm/RC}\right) m$. The result is $m_{opt} = (RC \log_e 2)/w = 0.6931\,RC/w$. All of the results thus far presented depend on RC and not on either R or C individually. This means that field shape is immaterial and only field size matters.

The staggered superimposed coding method was designed for the purpose of using the standard IBM code in superimposed coding. The words to be coded are punched into the card. The starting column of each word is determined by the first letters of the word. The columns are assigned in such a way that the groups of words with the corresponding starting letters have about an equal frequency of occurrence in English. The purpose of this "staggering" is to achieve as random a distribution of the punches over the card as possible. Let us estimate the probability of a match being false: On the basis of the number w of words punched, the expected number of holes in the two fields is calculated. Using these expected values, the number of additional words created in the field is calculated, and this is used to estimate the probability of a match being false. An estimate of the number of m letter additional vocabulary words created in the field is $Q(m) = (c-m+1)\left(\frac{G_z}{c} \cdot \frac{G_n}{c}\right)^{m-1} p(m)\ V/26^m$: an estimate of the total number of additional vocabulary words created in the field is $\sum_{m=1}^{M} Q(m)$, where M = maximum word size. Finally, the probability of a match being false is $P_{fm}(w) = \sum_{m=1}^{M} Q(m) \Big/ \left(w + \sum_{m=1}^{M} Q(m)\right)$

Until now, all estimated probabilities referred to the event of a single match being false. In general, inquiries consist of more than one word. It is therefore interesting to know the probabilities that among given number x of matches found there are at least d true ones. For some fixed x, let D be the number of true matches among those x matches, $0 \leq D \geq x$. Then D is binomial random variable, with parameters $(x, 1-P_e)$, where $(1-P_e)$ = the probability of a single match being true. A table can be constructed giving the probabilities $B_d(x, 1-P_e)$ that among x matches at least d are true and, for various values of the probability $(1-P_e)$, that a single match is true.

Superimposed coding can be a useful tool provided that it is carefully planned and optimized so that the number of false answers is kept at a predetermined minimum. Vocabulary size must be taken into consideration. The field size necessary to keep the probability of a match being false at a certain level and the optimum number of punches per word can be computed.

I-32. Evaluation of Information Systems for Report Utilization, by Mortimer Taube, in Studies in Coordinate Indexing, Vol. I, Washington, D.C., Documentation, Inc., 1953, p. 96-110.

Evaluation of information systems must be based upon internal criteria, characteristics of the systems themselves, and external criteria, facts of consumer acceptance or satisfaction. These two are interdependent and should be used together. Since the criterion of internal characteristics is the only one available it should be used as a measure of consumer satisfaction and an instrument of evaluation, until methods for testing the adequacy of any system in terms of consumer satisfaction can be developed. The elements or characteristics which can be included in a set of criteria (all the items are assumed to be independent) are: cost; size, of files and of cataloging apparatus; time, including that required to organize the system and that to search it for information; number of access points per item; rate of obsolescence, or rate of change needed to keep a system current; rate of growth; specificity; universality, or hospitality to ideas in different fidlds; logical structure; suggestiveness; neutrality, a sub-property of universality, but which would permit any item to be a member of any and all classes in the system and would render the system appropriate for both manual and machine handling; simplicity, both in structure and use; suitability for cumulative dissemination, as a card file reproduced in book form; and familiarity. All these elements have a relation to consumer satisfaction but we have no measures of their relative importance.

The Uniterm system, several classification systems, several subject heading systems, and several standard methods of indexing were compared in terms of these characteristics; the Uniterm system was judged superior in all respects except suggestiveness.

I-33. The Efficiency of Subject Catalogues and the Cost of Information Searches, by R. G. Thorne, Journal of Documentation 11, 130-148 (September 1955).

An expression for the efficiency of a subject catalogue or index is derived from the probability of success when using the catalogue, and the cost of making and using the catalogue compared with the cost of finding material when no catalogue is available. If no records were kept the total annual cost would be SCQ, where C = total number of documents in collection, Q = number of searches made in a year, and S = average cost per document of hand sorting, found by determining the time to search for typical questions through a reasonably large random sample of documents in the collection. If records are kept, processing costs include reading the documents to find their semantic content, converting the semantic elements into a vocabulary or code, and posting the code into a storage device. Then the annual cost of making and using the subject catalogue is PN +RQ, where N = number of documents received in a year, P = processing cost for one document, and R = operating cost for one search. The reduction in costs obtained by using a subject catalogue is thus SCQ-PN-RQ.

One hundred typical test questions based on material known to be in the collection can be used to assess the catalogue's probability of success. If A = total number of answers

which include reference to the original documents on which the questions were framed, then 100-A = questions in which the catalogue failed. Finding correct answers to these questions costs SC(100-A) and to all 100 questions costs 100 R +SC(100-A). The annual cost of finding correct answers = RQ +SCQ $\frac{(100-A)}{100}$ and the total annual cost including processing costs = PN +RQ +SCQ $\frac{(100-A)}{100}$. The efficiency of a recording system is defined as the cost saved by using the system divided by the cost if no system is used; i.e., $\eta = \frac{SCQ-PN-RQ-SCQ\frac{(100-A)}{100}}{SCQ}$ or $\eta = \frac{A}{100} - \frac{PN+RQ}{SCQ}$. With conventional cataloguing systems the efficiency of the system depends mainly on the percentage of successful retrieval; with sophisticated systems using mechanical or electronic devices the values of P and/or R will increase and may be a significant item.

This efficiency measure was applied to the Catalogue of Aerodynamic Measurements at National Aeronautical Research Institute, Amsterdam; Uniterm systems tested at Cranfield (cf. I-2) and at the Royal Aircraft Establishment (RAE); and the RAE Main Library's subject catalogue classified by the UDC. In test of the Uniterm systems, 40 test questions were prepared based on material known to be covered by the indexes; successful retrieval of the original paper was achieved for 34 and 33 of the questions respectively, and A is assumed to be 83 percent. The same questions were put to the RAE catalogue and the selected paper was retrieved for 20 of the 40 questions, so A = 50 percent. Subsequently 100 questions were prepared, with care in avoiding words in the titles of the papers selected; in the Uniterm systems only 14 percent successful retrieval was obtained. For the Netherlands catalog, 100 questions framed by typical users were based on papers known to be covered by the catalogue; successful retrieval was obtained for 88 percent of the questions. Some 64 of the papers used in this test were included in the UDC catalog in the RAE Main Library and were used in a test of this catalogue; for 80 percent of the questions a card for the report was found under the UDC numbers chosen for search. Other tests of the Netherlands system showed A = 85 percent and A = 60 percent.

The size (C) and growth (N) of the collections were determined, the number of searches per year (Q) were derived, the costs of hand-sorting (S) were suggested and the costs of processing (P) and operating (R) were calculated. From the various estimates of factors influencing search costs and efficiency the efficiency and cost per search of the various systems were calculated and plotted. In all cases the cost of the one search if no catalogue were available were also shown.

SECTION TWO

DESCRIPTIVE EVALUATION

The reports in this section deal with the testing of single systems, rather than the comparison of two or more systems. These testing and evaluative studies provide a description of the operating performance of the system.

As was the case in the previous section, reports included here deal either with study of total systems or with subsystems and components. The greatest interest has been focused on the indexing step, including the techniques and vocabulary of the indexing process. To reflect this predominant interest, the abstracts of reports dealing with indexes, indexing and indexers have been collected at the beginning of this section, items II-1 through II-16, in alphabetic order of author names. All of these reports may be roughly classified into those studying the effect of index language devices on system performance and those studying the problems of consistency in the indexing process.

The former group includes a report (II-1) on the latest phase of the ASLIB-Cranfield project, the study of indexing devices such as links and roles. This is an outgrowth of the project reported in Section I-1, which demonstrated the effect of the type of index language on the performance of systems being compared. Another study (II-11), undertaken at the Dupont Company, compares performance of three systems which differ in their use of links and roles, and is therefore included in this section rather than as a comparative evaluation in Section I. On the other hand, the reports (II-24 and -31) on evaluation of Patent Office files and the one (II-30) on the Bureau of Ships system contain data on the indexing phase, of interest in this grouping.

The question of the extent of consistency among different indexers indexing the same documents, or of one indexer re-indexing the same documents at different times, has great import in system design and evaluation. Quality of indexing depends on the level of reliability of indexers, which in turn depends on subject matter knowledge, training, and the use of aids such as term lists and lookup tools. Controlled testing of quality and reliability, both inter-indexer and intra-indexer, attempts to throw light on this problem.

Most of the descriptive evaluations included in the rest of the section (II-17 through II-56) deal with actual operating systems, and of these most are studying overall performance or effectiveness of the system. Only four reports cover experimental systems, three on the SMART system and one on ROUT, and these also measure performance. Four reports on actual systems detail costs of their operations (II-33, -40, -41, and -42).

The next biggest category of reports, and an area of growing interest, is concerned with the techniques of automatic indexing or classification, that is, analysis of text and selection of index terms by machine processes. These techniques are still experimental but their products, the output of their processing, are being subjected to critical study and evaluation. The techniques employed in such evaluations are of broader interest and application.

II-1. The Testing of Index Language Devices, by Cyril W. Cleverdon and Jack Mills, ASLIB Proceedings 15, 106-130 (April 1963).

(This work also exists as an ASLIB-Cranfield Research Project report. In addition, "The Analysis of Index Language Devices", by Cyril W. Cleverdon and Jack Mills, in Automation and Scientific Communication, Proceedings, Part 3 (papers presented at the 26th Annual Meeting of the American Documentation Institute, Chicago, Ill., October 1963), Ed. by P. C. Janaske, Washington, D.C., American Documentation Institute, 1963, p. 451-454, discusses generally the same subject matter as this paper. Also, "Testing Indexes and Index Language Devices: The ASLIB-Cranfield Project", by F. Wilfred Lancaster and Jack Mills, in American Documentation 15, 4-13 (January 1964), reviews the total project including the current work discussed in this paper.)

Current work on the ASLIB-Cranfield project attempts to measure the effect that particular indexing devices have on recall and relevance; it seeks also to use more precise measures of relevance. The project has demonstrated that there is an inverse relation between the two factors of recall and relevance, and that exhaustivity of indexing and specificity of the index language are the broad parameters that affect the performance potential of any retrieval system. The present work is concerned with index language devices, seeking to measure the impact on recall and relevance made by particular syntactical devices. Refinements in test methodology based on experience gained in the previous studies are being applied. A closed collection of 1,500 documents, indexed with great exhaustivity, is being used; 400 questions have been supplied for the test. Every document has been examined in relation to every question, and a relevance value within a five point range is given to each document for each question. Each index language device, such as multiple hierarchical linkages, coordination of terms, roles, links, etc., will be introduced into an uncontrolled vocabulary of single terms and will be tested in turn. Each device will have a certain impact, either to broaden class definition and so improve recall (but lower relevance) or narrow class definition and thereby improve relevance (but reduce recall). This impact will be measured; knowledge gained in this test should permit design of systems with maximum required performance characteristics.

II-2. An Evaluation of Links and Roles as Retrieval Tools, by Stanley M. Cohen, Carol M. Lauer and Bettina C. Schwartz, Journal of Chemical Documentation 5, 118-121 (May 1965).

The indexing and retrieval system described here uses the concepts of coordinate indexing; indexing is done from an authority list (glossary) to an average depth of 31 terms per document and an average number of three links per document. Term codes, role codes, link designations and document numbers are entered into standard IBM tabulating cards. Concept coordination is accomplished on a collator by matching document numbers after which an optical coincidence scheme is used to obtain matches at the link level. Two-digit role indicators are used, the main role expressed by the first digit and the subrole by the second digit. Both main role and subrole codes are assigned for all term entries. In indexing properties of materials, the indexer assigns the same main role code to both the property and material terms. Similarly, a term used adjectivally is made to agree in main role with the term it modifies. When the role agreement technique is used for retrieval, the main role is considered part of the document number, and cards are sorted and collated on the basis of this enlarged document number. The role agreement technique groups terms within a link and might be called sublinking.

To measure the value of links and roles, searches requested by members of the professional staff in a particular time period were subjected to a number of different search methods. These were: (1) links and roles not utilized, (2) links utilized, (3) links and main roles utilized, (4) links and subroles utilized, (5) links, main roles and subroles utilized, (6) links and role agreement utilized, and (7) links, role agreement and subroles

utilized. Each question was subjected to as many of the methods as applicable. The number of hits and processing times were recorded and the documents selected were examined for relevance. The number of relevant documents was recorded for each method used in a question, and the following parameters calculated: relevance, relative relevance (ratio of relevance to that of Method 1), relative recall (ratio of relevant documents found to those found by Method 1), productivity (relevant answers/minute), and productivity ratio (ratio of productivity to that of Method 1).

The relevance values obtained by use of the more sophisticated methods were substantially higher than those obtained in Method 1. Search methods which involved the maximum utilization of links and roles (Methods 5 and 7) resulted in the highest relevances. In other words, relevance increased as the use of links and roles increased. The introduction of search methods which use links and roles tends to decrease relative recall. Links alone have only a small effect on this decrease, but some answers were lost as a result of using roles. This effect is greatest in Method 7, where about 20 percent of the relevant answers found in Method 1 were lost. It appears that the use of roles results in a somewhat higher level of indexing inconsistency. The average loss of answers in the methods involving roles is about 10 percent and is primarily the result of these indexing inconsistencies. For average relative relevance, Method 6 had the lowest of the six methods involving links and/or roles, but still had relevances which averaged 1.56 times those of Method 1. The relevances of Method 5 averaged almost four times the relevances of Method 1 for the same question. The productivity for individual questions varies widely, and therefore comparison of the ratios of productivities is more meaningful than comparison of the actual productivities. The main factor affecting the relative productivities of search methods with and without the use of links and roles is the extent to which the size of the original card decks are reduced as a result of division by role. If machine time saved by use of smaller decks is sufficient to overcome that of the extra steps required, productivity is increased. The productivities of Method 5 averaged 1.78 times the productivities of Method 1 for the same questions.

In the system described, for any given search a number of options are available with respect to the use of non-use of links and/or roles. This flexibility of methods adds significantly to the intrinsic value of links and roles as retrieval tools.

II-3. Role Indicators and Their Use in Information Searching-Relationship of ASM and EJC Systems, by Marjorie R. Hyslop, in Parameters of Information Science (Proceedings of the American Documentation Institute, 1964 Annual Meeting, Vol. 1, October 5-8, 1964, Philadelphia, Pa.), Washington, D.C., Spartan Books and London, Cleaver-Hume Press, 1964, p. 99-107.

One of the most popular systems for information retrieval is based on use of a thesaurus for vocabulary control, links to show relationships of terms, and role indicators to show the function played by a particular term in the context of the document. The thesaurus is sometimes referred to as the indexing language. Links and roles are indexing devices, designed to refine, clarify and add precision to the indexing language. The present paper is restricted to a study of role indicators in the ASM and EJC systems and the actual use of the former in searching.

In both systems each role indicator is always used in conjunction with a particular indexing term; they are designed to provide rudimentary grammar Compatibility in the two sets has been improved and the roles can be placed in two groups --- those that are nearly identical or convertible and those for which there is no counterpart or where only a slight relationship can be established. Both systems require that each indexing term be accompanied by a role indicator. In determining how role indicators have been used in actual searching situations, the computer programs for 150 ASM searches performed

operationally in response to subscriber requests were analyzed. All were retrospective searches and had to be programmed for maximum effectiveness in one try. Fifty searches received in order, regardless of subject, were analyzed from each of three time periods: during the early period of ASM service, from the middle period, and representing the most recent experience. The analysis consisted in tabulating the total number of role indicators used in each computer program and the number of times each individual role indicator appeared in a computer program. Of the total 150 searches, 82 (54.7%) were programmed without any use of indicators. When roles were used, only 4 percent of the searches used just one and only about 5 percent used more than five. Figures showed a remarkably heavy use of certain role indicators and very slight use of others. Concentration on the "popular" role indicators is most marked in the recent period, with widest distribution of usage in the early period.

Certain facts can be summarized from the tallies: role indicators were used in roughly half of the search subjects requested, four role indicators did the lion's share of the work, a few were used so rarely as to question their necessity, and the total use of role indicators was somewhat greater in the later periods. A few explanations and conjectures can be offered: inconsistency in indexing militates against the use of role indicators in search programs, as programmers were hesitant to specify role indicators which carry precise shades of meaning; the tendency toward greater concentration on the popular role indicators might reflect some unfortunate experiences in attempting to formulate precise programs using a larger variety of roles; use of certain roles is unnecessary because their meaning is already built into the semantic code used as the indexing language; additional shades of meaning are built in by the process of infixing; and finally, the use of punctuation or links often ties together two concepts so closely and neatly that role indicators are rendered unnecessary. The ASM system carries a large amount of built-in redundancy among the functions performed by the vocabulary, roles, and links. Skillful manipulation of this redundancy by the programmer should insure precise search results if indexing has been properly and consistently performed.

Based on the analysis of 150 searches, the value of role indicators as an indexing device can be questioned. An obvious assumption is that a much simpler method of showing functional relationships would serve the purpose equally well. It would not be fair to extrapolate these findings to the use of role indicators in other systems which may not carry the redundancies which appear in the ASM system. The possibility that the ASM system could be improved by simplifying the role indicator complex and placing more emphasis on use of links deserves study. Finally, the lack of means of emphasizing the important topics in a document has proven to be a major defect leading to a large amount of irrelevance in searches on broad subjects. A new role indicator has been introduced into the ASM system, matching a role in the EJC list, which does not denote a function but is attached to any term which represents a major aspect of the document and can be used in combination with any other role indicator. It is anticipated that introduction of this new indicator will be of benefit to search results.

II-4. Indexer Consistency Under Minimal Conditions, by J. Jacoby and Vladimir Slamecka, Report to Rome Air Development Center, Contract AF 30 (602)-2616, Bethesda, Md., Documentation Inc., November 1962, 42 p., plus appendices.

The objective of this work is to improve the quality of indexing by establishing and then improving, if necessary, the level of reliability of indexers. Minimal consistency, i.e., agreement attained by indexers working without look-up aids, without mutual consultation of indexers, and without post-indexing editing, was determined for the Uniterm system of coordinate indexing. Two groups of indexers were used in the experiment: three experienced in Uniterm indexing and three indexer-beginners. The

inexperienced indexers used the Documentation Inc., Code Manual for chemical patents to guide their work. Since the experienced indexers also used the Code Manual by memory . the two groups were considered to be on a comparable level.

Inter-indexing reliability (consistency among indexers) and intra-indexer reliability (consistency by one indexer in re-indexing the same documents) were tested. A random sample of 75 chemical patents issued in 1962 constituted the collection of documents for the inter-indexer test. All indexers were required to index the title and claims of each patent; they were to index at their own judgment the specifications and examples in each. The number of terms used by each indexer to index the patents and the percent of terms matched by the experienced and inexperienced indexers were counted. For the intra-indexer test, the effects of memory or recall in re-indexing the same documents had to be controlled. Three batches, of 75 patents each, of "equated" documents were selected. (Patents were considered equated if they were classified in the same class and subclass of the Patent Office Classification and had enough similarity so that the mental processes for indexing were similar or substantially equivalent in each case.) A "general term" vocabulary was derived for each group of patents from a given subclass (a vocabulary common to all the patents within that group) and used as the basis for measuring intra-indexer consistency. Again, the number of terms used by each indexer, and the percent of terms used by each indexer in each batch, were counted.

In the inter-indexer test a significant difference was found in the number of terms assigned to the patents by individual indexers. The experienced indexers showed less variation than the inexperienced ones, and the defined indexing area (title and claims of a patent) exhibited more stability than the undefined one. As a group, the experienced indexers showed significantly higher degree of inter-indexer consistency than the inexperienced indexers; however, in general there is a large amount of individual variation and hence lack of inter-indexer consistency. In the intra-indexer test, all the indexers showed on the average no difference in the number of general terms used among the equated batches, and hence were consistent with themselves in this respect. The indexers tended to be more consistent in re-indexing an equated document in the defined area to be indexed, if they were inexperienced; but in the undefined section, if they were experienced. In either section, the inexperienced indexers showed more variability than the experienced ones.

Further work will be directed to investigating the possibilities of improving indexer reliability through the use of various indexing aids (cf. II-16).

II-5. The Modification of an Information Retrieval System by Improving Vocabulary Control, Indexing Consistency and Search Capabilities, by E. A. Janning, University of Dayton Research Institute, Report No. AFML-TR-65-20, Wright-Patterson Air Force Base, Ohio, Air Force Materials Laboratory, March 1965, 8 p. plus appendices.

The document retrieval system established by the University of Dayton Research Institute to handle an extensive collection of scientific and technical reports that pertain to materials research is a coordinate indexing system utilizing an NCR 304 computer for the processing of data. The use of deep indexing with links and roles formulated the basis of the system. The role indicators used were a slightly modified version of those used by the DuPont Company. The system's vocabulary was generated through the indexing process and very specific terms were used in naming materials. The indexing of about 10, 000 documents resulted in a thesaurus of over 18, 000 multidisciplinary terms. The use of roles in indexing, and vocabulary control, were two aspects of the system evaluated and changed; they are discussed in this report.

The primary purpose of using role indicators is to decrease the number of irrelevant documents retrieved in response to a search. Experience here showed that their use also prevented retrieval of relevant information unless all possible roles were used in every search and when this was done, the retrieval of irrelevant information increased. The primary problem encountered was the inability to assign unambiguous definitions to the role indicators because of the multidisciplinary nature of the material contained in the system. An evaluation of links and roles used in the retrieval of information was made as a thesis study (cf. II-15).

Steps had to be taken toward vocabulary control in order to maintain a manual searching capability and also a reasonable vocabulary for computer searches. It was obvious that the area requiring control was the naming of materials and trade names. The vocabulary was divided into classes of materials for analysis, to determine the possibility of generalizing the naming of materials without losing an undue amount of specificity. Each analysis involved consideration of compatibility of terminology between areas, effect on indexing and indexer qualifications, effect of terminology on searching, and the amount of specificity that could be retained while the vocabulary reached a semi-permanent plateau. The areas of organic compounds and metallurgical terms were the first analyzed. A special fragmentation system, employing essentially unit terms, was developed to handle the naming of organic compounds. The system uses connectors to show relationship between basic structure and substituent groups. The metallurgical terminology, including materials, processes, properties and products, was generalized to some extent without causing any measurable increase in false retrievals. In solving problems in other areas conferences were held with experts in the areas to determine choice and use of terminology in those areas. The efforts on vocabulary control resulted in reduction from 48, 000 terms to about 10, 000 terms without affecting search capabilities; rate of addition of new terms has been minimized.

II-6. A Method for Computing Indexer Consistency, by Arthur L. Korotkin and Lawrence H. Oliver, Bethesda, Md., General Electric Co., February 1, 1964, 8 p.

In studies of the human indexing process the task of objectively measuring indexer consistency presents a problem. The few techniques which have been presented do not seem to represent accurately the amount of agreement among indexers, i.e., inter-indexer consistency. It was concluded that consistency can be expressed in terms of the ratio of the number of pairs of indexers using the same descriptor for a given article to the total number of possible pairs, that is, the number of pairs which exist if there is perfect agreement. Other formulas require that some high level of consistency be reached before the formula can be applied. All the formulas agree at the extremes of consistency; with the formula presented here moderate degrees of consistency can also be measured.

II-7. The Effect of Subject Matter Familiarity and the Use of an Indexing Aid upon Inter-Indexer Consistency, by Arthur L. Korotkin and Lawrence H. Oliver, Bethesda, Md., General Electric Co., Information Systems Operation, February 14, 1964, 17 p.

Since the purpose of an information indexing system is to provide for retrieval, the efficiency with which information can be retrieved will depend on the techniques used in indexing. This study was designed to provide information on two parameters relevant to the indexing process, subject matter familiarity and the effect of a "job aid" or descriptor reference list, and to test the hypotheses that such familiarity and job aid use increase inter-indexer consistency. A total of 10 subjects, divided into two groups of five each,

indexed 30 abstracts from the American Psychological Association's Psychological Abstracts. One group consisted of psychologists, the other of non-psychologists. The task was to assign three descriptors to each abstract. After two weeks elapsed, both groups were given the same 30 abstracts and again asked to assign three descriptors to each, with the help of an alphabetized list of "suggested" descriptors.

The measure of consistency used was devised for the specific treatment of these data (cf. II-6). Consistency was expressed in terms of the number of pairs of indexers using a given descriptor out of the total number of possible pairs (here, 30 pairs per article). Results indicate that differences between the two groups were not significant (subject matter familiarity did not increase consistency) but that differences between the two tests were significant (the "job aid" did increase consistency).

II-8. Consistency Analysis of Two Indexers in Using K.C. for Political Science Material, by Barbara Kyle, London, England, National Book League, May 1962, 4 p.

The two indexers, both familiar with the Kyle Classification (a faceted system), indexed and classified by that system 246 titles from the International Political Science Bibliography. Seventy percent consistency was achieved. The cause for the 30 percent inconsistencies was analyzed, and 12 percent were found to be due to lack of attention to the system's rules. But 18 percent were due to difficulties in use of the system and the colon notations; e.g., in some cases the same symbols were used but in different order, in other cases different symbols were used. In still other cases different subject matter was indexed. The difficulties with use of the colon notations point to the need for revision and expansion of the system's rules.

II-9. Some Observations on the Performance of EJC Role Indicators in a Mechanized Retrieval System, by F. Wilfred Lancaster, Special Libraries 55, 696-701 (December 1964).

The majority of retrieval systems currently being established, particularly those employing some form of mechanical searching, are post-coordinate systems. To prevent false retrievals in such systems, the devices of links and role indicators have been introduced. An indexer "partitions" a document into its distinct themes, expresses each theme by means of a combination of terms, and links the terms together by assigning to them a common letter or numeral. Role indicators show the functional relationship between index terms, i.e., they add an element of syntax.

Project SHARP is an automated information storage and retrieval system at the Bureau of Ships Technical Library (cf. II-30). Documents input to SHARP are indexed by descriptors organized, where necessary, by means of links. Relationships between descriptors are defined by means of Engineers Joint Council (EJC) role indicators. The system was evaluated by test techniques of the type developed by Cleverdon (cf. I-2). The purpose of this paper is to consider the problems inherent in the SHARP indexing system, specifically some of the problems caused by the application of role indicators. It was not possible to obtain data on the effect of the role indicators in improving relevance ratios of searches, but it was possible to produce observations of their effect on the recall performance of the system.

Of the 12 source documents that were completely missed in the 50 searches, three (25 percent) were missed because of problems in the use of role indicators. Altogether nine of the 46 known relevant documents missed in searching, i.e., almost 20 percent, were lost because of role indicator problems. Of the total failures that could in any way be attributed to the system, approximately 70 percent were traced to the role indicators. In many cases, relevant documents should have been retrieved, with very little noise, by much more precise subsearches in which the descriptors selected by the searcher had been used in indexing the relevant documents, but the role indicators used by indexer and searcher did not match.

With most of the losses caused by role indicators, it was difficult to lay the blame clearly at the door of either indexer or searcher. Documents were missed because the indexer's interpretation of the relationship between concepts did not always agree precisely with the relationship demanded by the questioner. There are undoubtedly ways in which these problems can be minimized. In an ideal situation the same group of people should both index the documents and formulate the search programs. Under these conditions one could reasonably expect a greater coincidence between indexer and searcher in choice of roles. Using role indicators on a highly selective basis may prove the best solution. Only roles clearly mutually exclusive would be allowed, and these only in situations where their discriminating power was clearcut.

Role indicators are precision promoters. Their only use is to improve the relevance ratio of searches. If a system that will pull out everything on a particular subject is wanted, role indicators should not even be considered. If a system that will retrieve a small subset of highly relevant documents, with little noise, is desired, then role indicators merit serious consideration. An experimental investigation of the use of role indicators in various subject contexts may indicate that they are of definite worth in some situations but of doubtful utility in others.

II-10. A Study of Indexing Procedures in a Limited Area of the Medical Sciences, by Judith T. MacMillan and Issac D. Welt, American Documentation 12, 27-31 (January 1961).

Articles discussing the effect of chemical agents on the cardiovascular system have been handled successfully by a combined indexing-abstracting method developed by the Cardiovascular Literature Project. Training of subject-matter specialists to prepare the index-abstract has received particular attention; a training manual prepared for their use contains 234 main biological cross-referenced subject headings plus 72 "see" references, as well as detailed instructions and numerous examples and illustrations. In addition, the indexing is checked against the original documents.

Over a three-year period, 171 papers were indexed twice, either by the same indexer at different times, or by different indexers. The indexing was checked either by the same or by different checkers. The differences in indexing the papers were more obvious than the similarities; as the number of index entries per paper increased, so did the discrepancies among entries. Only 20 percent of the papers were indexed with the same number of subject headings; 6 percent used some synonymous headings rather than identical ones; a total of 74 percent were either duplicate or synonymous, but in each case one indexer added entries. In 48 percent of the cases, index entries themselves were more complete in one of the efforts. Of the 171 articles, 55 were indexed so differently that a correlation between the two instances was impossible. It seems that there is an optimal number of subject headings which may be used for indexing, but more detailed indexing leads to greater possibility for error.

II-11. Testing, Comparison, and Evaluation of Recall, Relevance, and Cost of Coordinate Indexing with Links and Roles, by Barbara A. Montague, American Documentation 16, 201-208 (July 1965).

The performance of three technical information systems in the Plastics Department of the DuPont Company was evaluated by a series of tests. System A contains 5,000 U.S. patents from the polyolefin art covering the period 1926-1960 and including composition of matter, process, articles of manufacture and apparatus patents. System B contains 811 patents from the same body of art covering composition of matter and chemical process only. The 433 patents common to the two systems served as the sample used in Test No. 1. System A indexed to an average depth of 90 terms per document; System B indexed the names of chemicals to an average depth of 70 terms per document. System A used links and a set of 11 roles, whereas System B used no links and five roles. A used an IBM 650 computer with magnetic tape and random access for searching and B used an IBM 305 RAMAC computer. Indexing and searching times were compared: System A's 83 minutes per patent for input includes technical time for indexing and editing, but B's 60 minutes represents time for indexing only, as no editing was performed.

The test consisted of drawing up 29 questions, both real and synthetic, and interrogating both systems. The patents retrieved by each system were screened for relevance by literature analysts. The total relevant was calculated to be the sum of the relevant patents retrieved between the two systems. An analysis of the data showed no statistical difference between the two systems based on a comparison of the ratios of relevant retrieved to total relevant (recall). System A was significantly better than System B based on a comparison of the ratios of relevant retrieved to total retrieved (relevance). System A missed patents due to indexing errors or searching too narrowly. Lack of vocabulary control and generic search capability seemed to account for patents missed by System B.

System C contains all 5,000 patents indexed by System A and consists of a subject index with one or two cross references on abstract cards. Test No. 2 consisted of preparing 33 questions: a random sample of 16 questions from Test No. 1 and 17 others, both real and synthetic. The questions were run by System A, followed by screening for relevance by information scientists, and by System C, with screening by patent attorneys. The high number of irrelevant patents retrieved by System A is due to the nonchemical questions which cannot be characterized as rigorously as chemical questions. The relevance figure of 80 percent for chemicals alone compares favorably with relevance data obtained in Test No. 1. Half the 92 patents missed by System A were due to indexing errors and the other half because the search questions were framed too narrowly. System C missed most of its patents because of insufficient information on the abstract card to alert the searcher and because of searching too narrowly.

An economic evaluation was made of Systems A and C by comparing input and search costs. Input costs for A were $15.10 per patent and for System C $4.00 per patent. Searching costs for A totaled $32 per question, as compared with $102 per question for C. These data reflect the philosophy of indexing by concept coordination, a one-time deep analysis of information at input which permits rapid, relatively inexpensive retrieval on demand.

Test No. 1 compared the performance of coordinate index systems: System A showed higher recall values due to the use of vocabulary control, generic search capability, and deep indexing; its higher value for relevance can be assigned to the use of links and roles to reduce false retrieval resulting from deep indexing. Test No. 2 compared a coordinate index with a classification index: coordinate indexing is faster and retrieves more relevant information.

A series of tests of System A was designed to measure the effect of links, roles, and depth of indexing on recall and relevance. The search questions employed were selected from the ones used in the Tests No. 1 and 2 above. Test No. 3 was run without links, with roles and deep indexing using a random sample of five chemical plus five nonchemical questions, which retrieved a total of 500 patents. Two singificant observations can be made: indexing without links had no significant effect on relevance or preventing false retrieval, and use of links did not affect recall by blocking retrieval of relevant patents. Test No. 4 was run without roles, using the same 10 questions. Recall for chemical questions increased from 83 percent to 93 percent. That is, indexing with roles blocked retrieval of relevant patents. No such reduction was observed with nonchemical questions. This fact points out that inaccurate assignment of roles affects recall of relevant references. But the use of roles does reduce false retrieval and costly screening time.

Test No. 5 was run using the same 10 questions without links and without roles. Their appears to be no interaction between links and roles, and their effects are additive. To test the concept that roles may not be mutually exclusive, each of the 10 questions was interpreted strictly and a single best role selected for each term searched. In this Test No. 6, links were used and depth of indexing held constant. For the chemical questions no appreciable effect is observed on recall or relevance using strict application of roles. The limited application for nonchemical information, however, results in a drastic reduction of recall with only a small increase in relevance. Test No. 7 measured the effect of indexing less deeply on recall of relevant patents. Three levels of indexing were selected: the 40 terms per patent used in System A, the terms used to index the claims plus the first half of the disclosure of the patent, and the terms used to index the claims alone. A random sample of 90 patents was examined at each depth to determine whether they would be retrieved in answer to the question which retrieved them at the first depth. Recall for the six chemical questions sampled was 97 percent at depth 1 and successively lower at depths 2 and 3. A similar trend of reduction in recall with indexing depth is observed for nonchemical questions. An observation made is that increasing depth has a definite effect upon recall but as you progress to deeper indexing levels, recall is less drastically affected. This gives a quantitative measure of the depth of indexing required and shows that there is a point of diminishing return on the depth of indexing investment.

A study of the incremental cost of indexing with links and roles was made by measuring the time required to add these two controls by five literature analysts who indexed 41 documents. The average indexing time per patent was 36 minutes; at a rate of $8.10 per hour, links cost 21 cents per patent and roles cost 54 cents out of a total investment of $15. The last test measured the variation of indexing depth with time. Five analysts indexed a total of 50 documents, noting the number of terms indexed at two-minute intervals; 90 percent of the indexing is completed in 80 percent of the time invested.

II-12. An Analysis of Duplication and Consistency of Subject Indexing Involved in Report Handling at the Office of Technical Services, U.S. Department of Commerce, by Ann F. Painter, Washington, D.C., Office of Technical Services, March 1963, 135 p.

As part of an investigation of the possibility of mechanized information and report handling in the Office of Technical Services (OTS), a study was made of the amount and frequency of duplication, and the consistency with which subject terms are assigned, within several of the systems contributing to OTS: Armed Services Technical Information Agency (ASTIA, now Defense Documentation Center), Atomic Energy Commission (AEC), OTS, and the National Agricultural Library (NAL). Consistency was determined by re-indexing reports and comparing the first indexing with the second. Manual systems, represented

by the AEC and NAL, and mechanized systems, represented by ASTIA, were studied. Each agency was asked to re-index a number of items recently announced, often with a one- or two-month lapse: OTS did 32 items; ASTIA, 94; AEC, 96; and NAL, 99 items. The investigation indicated that the consistency of indexing varies between 65 and 72 percent in both the manual- and machine-oriented systems. There was apparently little difference in the rate in the systems which averaged 2-3 index terms per document and those which averaged 12-16 terms per document. A 65-72 percent consistency rate is not good and would affect the quality of retrieval as well as the efficiency of a table of equivalents to be used for convertibility between all the systems.

There are perhaps four ways to raise the level of consistency: insure adequate subject training of indexers; allow for gaining of experience in indexing specific types of literature; standardize the indexing code or rules to be followed; and provide fairly close supervision of input. Convertibility among systems is dependent on the consistency of indexing. In addition, each of the possible solutions is currently feasible in all systems and should be attempted for quality within an agency's program even without consideration of convertibility.

II-13. A Study of Inter-Indexer Consistency, by Dorothy J. Rodgers, Washington, D.C., General Electric Co., September 29, 1961, 59 p.

The problem under study here, a fundamental problem independent of the specific indexing system used, was what subject content from a document is chosen to be indexed. The indexer must determine what topics are significant enough to be selected as keywords. This study sought information on the degree of consistency with which indexers make such selections. A keyword index system was used because the data collected would provide a minimum estimate of consistency and the system was more easily used without extensive training. Individuals with experience relevant to the subject matter of the documents were chosen as test indexers, to provide some basis for agreement about the selection of significant topics. Sixteen test indexers were chosen; only eight were able to complete the assignment. Twenty articles from Area 5 papers of the International Conference on Scientific Information were chosen, to match the scope of subject specialties and interests of the test indexers. Computer-generated "auto-abstracts" and keywords selected on the basis of frequency had been published for this set of documents, making it possible to compare the words selected in this test with those selected by a statistical system.

The keywords selected were analyzed to determine the degree of agreement among indexers in choice of keywords. A great deal of individual variability was found: the number of words selected unanimously is practically zero; only 11 words were selected by all the indexers for the 20 articles; the proportion of words on which 50 percent agreed was not high; and for most articles over 50 percent of the words were selected by only one individual. The main source of the small proportion of words on which the indexers agreed was the titles and subtitles of documents. The words selected were also compared to the words selected by the statistical system mentioned above, to indicate the degree to which the indexers were affected by the frequency with which a word appeared; the results indicate that this did not have an important effect.

The recommendation is made that effort be directed to development of standard forms in which the indexer's task is limited to selecting specific pieces of information, reducing the opportunity for the wide degree of individual interpretation.

II-14. A Study of Intra-Indexer Consistency, by Dorothy J. Rodgers, Washington, D.C., General Electric Co., January 1961, 25 p.

The two most relevant questions about an indexer's performance are how consistently and how accurately he selects key words for indexing. Explicitness of instructions to the indexer at least partially determines such consistency and accuracy. The indexer studied in the work reported here was instructed to use a Significant Word Indexing System based on the Uniterm system of coordinate indexing, but differing from it in that (1) the terms selected were to be divided into groups of associated terms and listed in order of importance, and (2) terms were to include all names of people, places, organizations, etc., in addition to key words involved in the main thought of the article. Although no absolute limit was set on the number of terms to be selected from a document, the aim was to limit the number to no more than 20. With this set of instructions, the indexer selected significant words from all articles and editorials appearing in the New York Times about the Arab Republic from November 1959 through July 1960. In August 1960 a subsample was selected of 15 documents from November, December, March and June, resulting in a total sample of 60. This sample was also classified into short, medium, and long documents. The indexer selected significant words from these a second time. A consistency measure was derived counting the proportion of words selected twice from the total words selected on both indexings. Analysis indicated that the shorter documents were indexed more consistently. There seemed to be no systematic effect on the consistency measure as a result of the effects of learning and memory. The over-all mean consistency for all documents was .59, not high enough to give confidence in stable indexing. Relationships between number of words selected and length of document were high, as were relations between time to index and length of document.

Subjective analysis suggests that instructions to the indexer were not sufficiently detailed and precise, so that the indexer was forced to rely on his own interpretations which were quite variable. More work needs to be done in developing criteria for selecting key words, so that these criteria could be reflected in precise and objective instructions to indexers.

II-15. An Evaluation of Links and Roles Used in Information Retrieval, by Jefferson D. Sinnett, Thesis presented to the Air Force Institute of Technology, Air University, Wright-Patterson Air Force Base, Ohio, Air Force Institute of Technology, December 1963, 300 p.

The object of this evaluation was to determine whether use of links and roles will cause a favorable reduction in the retrieval of irrelevant information without incurring an unfavorable loss of relevant information. The information retrieval system tested was under development for the Air Force Materials Laboratory of the Air Force Systems Command, Wright-Patterson Air Force Base. It is a coordinate index system using an NCR 304 computer for retrieval. At the time of this study 6,280 documents had been indexed and stored in the system. A total of 24 questions was put to the system: 18 prepared from 16 documents randomly selected from the store, 4 "line" requests made to the system, and 2 originated by the author. Four retrieval methods were followed: the basic coordinate index system unmodified, the basic system modified with links, the basic system modified with roles, and the basic system modified with both links and roles. Documents in the common set produced by the four searches were assumed to be relevant (samples were analyzed to test the validity of this assumption), and documents not in the common set (called complementary sets) were analyzed to obtain a judgment of their relevance. The first method (basic system unmodified) was considered the standard and used to approximate an absolute measure of efficiency. A relevant document was defined for this study as one in which the questioner found information in relation to his question.

Values were computed for number of documents retrieved, the relevant set from the first method to be used as the standard, the relevant sets from the other methods, and the irrelevant sets produced. The retrieval effectiveness for each of the four methods was scored by a modification of Borko's formula (cf. IV-9), R=K (r-i), where K = a constant factor, r = relevancy score, and i = irrelevancy score. The conclusion reached was that the four methods are ranked from highest to lowest in the order: second method, first, fourth, and third method. The second method (modified with links only) attained the highest average retrieval effectiveness score, and gave a greater than 56 percent reduction in irrelevant documents with a less than 5 percent loss in relevant documents retrieved. The recommendation is made that the use of roles in this system be discontinued, resulting in increases in the rates of indexing and retrieving answers.

II-16. Classificatory, Alphabetical, and Associative Schedules as Aids in Coordinate Indexing, by Vladimir Slamecka, American Documentation 14, 223-228 (July 1963).

In studying inter- and intra-indexer consistency, three types of situations are apparently associated with low indexing reliability: (1) those inherent to the indexer, his background, his psychological makeup, and his experience; (2) those contained in the indexing situation, the choice of system and the kind and thoroughness of rules and instructions; and (3) those inherent in information contained in documents, the variety of relations possible between concepts and the different emphasis accorded to the same subject by different authors. One kind of indexing control, the purpose of which is to reduce disagreement among indexers, is indexing aids, including lookup tools and devices providing feedback to the indexers. Indexing aids can prescribe the terms to be assigned, and (or) suggest alternative or additional concepts and terms to be considered. Prescriptive indexing tools lead to standardized use of vocabulary by guiding the indexer in consistent use of terms, by substituting synonyms, and by indicating maximum degree of specificity. The suggestive type of aid draws the indexer's attention to related terms but does not instruct him to apply them mandatorily. At present, the effect of indexing aids on indexer reliability depends on the compatibility of the aid with the indexing system, the specific role of the aid, the prescriptive force of the aid, and the frequency of use of the aid as stipulated by indexing instructions. With more knowledge of the relative weight of factors causing indexer inconsistency, indexing aids could be designed to focus attention on specific causative factors.

II-17. A Multiple Testing of the ABC Method and the Development of a Second-Generation Model. Part 1. Preliminary Discussions of Methodology, by Berthold Altmann, Report No. TR-1295, Washington, D.C., U.S. Army Materiel Command, Harry Diamond Laboratories, April 1965, 83 p.

The first-generation ABC storage and retrieval method is an HDL-developed method that utilizes appropriate standardized English-language statements processed and printed by a KWIC-type computer program. It is a completely automated library system; this report deals with only one major segment of that system, the method of analyzing items and the procedures for organizing, storing, and retrieving them. The system, called ABC (Approach-by-Context), was designed for dissemination and recall of very specific technical information. The method uses the natural language of the scientist and engineer to index and retrieve information. A key to this method is a very specific index (dictionary) of the collection. In obtaining a response to specific problems or questions, the investigator himself consults the dictionary of concepts. The lines are referred to as "concepts" and are concise, meaningful, self-explanatory statements of control. The investigator compares his formulation of a problem with the standardized descriptions of available information.

The study and further development of the ABC method concentrated on three tasks. The first was to design and perform an objective test and to determine the validity of major assumptions and claims about the system. The second task was a critical assessment of the first-generation method, based on several typical retrieval operations; the third task deals with a second-generation model based on experience gained in the test.

The ABC method was subjected to a performance test at the request of DOD. The work was performed within the Technical Information Office in HDL, using largely "volunteer" help. A requirement imposed on the test was adherence to certain procedures that personnel at Cranfield had introduced (cf. I-2). Only by following the general outline of their tests could HDL produce results amenable to comparative analysis. To circumvent certain disadvantages in the approach, however, additional control test runs and changes and adjustments to duplicate situations more common to local experience were made.

This report covers the first descriptive part of the test report; a statistical analysis is in preparation.

A test collection was established of 3, 650 journal articles and technical reports on solid state devices. The subject area was small enough to permit comprehensive coverage with that number of items. Selection of that particular subject expedited critical analysis for indexing and facilitated establishment of an additional tool for checking on the completeness of the retrieval and determining the recall ratio. This auxiliary control was based on catalog cards and abstracts supplied by DDC for technical reports and similar cards for journal articles supplied by Cambridge Communications Corp. (CCC). This card catalog was organized by a detailed subject classification scheme prepared by CCC.

The selected papers were sent to the information analysts for evaluation and formulation of ABC concepts. The concepts were standardized and punched into cards. In a parallel effort, the items were cataloged and the descriptive titles key punched. Both types of information, concepts and titles, were transcribed on a catalog file tape. An IBM 7090 computer produced eight catalogs from that tape: two ABC dictionaries listing concepts and term-letter code combinations, differing in numbers of alphabetized key words; a catalog with titles filed under appropriate term and code; a list and a card file of letter codes with the concepts they signified; a card file of accession numbers with titles; a KWIC title list; and a file of reports listed numerically by AD numbers. In a separate effort (not a machine process) the abstract cards from CCC and DDC were organized into a subject card catalog.

Two groups comprising about 40 scientists and engineers formulated two sets of test questions. The computer selected at random 400 titles from the catalog of the collection, and printed them as a list providing the subject classification supplied by CCC, and the title of the periodical or the DDC accession number. From this list each of the scientists and engineers selected subject areas consistent with his interests from which to derive test questions, and then selected documents of interest, examined the text, and formulated a test question that could be answered by its content. About 225 questions were obtained in this manner. To provide a degree of realism, 36 additional questions were formulated by other scientists and engineers, based merely on general knowledge of the subject areas of the collection and on their own experience.

The questions, prior to use in actual retrieval operations, were evaluated, combined, and edited to eliminate deficiencies and redundancies. These evaluative and editorial responsibilities were assigned to a group of senior scientists and engineers. As a result of this process, 100 questions formulated by the first group were approved in addition to the 36 questions formulated above.

Three methods, two variations of the first-generation ABC method together with a KWIC-title list, were tested. A control group was established to discern the bias possibly introduced by the use of the test operators' own questions and the sequence of the three test runs. In addition to the 40 scientists and engineers (divided into subgroups 1A and 1B) mentioned above, six research analysts and six librarians searched the 136 questions by all three methods. Therefore, 1, 632 retrieval sheets containing more than 6, 000 documents were obtained. Each searcher was asked to retrieve documents using tools in what was called the normal sequence: the short ABC dictionary, the KWIC list of titles, and the long ABC dictionary. To provide for additional evaluations, subgroup 1B was asked to use the KWIC title list first and the short dictionary second. The individual operator received at one time all his questions for testing one of the three approaches, and turned in those answers before receiving a duplicate set for processing with the second tool. The interval of at least one day between the two retrieval operations was allowed on the assumption that the retriever had forgotten the codes and accession numbers recorded during the first run and his previous experience would not influence the results of the second run. In a major deviation from the Cranfield test the retrieval was conducted freely without knowledge of the "answer". It was felt that the retrieval of the basic paper could not and did not signify the success or the end of the operation.

The relevance and recall ratios will be calculated and used to evaluate the retrieval method. What was attempted was a test of the ABC method and an evaluation of its efficiency in terms which permit comparison with similar systems. The relevance and recall ratios are reasonable measurements if we are in a position to rate an average performance in terms of user satisfaction and capability of locating all appropriate titles in the storage system. In practice, however, this cannot be easily accomplished. The factor of human fallibility poses another obstacle to consistent evaluations and objective comparisons of different storage and retrieval systems. The views on the validity or more usefulness of tests vary drastically. But test design and test procedures have been greatly refined. In the test here, the entire operation was an experiment and an opportunity to analyze the bread-board model that was built, to analyze the failures of the total system as well as of its individual components, and to redesign or adjust the retrieval methods.

To insure the reliability of the data obtained during the official test runs, the four groups of retrievers were organized so as to resemble the profile of actual users. The human error factor was limited, the bias eventually introduced by the testing procedure was identified, and the evaluations were "objectivized". The methods used to accomplish these ends were critical analysis, generation of multiple test data, and the introduction of control groups and stringent controls. The disquieting connotation of relativity and subjectivity generally inherent in the term "relevance" was removed by comparison with the contents of a document or the substance of a question. The conceptual substance of the question was analyzed, relative weights attached to the various conceptual components and their combinations, and the resulting scales used as a measuring stick.

The test questions were processed eight times according to the ABC method. The eight different runs may help to determine whether the causes for failure may be traced to the human element or to the system. Control factors and control groups have been used to insure realistic conditions, reliability and accuracy of data and consistency of testing procedures. The three retrieval loops developed are control "mechanisms" to assist the evaluator to recall from the collection all titles of relevant papers which the operators had failed to retrieve. The retriever evaluated the contents of the retrieved documents, compared them with the basic document, and indicated their value (=, +, -, or 0). Questions not based on specific documents were graded on the quality of the answer. The senior scientists and engineers who had approved the questions evaluated the results. Using the CCC card catalog, pertinent titles not retrieved during the test runs were determined and graded; this provided the basic data necessary to determine the recall ratio.

To determine the cause of failure, an evaluation of all concepts used in a given run was made, the essential elements signifying the contents of a question identified, and the combination of them that would have appeared in appropriate concepts were graded (+, -, =, or 0). The concepts used in retrieving were graded by comparing them with the theorized combinations.

The evaluator, in locating all titles related to the formulated questions, was to first turn to the ABC dictionary to find additional approaches missed by the members of the retrieval groups, and if successful was to follow the standard retrieval method, checking pertinent titles in the ABC card catalog. In the second loop, the evaluator was guided to the respective subdivisions of the subject card catalog where all related information was combined. He examined the ABC card catalog for the location of secondary concepts and exhausted this loop as before. Finally the evaluator entered the third loop, screening the KWIC title list under all possible pertinent and significant terms, leading first to the title catalog and second through the code to the ABC card catalog where he could find secondary concepts to close the loop. This rather elaborate scheme was provided and followed in a number of cases to obtain all applicable documents on a given question from the collection.

The forms used permit question-by-question comparisons of the results by the four operator groups in numerical and qualitative terms; determination of the relevance ratio; organization of the results by individual retrieval operators for each of the tools used and each of the questions retrieved; determination of the recall ratio, the relationship between quality of results and length of retrieval time, and the percentage of instances when the basic document was not recovered; and fast review of the results obtained by the four groups using each of the three retrieval tools.

A cost account of the test, in which all expenditures were reduced to unit cost, shows a total of $3.55 per title spent to organize the test collection in accordance with the ABC storage and retrieval method.

II-18. Thesaurus Controls Automatic Book Indexing by Computer, by Susan Artandi, in Automation and Scientific Communication, Short Papers, Part 1 (papers presented to the 26th Annual Meeting of the American Documentation Institute, Chicago, Ill., October 1963), Ed. by H. P. Luhn, Washington, D.C., American Documentation Institute, p. 1-2.

(The work reported here appears in greater detail in Book Indexing by Computer, by Susan Artandi, Thesis presented to the Graduate School of Library Service, Rutgers, New Brunswick, N.J., Rutgers, The State University Press, 1963, 207 p.)

A method was developed for automatic derivation of index entries from natural language text, based on a dictionary of terms for a given subject area (inorganic chemistry) and a computer program to compare terms of the list with words of the text (portions of an inorganic chemistry textbook). Measures were developed to evaluate the product of the method, based on a comparison with an existing generally accepted manually produced index (average of indexes found in inorganic chemistry textbooks). Completeness of indexing, including qualitative and quantitative criteria, was found to be practically identical for the experimental index and the average of the published manual indexes checked. Entry density (ratio of number of page references to number of pages) was 63.8 percent higher for the mechanized index; heading density (ratio of number of entries to number of words in the book) was 8.8 percent lower than the manual indexes. Cost of indexing was also calculated. Cost-per-page figures are too high to be competitive with manual indexing. A major cost is the conversion of text for machine input; general availability of texts in machinable form would increase the competitive efficiency of the experimental method. High cost may be justified, however, where uniformity of indexing is important.

II-19. Measure of Indexing, by Susan Artandi, Library Resources and Technical Services 8, 229-235 (Summer 1964).

In conjunction with a project on automatic indexing of natural language text (cf. II-18) measures were developed for comparison of the mechanically-produced book index with the average conventionally-produced index for the same type of material. It was necessary to establish an average level for manually-produced book indexes, against which to test the computer-produced index. Evaluation is based on three criteria: density, distribution, and completeness of indexing. There are two measures of density: heading density, the ratio of the total number of index entries to the total number of words in the book, and entry density, the ratio of total number of page references to total number of pages. Distribution considers the ratios of index entries for chemical compounds, proper names, and other types of subjects to the total number of index entries. The measure for completeness is not based on objective data but includes the indexer's point of view concerning the level of indexing he wants to achieve, which may be due to such factors as cost and space considerations or editorial policy. Completeness of indexing is calculated on a statistical basis by taking a random sample of the pages of the book and expressing the ratio of the number of entries in the index for the sample pages to the sum of that number of index entries plus the errors found in the index for the sample plus the entries not found in the index (i.e., where an entry should occur according to the indexer's pattern for the choice of entries but does not).

For the purposes of the mechanical indexing project the average of the three evaluative criteria calculated for five experimental titles was taken as the level of indexing for comparison with the mechanically-produced index.

II-20. Application of a Telereference System to Divisional Library Card Catalogs: A Feasibility Analysis, by F. R. Bacon, N. C. Churchill, C. J. Lucas, D. K. Maxfield, C. J. Onant and R. C. Wilson, ERI Project 2733-1-F, Final Report to Council on Library Resources, Inc., Ann Arbor, Mich., University of Michigan, Engineering Research Institute, May 1958, 91 p.

A method of using closed-circuit television equipment for remotely controlled catalog card viewing was studied for its feasibility in terms of the amount of equipment needed and any cost savings which might result from replacing card catalogs in divisional libraries. The amount of equipment needed to replace catalogs was estimated for 16 of the 43 divisional libraries in the University Library system, plus the Undergraduate Library and the Public Catalog in the General Library. Cost savings were estimated on the basis that procedures would be eliminated or modified; costs were calculated in terms of direct labor, direct labor overhead, and materials involved. Characteristics of card catalog use studied were arrival rates, catalog holding time, and peak load demands. The conclusion was reached that use of Telereference equipment cannot be justified on the basis of savings in direct costs, which would amount to only 50 percent of the costs of maintaining the equipment. No attempt was made to estimate the value of added services which might be provided.

II-21. A Comparison of Relevance Assessment by Three Types of Evaluator, by Gordon C. Barhydt, in Parameters of Information Science (Proceedings of the American Documentation Institute, 1964 Annual Meeting, Vol. 1, October 5-8, 1964, Philadelphia, Pa.), Washington, D.C., Spartan Books and London, Cleaver-Hume Press, 1964, p. 383-385.

Fourteen questions, submitted by members of a pilot user group of educational researchers, were searched over a file of 4,000 documents to investigate the effectiveness of measures of relevance based on non-user evaluation. A sample of 50 responses, randomly selected from the total number of responses, was given three evaluations: by the user, the individual posing the question; by another subject specialist; and by a system specialist, neither user nor subject specialist. The three were asked to indicate for each response in the sample whether it was relevant or non-relevant. Both expert and system specialist were given copies of the questions as submitted by the user. No additional information was provided and no contact was made with the user.

The effectiveness of the relevance assessments by expert and system specialist was determined by the method developed by Goffman and Newill (cf. III-9). The results indicate that for most questions the effectiveness of both expert and system specialist was quite low. The results also indicate a high correspondence in the measures of sensitivity, specificity, and effectiveness, between expert and system specialist. They are apparently not making random selections.

Because of the small number of questions used for this report no definite conclusions can be reached.

II-22. Selective Dissemination of Information.SDI 2 System, by W. Brandenberg, H. C. Fallon, C. B. Hensley, T. Savage and A. J. Sowarby, Yorktown Heights, N. Y., International Business Machines Corp., Advanced Systems Development Division, April 18, 1961, 95 p.

The SDI 2 System sends new items of information to users who need such information by matching characterizations of users' interests with characterizations of documents' contents. Notifications of some documents are sent to each user on a random basis. Monthly reports show the ratio of positive and negative responses of users. These random selections and reports are used to evaluate the system's performance. The two variables, proportion of negative responses ("trash") and estimated proportion of notices not sent ("miss"), are affected by document and user population, document and user profiles, and the setting of the matching criterion p (p = the number of a user's keywords which must match the keywords of a document). Raising p reduces trash and increases misses; lowering p has the opposite effect. Besides miss and trash, the setting of p can control the flow of notices and documents through the system: raising p will decrease the number of notices on a document and increase the probability of these notices being positive responses. SDI 2 has experienced between 50-70 percent positive response to notices where p was set between 12-24 percent and document profiles were 26 keywords or less.

II-23. A Progress Report on Search-System Evaluation in the United States Patent Office, by Edward C. Bryant, Report No. WRA-PO-14, Denver, Colo., Westat Research Analysts, Inc., May 1964, 27 p.

The three considerations of backlog, accuracy of examination, and cost are the criteria by which the magnitude of the Patent Office problem is measured. The system problems relate generally to the effectiveness of the patent examination system --- how well it operates and what would be the effect of proposed changes. Study of the system as a whole has been done through the construction of a mathematical model and estimation of its parameters for the Chemical Operation. The model considers the flow of applications through the system in terms of probability; the model is therefore referred to as a "stochastic model". Actual data have been collected tracing the flow of 493 applications through the Chemical Operation. One use of this study has been to project the expected changes in composition of the patent backlogs in view of changes in procedures.

A program was begun to "quality rate" a random sample of the actions completed by the Examining Corps. Quality ratings were made by group supervisors, resulting in immediate feedback to the examiners. A preliminary study of the system of reclassification and the advantages to be gained from it points out the need for carefully designed experimentation in that area.

All of the non-conventional search systems in operation at the Patent Office involve some sort of coordinate indexing. A fundamental understanding of the limitations of such systems is necessary to any evaluation. Models have been constructed to show the effect of indexing errors on retrieval. One can improve the efficacy of the systems by modifying the indexing to guard against errors, by modifying search strategy in view of known errors in indexing, and by training the examiners in most effective use of the systems. Some evaluation of specific systems has been accomplished and it can be concluded that coordinate indexes can be constructed for well structured arts which will assist the examiners. An experimental index is being constructed for heterocyclic compounds involving around 30,000 documents. This represents the first opportunity to investigate problems of scale in coordinate indexes.

The evaluation program must be divided between search and retrieval studies and operational studies. The latter can be lumped into a category called systems evaluation and modification studies. In the search and retrieval area are relatively small systems demonstrated to be useful to the examiner and also research into automatic indexing and searching of natural language text. Studies are needed to determine the costs of various potential systems and the gains to be expected from them. At the same time programs at the intermediate level of replacing human judgment by machines should be statistically evaluated.

II-24. Analysis of an Indexing and Retrieval Experiment for the Organometallic File of the U.S. Patent Office, by Edward C. Bryant, Donald W. King and P. James Terragno, Report No. WRA-PO-10, Denver, Colo., Westat Research Analysts, Inc., August 1963, 52 p. plus appendices.

Research reported here focuses on the accuracy and consistency with which documents are indexed and coded, and the effect of errors or inconsistencies in indexing on the retrieval of documents. The particular file serving as the basis for the experimental work is the organometallic chemical file of the U.S. Patent Office, consisting of 3,625 patents. The indexing system is based on fragments of chemical compounds and their descriptions. Standard practice called for an analyst to index the document and one of the most experienced analysts to review the indexing, adding or deleting codes in accordance with his best judgment.

The primary characteristics of interest are a measure of accuracy of indexing, two measures of consistency (or inconsistency) of indexing, number of codes involved, and time required to index a patent document. Several analyst or indexing modes were evaluated: a single analyst, a single analyst with a reviewer, two independent analysts with their indexings combined either by set sum or intersection, two independent analysts with a reviewer, and three independent analysts. A random sample of 201 patent documents was chosen for the test, and a random subsample of 24 documents was chosen for an intensive study of indexing. Of the twelve analysts in this study, six had experience in working with organometallic compounds; therefore half of the documents were indexed by analysts in each of the experience groups to permit evaluation of the experience factor. Four reviewers, two with extensive experience, were used, so the effect of their experience could be assessed also. Accuracy was measured by estimating the conditional probabilities that a code will or will not be selected given that it should be. The "correct" code was determined from all the indexings in the experiment plus the original indexing and review for the 24 documents. The two most senior analysts performed this reconciliation. The experiment provided estimates of conditional probabilities for specific codes and individual analysts as well as averages over all analysts and over all codes. It was possible also to estimate the probabilities for each of the analyst modes.

Two measures of consistency or inconsistency were computed for one third of the codes, after eliminating those which appeared with very low frequency. The results, plus discussion of the measures, appear in another report (cf. II-25). The consistency coefficient is essentially the number of documents in which the given code was indexed by both analysts divided by the number in which the code was indexed by either. The index of inconsistency is based on the ratio between the number of documents in which there is disagreement between the indexers and a scale factor dependent on the proportion of documents indexed with a given code. The high correlation between consistency and accuracy, as determined by comparison with a carefully prepared indexing, provides a basis for controlling the quality of indexing by the use of consistency coefficients.

When compared against the "exact" code, an indexer selected about 86 percent of the codes he should have selected and only added about 3 percent more codes than he should have used. In terms of consistency, about 73 percent of the codes selected by either indexer were selected by both. Differences due to experience in indexing organometallic compounds were generally found to be non-significant. Document-to-document variation was high, causing the test to be insensitive to the experience factor. There is some evidence of interactions between analysts and reviewers. The hypothesis that a reviewer will cut out superfluous codes, thereby reducing false retrieval, is discredited --- it appears to be better to use a set sum of two indexers rather than an indexer and a reviewer.

To determine the effect of differences of indexing by different analysts in terms of the ultimate document retrieval, the indexing of the 201 documents chosen in the random sample was repeated and mechanized searches of the sample file were conducted. (One of the principal reasons for repeating the indexing was to determine measurements of consistency and to compare these with measures of accuracy, using the reconciled coding as the "correct" coding.) The search questions used were 184 actual questions formulated by examiners in the past and 172 synthetic questions formulated by analysts from patent claims. Again, a "true" set of codes was selected by two senior analysts for the 201 documents and this set was used to measure the number of missed documents and false drops resulting from errors in indexing. Five separate decks of punched cards were prepared: "true" codes, single-analyst mode, repeated single-analyst mode, double-analyst mode, and single-analyst-reviewed mode. Searches were conducted simultaneously on these five decks on a Bureau of Census Multiple Column sorting machine. The observed retrievals provide estimates of the distribution of the number of documents in the retrieval, the proportion of correct retrieval and the distribution of the number of false drops resulting

from indexing errors. Correct retrieval is defined to be the retrieval of the "true" deck for the documents containing the codes used in the search questions.

Indexing consisting of all of the codes selected by either of two independent analysts achieves a retrieval of 91 percent of the documents which a "perfect" indexing would yield, with a false retrieval of only 33 percent more than the perfect indexing. Although there were differences in the results of searches using actual questions and those using synthetic questions, the principal conclusions would have been the same had either set of questions been used alone. It was postulated in another report (referred to above, cf. II-25) that the proportion of missed documents due to indexing errors is $1-p_3^r$, where p_3 is the probability that a code will be selected given that it should be, and r is the number of codes used in the search question. The model was found to be representative of actual practice for indexings accomplished by two or more analysts. For single analysts, an adjustment must be made for dependence of indexing accuracy from code to code within the document.

II-25. Some Technical Notes on Coding Errors, by Edward C. Bryant, Donald W. King and P. James Terragno, Report No. WRA-PO-7, Denver, Colo., Westat Research Analysts, Inc., July 1963, 30 p.

By "coding" is meant the selection of a term, commonly called "indexing", and its translation into a numerical code, commonly called "encoding". Failure to select a term that should have been selected will be called an error of Type I, while selecting a term which should not have been selected will be called an error of Type II. We envision a search in which all of the documents coded by particular terms are retrieved, where it may be specified that the documents of interest must possess all of the terms (set intersection), any of the terms (set union), or any intermediate combination of unions and intersections.

For a search based on a single term, we can calculate the expected retrieval of desired documents plus not desired documents, and the expected number of documents not retrieved which should have been. For a search based on two terms, we calculate as above for both set intersection and set union. Retrieval under an intersection strategy is more sensitive to coding errors of Type I than is retrieval under a set-sum strategy (i.e., the percentage error is less). Coding errors of Type II increase the percent retrieval under a set-sum strategy. If one can control Type I errors, but not Type II, it is better to use an intersection strategy than a set-sum strategy. The reverse is also true.

For searches based on three terms, the calculations show the sensitivity of the intersection strategy to Type I errors by the small fraction of desired cases retrieved. The false retrievals are excessively high for the set-sum strategy.

The conclusions are that when intersection strategies are used it is important to reduce Type I errors; set sums of codes of several coders might be used; coders should be instructed to use a term if there is any doubt as to its applicability; perhaps multiple searches, each one composed of different sets of terms, should be used; the effect of coding errors on retrieval and the necessity for control of such errors should be experimentally verified.

The importance of accurate coding is difficult to assess. Consistency of coding is relatively easy to measure, and it seems that consistency and accuracy should be closely related. Some experimentation in the organometallics file of the U.S. Patent Office is studying relationships among various measures of consistency (cf. II-24). In this file most coding errors are of omission rather than of commission. Also the coding structure,

based largely on chemical formulas, causes less error due to interpretation than might be true in less formal subject matter. One can feel that for carefully structured indexing schemes measures of consistency will yield useful information about the accuracy of the file.

The calculations for a coefficient of consistency use the number of times a term was coded by both members of a pair of analysts as the numerator (divided by the average number of times it was coded by either). This variable was averaged over all possible pairs of analysts; an adjustment factor for the variances of such an average, to put them on an "independent-pairs" basis, has been computed. The dependent-pairs variance is multiplied by the correction term.

II-26. Is Automatic Classification a Reasonable Application of Statistical Analysis of Text? by Lauren B. Doyle, Report No. SP-1753, Santa Monica, Calif., System Development Corp., August 31, 1964, 34 p.

This article traces the development of statistical analysis of document collections, from the work of Luhn, Maron, and Stiles to the recent trend toward automatic classification study. A debate over classification versus coordinate indexing had been going on for some time, but recognition of the inadequacies of coordinate indexing for large collections led followers of the statistical approach to seek to rescue or even replace it with statistical analysis and classification. Classification and coordination need not be regarded as mutually exclusive alternatives --- in computerized systems they can be complementary.

The author's early work took the form of a two-dimensional layout of the most strongly co-occurring word pairs, called an "association map". These maps could be used in preparation of a thesaurus, as an aid in thinking up search request statements, or as an integral component of a retrieval system. The first change in the map idea came when it was realized that word-hierarchies could be generated by linking only unequally frequent words. Hierarchical association maps were more orderly and probably more useful for retrieval. Since classification and coordinate machine-searching can be combined, and since construction of such a system (given that the problem of automatic classification is solved) is an engineering problem, the remaining question is of quality: can automatic classification equal or exceed classification of human judgment? But human judgment may not be the best standard, so a simple, small-scale test involving non-judgmental criteria was conducted. The hypothesis under test was that automatic classification should improve as the amount of information per document increases.

A corpus of 100 items on the topic of computer processing of natural text was used; each item (a document or a portion of a document) was represented by a 36-word list whose words were arranged in decreasing order of frequency. These lists served as input to a Philco 2000 version of the Ward and Hook hierarchical grouping procedure. Six runs of the program were made, involving successively 12, 15, 19, 24, 30 and 36 words from each list. Five criteria were studied: (1) time ordering, since 50 of the items were taken from daily work-records and diaries; (2) grouping of portions of documents; (3) grouping according to author; (4) another time criterion, since the other 50 items were orderable by date of publication; and (5) increased stability in the groups formed as the number of words increased. It was arbitrarily decided to compare the six runs at the point where seven entities remained to be grouped in the time-ordered material (i.e., the items had been distributed into seven categories). It was hoped that each of the categories would contain an unbroken time-ordered series, and that none of the time-ordered items would be filed outside of the seven categories. The number of time-breaks diminished as the words increased from 12 to 36, and the number of items filed outside decreased to the vanishing

point. Two more criteria showed improved classification with increased number of words per list, namely regrouping of segments and grouping by author. In progressing from 12 to 36 words, substantial progress was made toward an ideal or maximum attainable number of such groupings.

The 100-list collection was sufficiently topically homogeneous that these criteria variables had a fairly decisive effect on classification. The other two criteria gave rise to no significant confirmation of the hypothesis.

The major barriers to the application of automatic classification are expense and the problem of representation or display of the output of the classification process. The former can be resolved with newer equipment and refinements of technique; the latter needs to be worked on.

II-27. Problems in Automatic Abstracting, by Harold P. Edmundson, Communications of the ACM 7, 259-263 (April 1964).

There are several major classes of problems in automatic abstracting: conceptual, input, computer, output and evaluation problems. The brief discussion of the latter is of interest here.

The first problem concerns the acceptability or utility of the final product, usually requiring a comparison between an automatic abstract and an "ideal" human abstract. However, the linear coefficient of correlation among human abstractors varies from .2 to .4, even when they operate under moderately well-defined abstracting rules. This is due in part to the fact that the correlation coefficient is not the best measure. If two individuals happen to select different but co-intensional sentences, then the correlation coefficient will be low. The problem of what sentences of a document are co-intensional is solvable only by further semantic research.

The second problem of evaluation is that of system cost in dollars and in time. Insufficient data have been collected to permit reliable estimates of cost per document and estimates of bounds on the error for an operating system. Such information does exist for research systems.

II-28. Evaluation of the Performance of an Information-Retrieval System by Modified Mooers Plan, by E. M. Fels, American Documentation 14, 28-34 (January 1963).

The mechanized system for retrieval of legal documents, established at the Health Law Center of the University of Pittsburgh, was evaluated statistically by a modification of the Mooers plan (cf. IV-27). Comparison was made between one system based on the index parts of Purdon's Pennsylvania Statutes Annotated (hand or manual search) and the mechanized system based on machine scanning of the full text, without intervening indexing. The modifications to Mooers' plan consisted of using his absolute measures as comparative measures and of adding an "umpire" to the "customers" who supply questions based on samples of documents shown to them. The customers (called "writers" here) submit their questions to the umpire who corrects and controls the assignment of the relevance values. A probability sample of 20 units was chosen with the help of a table of random numbers; three questions for each of the units gave 60 ordered pairs (the number was dictated by economic feasibility). A relevance-value matrix was used to depict accidental relevances in the sampling units. Print-outs of retrieved sections pertaining to the

questions were scanned for those sections picked in the random sample. Computed document ratios (point estimates of corresponding retrieval probabilities), confidence limits, exact tests of significance for the differences between corresponding ratios and the power function of these tests, were computed. One effect was to show up the insufficient sample sizes in the test. Although one table showed that machine retrieval is not better than hand search, another table showed that it is premature to assert that the hand search is clearly better. Much larger sample sizes are needed.

II-29. Evaluating Coordinate Indexing Systems, by J. Jaster, Barbara Murray and Mortimer Taube, in State-of-the-Art of Coordinate Indexing, Washington, D.C., Documentation, Inc., February 1962, p. 81-107.

There are several factors to be considered in evaluation of information systems. These include acquisition, indexing, the index, constraints (categorization, cross-referencing, use of roles and links, etc.), search procedure, and user education. Attention is directed here to studies of indexing and of indexes with and without constraints. Indexing studies include an informal experiment at the Man-Machine Information Center at Documentation, Inc., in which several people indexed the same documents with no limit on time or number of terms to be used. Point of view was a significant factor; the importance of depth of indexing was obvious; overindexing can lead to retrieval of excess material; restriction in number of terms used can lead to more careful selection of terms. MacMillan and Welt's study of indexing of cardiovascular literature (cf. II-10) also points up indexing differences among subject specialists, and gives figures for doubly-indexed papers which were indexed the same or differently. The work represents an attempt to make deductions from actual experience. A large-scale statistical investigation into indexing consistency among indexers, and any one indexer's consistency, was conducted by Documentation, Inc., for Rome Air Development Center (cf. II-4).

There have been many studies of indexes, from comparisons of coordinate indexing and classification systems to measuring the value of adding constraints to an index. Two major studies are those of the ASLIB-Cranfield project (cf. I-2) and Schüller's study of UDC and Uniterm (cf. I-27). The Cranfield report gives results of five tests of retrieval efficiency and gives efficiency percentages for UDC, alphabetic, faceted, and Uniterm indexes. (The Uniterm system displayed no significant superiority over the others.) Schüller's more modest study recommends the use of the Uniterm system for technical reports (as a complement to the UDC system). An important point is repeated here: test results are, even when correct, not universally applicable.

Three methods of evaluating indexes could be (1) analysis of false drops, (2) use statistics and (3) user studies and determination of user requirements. But it is difficult to define a false drop, and no one has disclosed a means of analyzing the reason for non-pertinency of material. As for use statistics, studies in frequency of term usage and studies of the "association factor" (i.e., all terms indexed along with a particular term) are underway and should be useful not only in evaluating but also improving a system. The work of Stiles (see Stiles, H. Edmund, The Association Factor in Information Retrieval, Journal of the ACM 8, 271-279, April 1961), Stevens (see Stevens, M.E., A Machine Model of Recall, in Information Processing (Proceedings of International Conference on Information Processing), Paris, UNESCO, 1960, p. 309-315), and Schultz (cf. I-2 and II-50) in this area are noted. With user requirements, although Taube's study (see Taube, M., An Evaluation of "Use Studies" of Scientific Information, in Emerging Solutions for Mechanizing the Storage and Retrieval of Information. Studies in Coordinate Indexing, Vol. V, Washington, D.C., Documentation, Inc., 1959, p. 46-71) concluded that such studies based on interviews with scientists were not of much value in design, other studies

based on analyses of reference questions might more nearly apply to the design and evaluation problem (see Herner, S. and M. Herner, Determining Requirements for Atomic Energy Information from Reference Questions, in Proceedings of the International Conference on Scientific Information (Washington, D.C., November 1958), Vol. I, Washington, D.C., National Academy of Sciences-National Research Council, 1959, p. 181-187).

II-30. Project SHARP (SHips Analysis and Retrieval Project) Information Storage and Retrieval System: Evaluation of Indexing Procedures and Retrieval Effectiveness, by Walter F. Johanningsmeier and F. Wilfred Lancaster, Report No. NAVSHIPS 250-210-3, Washington, D.C., Department of the Navy, Bureau of Ships, June 1964, 49 p.

(The subject matter of this report was contained in "A Case Study in the Application of Cranfield System Evaluation Techniques", by Saul Herner, F. Wilfred Lancaster and Walter F. Johanningsmeier, paper presented to 148th National Meeting of American Chemical Society, Chicago, Ill., September 1964, Washington, D.C., Herner and Co., 14 p.)

The automated information system of the BuShips Technical Library was studied and evaluated by indexing documents according to the coordinate indexing system, then addressing queries to these documents and analyzing the search results. The subject area was undersea warfare; the 1,000 documents selected for indexing reported on research sponsored by BuShips and dated 1956-1961.

Coordinate indexing with links and roles as advocated by the Engineers Joint Council (EJC) was followed, using the BuShips Technical Library Thesaurus as the authority list. (Prior to the study, the indexers took the EJC-Battelle one-week indexer-training course.) After the subject indexing was completed, the descriptors were coded for entry into the IBM 7090 computer at the David Taylor Model Basin, including codes for broader terms where applicable and for link designations. Testing was accomplished by a modification of the techniques developed by Cleverdon in the Cranfield project (cf. I-2).

Technical personnel at BuShips compiled test questions based on documents in the collection (source documents). Fifty queries (questions translated into the system language with descriptors and roles) were formulated, with a series of subqueries involving all major combinations of appropriate descriptors and role indicators. For all searches, the documents retrieved were reviewed by the compilers of the questions, to decide whether they were relevance A (as useful as the source document), relevance B (of some relevance to the question), or non-relevant documents. On the basis of these judgments it was possible to compute relevance ratios. In addition, for 10 questions selected at random the entire collection was searched manually to locate all documents of any possible relevance to those requests, in order to compute recall ratios. For all searches, reasons were sought as to why relevant documents were missed and why non-relevant material was retrieved.

Considering documents of any relevance (source documents + relevance A + relevance B documents) the system achieved 54.3 percent relevance ratios for the 50 searches and 56.4 percent for the 10 searches; 53.8 percent recall in those 10 searches was achieved. Indexing error (failure to index a concept in the text) accounted for 28.3 percent of all failures to retrieve relevant items; 43.5 percent of such failures are attributed to searching errors (lack of perseverance or ingenuity); 28.2 percent were due to system difficulties, particularly to differences in role indicators as selected by indexer and searcher. Although no definite conclusion as to the value of role indicators can be drawn, they should be studied and possibly dropped or their use modified to reduce these errors.

Analysis of reasons for retrieval of non-relevant documents showed that searching errors were the greatest cause of "noise", particularly the failure to select crucial terms or concepts for search. Approximately 98 percent of the unwanted items were retrieved through searches too generic or which omitted important concepts.

The Project SHARP system has sufficient discriminatory power and can retrieve a small subset of high relevance documents with little noise. Recall might be improved through greater depth of indexing. Further attention needs to be directed to coding, depth of indexing, and optimum application of role indicators.

II-31. Evaluation of Coordinate Index System During File Development, by Donald W. King, Journal of Chemical Documentation 5, 96-99 (May 1965).

There are three periods in the evolution of an information retrieval system in which evaluation may be helpful: the preliminary phase, the developmental phase, and the operational phase. During the developmental phase critical decisions are required and the system can best be modified and corrected. The principal objectives of evaluation during file development are to provide management with an effective tool for making decisions, to serve as a guide for using the system optimally, and to provide a control to insure that the system will perform satisfactorily. Decisions must be made concerning alternative system software, system hardware, and techniques for implementing the system. The major decisions include selection of information sources to be included; selection of information concepts, specific term descriptions, and structure of the term list, including the depth and breadth of indexing and inclusion of links, roles, etc.; selection of search procedures, including possible use of several screens and alternative search strategies; selection of document retrieval or presentation devices; selection of the optimum means of indexing the file; selection of optimum search strategies; and establishing a means for updating and maintaining the file.

Most management decisions regarding file and system development require information on cost, system reliability and time. The reliability may be described in terms of the system's ability to retrieve the documents required by the user. The measure of the missed relevant documents should be evaluated. Errors in output can be attributed to input errors, to inadequacy of search, or to an interaction of the two. If the number of missed documents is excessive, a means must be found to minimize and control the amount of error. System costs can be subdivided into the cost of file preparation, the cost of system equipment, and the cost of operation. The time required to prepare a file is important as a constraint on system design, since long delay in file preparation can reduce the effectiveness of a system. Also, lag time between request and retrieval is important, as is the total amount of time devoted to search, which is a function of production requirements, the overall system, search requirements, etc. These three measures (cost, reliability and time) are highly interrelated. The extent to which a change in one factor results in changes in the others is important. Equally important is the ability to estimate the effect of changes prior to effecting them.

To evaluate system performance, one must have a way of characterizing document retrieval. The retrieval profile provides such a device, where the set of file documents is classed into the cell entries of the two-by-two table, as retrieved/not retrieved and relevant/not relevant. The cell entries of the retrieval profile are a function of indexing accuracy (and consistency), the relationships among the terms, the number of terms used in a search query and the total number of documents in the file. One means of estimating the cell entries of the profile prior to indexing the entire file is by use of a mathematical

model constructed as a function of the variables just mentioned. The parameters of these models can be estimated during the developmental phase; and their validity for individual files can be checked by experimentation.

The evaluation program outlined here is currently being used in the development of the heterocyclic and magnetic core files of the U.S. Patent Office. Three phases of this program are a document analysis and indexing experiment, a document search experiment, and an indexing quality control program. The indexing experiment consists of two parts. The first involves re-indexing by various indexing modes (a single indexer, a single indexer reviewed, two independent indexers) a sample of 50 documents chosen randomly from the entire file. In addition, the documents are reconciled to formulate a set of "true" or "exact" indexed documents. The second part of the indexing experiment involves indexing an additional 250-500 documents also chosen randomly from the entire file, and indexed under normal production procedures. Values for the conditional indexing probabilities and indexing time (cost) are measured for the various indexing modes in the first 50 documents. This information alone may not be sufficient to decide which mode is optimum or indeed to decide if any or all are acceptable. For this reason, cell entries for the retrieval profile are estimated using the mathematical retrieval models and estimates of their parameters. The 250-500 documents are used to test the validity of the model.

The search experiment requires formulating a number of representative mechanized search queries, say 50-100. In certain cases synthetic questions can be asked. The validity of the models is tested by comparing the model estimates of the total retrieval, false drops and relevant documents retrieved with observed values of these variables estimated by searching the indexed sample of 250-500 documents and the documents in the indexing experiments which were indexed using a comparable indexer mode. The mathematical models provide good estimates for the cell entries of the retrieval profile. The decision concerning the optimum indexer mode is made easier with information from the indexing and searching experiment. With a given number of documents to be indexed, a given number of searches to be conducted per year, and some rough indication of the relative importance of cell entries in the retrieval profile one can determine the optimal indexer mode.

The two previous experiments should yield information concerning an acceptable level of indexing accuracy. The quality control program is designed to insure management that this acceptable level of indexing accuracy is maintained throughout the preparation of the file. The consistency coefficient is found by averaging the term selections by both of two indexers divided by the selections of either of them. The value of the probability that a term is selected given that it should be (p_3) can be found from the consistency coefficient using linear regression estimation. The consistency coefficient is more readily measured since it need merely be observed from two independent indexing operations, while the indexing must be judged to provide an estimate of p_3. Quality control acceptance tables can be constructed for various tolerance values for the consistency coefficient, various sample sizes and numbers of independent indexers. The entire area of quality control should be particularly important in the field of information retrieval, since retrieval errors are highly sensitive to indexing errors.

II-32. An Evaluation of Automatically Prepared Abstracts and Indexes, by William John Kurmey, Dissertation presented to Graduate Library School, University of Chicago, Chicago, Ill., University of Chicago, June 1964, 64 p.

The basic hypothesis for this study is that the standard of quality of automatically produced abstracts and indexes must meet the standards of manually produced abstracts and indexes, standards dictated by editorial policy and customer acceptance at minimum cost of preparation. No absolute scale exists for measuring the quality of abstracts and indexes, so the technique of establishing a recognized reference standard and measuring relative quality has been applied. Chemical Abstracts (CA) is acknowledged as outstanding in the field of chemistry and was adopted as the standard. With it as the base, the degree of conformity and difference of the automatically produced abstracts and indexes may be ascertained and their effectiveness evaluated.

Production of machine abstracts is based on counting the words in the text of the original article. On the basis of statistically measuring the number of words occurring frequently in an article and their proximity within a sentence, a numeric value is computed for each sentence. The sentences are ranked and a specified number with the highest values are selected to form the "extract" of the article. Twelve statistical methods of producing such abstracts were applied. Thus for the 50 articles selected for the sample from CA for the years 1951-1955, 600 extracts were created for comparison with the human abstracts. A byproduct of the production is a list of frequency of occurrence of words in the article. A machine generated index was obtained by using each word occurring more than a specified number of times in an article as an index entry. The adequacy and quality of these entries were then compared with the manually assigned subject entries for the same article.

Criteria suggested for an adequate abstract include that the content should show purpose, method, results, conclusions, and specialized content as required; that the abstract be brief and nonrepetitive, and that the form have clarity of content and conciseness of expression; that informative abstracts be in active voice, past tense and discuss the research while indicative abstracts have passive voice, present tense and discuss the article describing the research. Autoabstracts fall into the category of informative abstracts. If the introduction, method, results and conclusion sections are well written and concise then it may be expected that such sentences would be selected for the autoabstract. But no consistency in the choice of sentences from these sections occurred in the sample studied. Also for the treatment of form and content, no consistent results are obtained. A notable feature of the autoabstract is the absence of overall organization, as the abstract sentences often lack the context of preceding discussion. Most exacting control is exercised over the size of the autoabstract. The number of sentences to be selected was controlled so that the same number as appeared in the human abstract was selected. CA terms that were recorded for each abstract are considered to be the complete set of terms assigned to each abstract.

Comparison of the autoabstracts and human abstracts was carried out manually. Each sentence in the human abstract was analyzed into component concepts consisting of a subject, verb and object. The same process was carried out on the autoabstracts. The two were compared and equivalent concepts were counted. Results were considered as percentage agreement and percentage nonagreement between abstracts. Normalized percentages are calculated on the basis of percentage equivalents + percentage non-equivalents = 100 percent. Normalized percentages compensate for one concept in one abstract matching more than one concept in the other abstract.

Comparison of the index words was carried out manually. First the CA entries were broken into single word entries. Then the entire word list for an article was scanned to see if the word used by CA was in the text of the article. The agreement between the CA words and the word list was taken on a straight percentage basis. Next the number of words used by CA was used as a cutoff to obtain the words with the highest frequencies. Agreement was also taken on a straight percentage basis.

Normalized percentages of agreement for the 50 articles in the sample by each of the 12 methods of abstracting, average percentage agreement over the entire sample for each method, and average percentage of agreement over each document by all 12 methods are given. Percentage of agreement of words in the article and index entries (analyzed into single words) is also shown. The average overall conformity between word list and CA entries is 81.76 percent. The average overall conformity between the subset of words of highest frequency and CA entries is 27.63 percent.

The performance of the 12 abstracting methods varies considerably, from 0 percent to 100 percent agreement as to number of concepts included in the CA abstract. Cases with 100 percent agreement occurred where the human abstract consisted of sentences lifted directly from the article and the autoabstract selected the same sentences. Cases with 0 percent agreement occurred where the human abstract was a general summary and the autoabstract gave details of the research, or vice versa. The use of the best of the abstracting methods provides abstracts containing at most 37 percent of the concepts included by CA in abstracting the same article. It appears that, in the application of statistical word frequency techniques, equivalent results are obtained from either complex chaining techniques or by a simple gross counting method. The application of such techniques to abstracting chemical literature does not appear to be justified at this time. It is apparent that maximum-depth indexing would cover most of the entries used in the CA indexes for the articles. However it is not justified to select a subset of the most frequently occurring words to determine the index entries for an article. Apart from constructing maximum depth indexes, there appears to be no straightforward statistical method of arriving at index entries derived from a word frequency model of text. Possible improvement in the indexing entries may be obtained by utilizing a thesaurus applicable to the field of chemistry.

Conclusions as to performance of the abstracting and indexing methods are based on acceptance of CA as the high quality reference base and are valid only for the field of chemistry.

II-33. Indexing Costs for 10,000 Documents, by L. H. Linder, in Automation and Scientific Communication, Short Papers, Part 2 (papers contributed to the 26th Annual Meeting of the American Documentation Institute, Chicago, Ill., October 1963), Ed. by H. P. Luhn, Washington, D.C., American Documentation Institute, 1963, p. 147-148.

Costs have been compiled for input processing of report literature for a Termatrex-equipped coordinate indexing system; they are based on 10,000 items indexed with an average of 9.2 subject descriptors and 3.4 additional access points (corporate author, accession number, series or project name). Equipment costs, including office furniture and machines and the Termatrex Units, total $2,800; amortization is computed at $771.15, regarded as part of the cost of indexing these reports. Supplies, such as Radex cards and standard file cards, total $1,066. Labor charges including fringe benefits, for 8,005 hours actually worked, total $28,107. The total cost for indexing the 10,000 reports is thus $29,944.15; the cost per report averages $2.99. By standardizing the indexing language (2,600 descriptors) and using an existing vocabulary (the Thesaurus of ASTIA Descriptors, second edition, only slightly modified), the costs have been kept to a minimum.

II-34. Preliminary Studies for Machine Generated Index Vocabularies, by A. R. Meetham, Language and Speech 6, 22-36 (January-March 1963).

In the National Physical Laboratory program of research on information retrieval, the computer is to use text to generate an "index vocabulary" consisting of "descriptors" each of which is a group of words or stems in natural language. Then it will index the text in terms of the descriptors, for study of actual retrieval problems. Independently of the main project, the questions posed by such a program have been studied in relation to a short text with a limited and precise technical vocabulary. Each proposition of Euclid as published in Hall and Stevens, A School Geometry, Parts I-VI (MacMillan 1903) was treated as a separate document; in all there were 156 documents with a total technical vocabulary of only 146 words. The vocabulary of descriptors was generated by first finding all possible pairs of words with high rate of co-occurrence within the same document, then amalgamating these "significant word-pairs" into groups of 3-8 words. For the further breakdown of the two largest groups the word co-occurrences were expressed in the form of a matrix and two different strategies used alternatively. The two descriptor systems found by the two strategies were called the A-system and B-system. Both the descriptor systems were used with a threshold of 2 (at least two words from a descriptor must be present in a document for a descriptor to form part of the index of that document). The propositions fell into 107 index categories in the A-system and 112 index categories in the B-system. Inquiries were encoded in the descriptor language using a threshold of 1 (if one member of a descriptor was present in an inquiry, the documents indexed by that descriptor were assumed of possible relevance).

One set of 156 inquiries was examined; it was generated by taking at random one fourth of the words of each proposition in turn. Direct measures of relevance and recall efficiency were compared with other methods of evaluating the descriptor systems. The object was to find any simple way of testing, at as early a stage as possible, how useful a particular descriptor, set of descriptors, or index was likely to be. The methods of evaluation were those applied to descriptors before any further work is done, including number of words in a descriptor (called method 1), total number of occurrences of all the words of a descriptor (method 2), total of 1's in the word-pair matrix (method 3) and total of numbers in the second matrix of a descriptor (method 4); those applied to descriptors and indexes when the latter are complete, including number of documents for which a descriptor is applicable in relation to the total number of documents (method 5), and the "information content" of a descriptor with reference to the set of descriptors in the index of a set of documents (method 6); those applied only after fully indexing the documents and processing a suitable number of inquiries, including the number of times a given descriptor appears both in a successful answer and in the inquiry (method 7), the number of inquiries of more than average success in which the descriptor is included (method 8), and the number of times each descriptor occurs in both halves of a document (method 9); and those that can be applied when the documents in answer to each inquiry can be specified uniquely, namely the relevance and recall ratios of documents retrieved (method 10). Method 10 is an accepted method of evaluating complete retrieval systems; methods 1-8 were intended for assessing descriptors; method 9 was an attempt to introduce nearness of meaning; methods 6 and 9 are valuable for special reasons but the other seven methods were examined to find short cuts to the evaluations provided by methods 6, 9 and 10. The first nine methods were used to assess the relative values of the 20 descriptors in the B-system, and the agreement between sets of assessments was measured by correlation.

Conclusions drawn include that when there are choices of demarcation within a set of descriptors the most useful evaluation methods appear to be 5, 6 and 8; method 6 may be used to examine whether any descriptors are redundant; and when there are choices between sets of descriptors, method 10 is recommended, using the criterion that a document is relevant if it contains all the words of the inquiry. There is no proof that the methods will succeed when applied to a larger text with a larger vocabulary. Also the

mechanical tests of methods 8 and 10 are objective; it remains to be seen whether systems which are "good" by their tests are still good when used subjectively by actual inquirers. In addition, the evaluation methods 1-10 made no allowance for semantic associations and therefore if anything put too low a value on the retrieval system. It would be extravagant to claim that machines will perform tasks requiring some semantic understanding. What has been done here shows that by manipulation of words alone documents can be partially matched to inquiries, resulting in references to some of the wanted documents and not too many unwanted documents.

II-35. Automation of the Penn State University Acquisitions Department, by Thomas L. Minder and Gerald J. Lazorick, in Automation and Scientific Communication, Proceedings, Part 3 (papers presented at the 26th Annual Meeting of the American Documentation Institute, Chicago, Ill., October 1963), Ed. by Paul C. Janaske, Washington, D.C., American Documentation Institute, 1963, p. 455-460.

Prior to automating the operations of ordering and receiving library materials, a step-by-step analysis of each job within the Department was undertaken, including a diagram of the paper work and materials as they flow through the Department. This task brought out a number of defects such as bottlenecks, poor distinction between professional and clerical duties, and excessive checking procedures. The 50 essential operations worthy of a cost analysis were determined, and a two-week study was made of each, using random sample methods. Ordering costs are somewhat more than receiving costs; in summary, it costs $2.15 to purchase a book, $8.94 for a serial, and $6.69 for a periodical. The analysis of the flow study and the cost analysis supplied some basic tools for automation planning. A computer program for the entire processing to just short of the card catalog is being written and tested, to be followed by establishment of a prototype acquisition department. This will be evaluated before conversion to the permanent installation is started.

The study pointed up the need for standards in the field: standardization of computer languages on a profession-wide basis, standardized alphabetic symbols for a computer-based library system, standardization of acquisition input forms, and more realistic courses in library schools to train students in systems engineering and data processing techniques.

II-36. Machinelike Indexing by People, by Christine Montgomery and Don R. Swanson, American Documentation 13, 359-366 (October 1962).

A brief study of an existing system for indexing medical literature was carried out through detailed examination of its end product --- a monthly index of medical journals. This approach does not involve evaluation in the sense of measuring retrieval effectiveness, but rather is based on the question, "To what extent can the human indexing operations that take place in an existing system be simulated by machines?" The indexing entries studied were selected from the September 1960 issue of Index Medicus published by the National Library of Medicine. This indexing system is based on an alphabetic subject heading list whose 5,000 terms vary widely from the specific to the general. The subject portion consists of title citations cross-filed under each assigned subject heading and subheading. The process is perhaps better described as "assignment of articles to pre-established subject categories." Those words of article titles which are identical to --- or near synonyms of --- the subject heading (usually one word) under which the title appears could have been assigned to the proper subject headings by a machine procedure.

The objective of the study was re-stated by asking which subject headings within the Index Medicus are amenable to automatic assignment of titles. The study proceeded on the basis of examining entries in Index Medicus. The first sample of subject headings was random but excluded those with fewer than 30 title citations. Each title was inspected to determine whether it contained words identical to or essentially synonymous with the corresponding subject heading. Matches were found in 4, 093 of the 4, 770 entries studied, 149 were doubtful, and the remaining 528 did not contain matching words. A second sample of an equal number of subject headings was taken alphabetically (excluding those previously studied). Of the total of 451 entries examined, 392 contained words matching elements in the subject headings, 12 were unresolved, and 47 were judged difficult or impossible to index by an automatic process.

If a machine were carrying out this matching process, it would have to be provided with a thesaurus or dictionary containing groups of words which either are synonyms or which function as synonyms for purposes of information retrieval. The size of the synonym groups is a function of the inclusiveness of the individual subject heading rather than of the number of titles listed under it. The conclusion is drawn that if a computer were provided with the title of each of these articles, the Index Medicus subject heading list, and a synonym list or thesaurus group of each subject heading, it could then be programmed to determine which subject headings should be assigned to each article. About 86 percent of such assignments would be the same as assignments made by human indexing.

It is of interest to examine the 14 percent of the cases that failed to conform to these specifications. About half could not be resolved without more extensive research. The remaining titles fell into two groups, those which were inherently ambiguous and those which were ambiguous within the framework of the Index Medicus system. The subject headings vary in generality and many specific headings can be subsumed under more general ones. This ambiguity of overlap is only partially resolved by the cross-reference structure. In this study, articles containing a word identical to a specific heading but listed under the general heading were labeled as not indexable by an automatic procedure. There was also the reverse situation. Such articles were ambiguous with respect to their assignment to one or several potential subject headings. Presumably, reference to the full text is necessary in those cases.

A study similar to the above was made of a sample of 83 bibliographies compiled in the UCLA Biomedical Library in response to requests for information on a variety of topics. The 83 requests were first matched against the Index Medicus subject headings to determine whether comparable bibliographies could have been compiled from a search of listings under those titles in the Index. Only 16 of the 83 requests were identical to subject headings in the Index, while 59 were phrased in more specific terms than the corresponding headings. The remaining eight involved terms neither listed among the subject headings nor defined by the cross-reference structure. An alternative search procedure was then tested which involved direct matching of the words (expanded by lists of synonyms) used by the requestor to corresponding words of article titles in the bibliography compiled in response to the particular request, to establish the number of citations per request containing one or more of the concepts. Of the total of 3, 145 bibliographic citations evaluated, only 5. 3 percent did not contain any terms which corresponded to those concepts. This seems to indicate that the particular special bibliographies could have been compiled on the basis of title alone.

A project on text searching at Thompson Ramo Wooldridge Inc., (cf. II-53) included the development of a set of estimates of the relevance of each of 100 physics articles to each of 50 questions, based on direct examination of the full text of the collection by experts in the subject matter. A similar estimate has been made in which only titles were made available for relevance estimates. Comparison indicates that titles are only about

1/3 effective as a basis for estimating the relevance of the article to a given question. This title relevance study also included an analysis of why relevant information was judged irrelevant and why irrelevant information was judged relevant. In most cases of missed relevant information, the title did not adequately describe that portion of the article relevant to the question. It was observed that 80 percent of the successful retrievals could also have been performed by a machine through a process of matching words in the retrieval questions to words in the title. In those instances of successful retrieval in which no such word match was present, the requester's knowledge of physics permitted associations which could not be incorporated in a thesaurus or synonym dictionary. In about 1/2 of the instances of irrelevant associations, the article title contained one or two words in common with the search question, which made that article a reasonable candidate for being considered as possibly relevant to the question.

This study suggests that indexing should be based on more than titles and that a bibliographic citation system should present to the requester something more than titles.

Comments and questions on this paper are contained in Letters to the Editor, by Gary Carlson, American Documentation 14, 328-329 (October 1963). He comments on the general intent of the paper which seems to suggest that Index Medicus could be published in the form of a permuted title index. This leads to over-indexing, which has disadvantages. What is needed is a basis to determine the depth-of-indexing best suited to the needs of the users. Examination of seven subject headings which include 22 articles shows that these articles were indexed under a total of 33 entries. A permuted title index would have created 102 entries.

Few would question that a machine process for creating bibliographies by matching request terms with title terms would include wanted items, but the question is how much other material would also be retrieved by the machine? The original work at UCLA Biomedical Library was checked, including not only the finished bibliographies but also the worksheets used by the librarians to compile them. In one example, the reference librarian's worksheet originally had 43 citations, reduced to the 13 given in the final bibliography. If a machine routine were used as suggested, matching only single terms separately, 2,687 articles would result. Checking entries for joint 2-term occurrence together gives only 2 to 4 articles (depending on how synonyms are defined). Both results are in sharp contrast to the 13 or 43 arrived at by the reference librarian, and demonstrates that a simple machine processing of titles would give way too much or practically nothing.

The Montgomery-Swanson article points out the need for closer cooperation between computer people and librarians and for accurate manipulation of experimental data and clear statements of procedure, and that proposals to mechanize information retrieval should consider the problem of false drops.

II-37. The Metallurgical Searching Service of the American Society for Metals-Western Reserve University: An Evaluation, a report by an Ad Hoc Committee of the Office of Documentation, National Academy of Sciences-National Re search Council, Publication 1148, Washington, D.C., National Academy of Sciences-National Research Council, 1964, 96 p.

Early in the 1950's the Center for Documentation and Communication Research at Western Reserve University began development of a pilot mechanized searching system for the American Society for Metals. In 1960 the National Science Foundation made a grant for a large-scale test of the system's procedures, and requested the National Academy of Sciences-National Research Council to assist in planning of the test program and to evaluate its results. An Ad Hoc Committee was appointed, which began simultaneously exploratory studies to develop a framework for measuring the effectiveness of information retrieval systems, and evaluation of the Center's service itself. Specifically, the Committee recommended two operations research studies on criteria for evaluating information retrieval systems, one by Arthur Andersen and Co. (cf. IV-2) and the other by Stanford Research Institute (cf. IV-11). The Committee also arranged a series of parallel searches performed by different searching services competent in the field of metallurgy: the Research Information Service of the John Crerar Library, the Science Information Service of the Franklin Institute, the Office of Technical Services of the Department of Commerce, and a registered patent attorney with metallurgical competence. The Center picked ten current-awareness searches and six retrospective searches from its operations; the number of documents found by the Center, the number found by the parallel search, the number common to both searches, and the number of documents in the Center's files cited by the parallel searches but not by the Center, were tabulated. In addition, the number of documents judged not pertinent by a reviewer was noted. The Committee called attention to the surprisingly small number of common citations in all the searches.

The results of the ASLIB-Cranfield experiment involving the Center's system (cf. I-1) were also considered by the Committee in its final evaluation.

The Committee had a survey made of users' reactions to the Center's service (cf. II-56). The Bureau of Social Science Research Inc., conducted 111 questionnaires or interviews, in early 1962, out of a total population of 131 clients of the Center. The important characteristics to users are coverage, speed, and relevance --- coverage is by far the most important. About half the users found the Center's service satisfactory; seven out of eight said the Service covered sources in metallurgy very well; and almost two-thirds were confident that little of importance was missed by the Service. Three-fifths of the subscribers found that the service performed as fast as they had expected; 60 percent of the abstracts supplied were considered relevant. Another finding of the survey was that users who had opportunity to compare the Center's service with others found it better in only one-fourth of the comparisons.

During the tests, the personnel of the Center conducted a study of search strategies, to determine whether means could be developed to reduce the percentage of irrelevant answers and thus the amount of evaluation necessary on the output from a machine search. The procedure consisted of searching eleven questions against 3 percent of the file; after the first search, results were analyzed, the strategy changed, and the file re-searched. Repeated re-searching resulted in marked improvement --- a 50 percent increase in the ratio of relevant to irrelevant material by adjustment of association levels, a 33 percent improvement in the ratio by use of logical negation, somewhat less than 33 percent by providing terminal numerical suffixes and different semantic factors, and 20 percent improvement by use of role indicators. The Committee concluded that the Center does not always make full use of the sophistication of its system, raising a question as to the value of that sophistication.

The Committee recommended, in general, continued efforts to identify user requirements, a critical review of the Center's internal operation with the objective of improving its deficiencies, and that the Center account for its failure to utilize the sophistication of its system.

II-38. The Keyword-In-Context Index: An Evaluation, by A. Neelameghan, in Documentation Periodicals: Coverage. Arrangement. Scatter. Seepage. Compilation, Bangalore, India, Documentation Research and Training Center, 1963.

Chemical Titles, the American Chemical Society current awareness bibliography in the form of a permutation index, can be judged as to coverage and as to efficiency as a service. This latter criterion depends on speed in covering current articles and on the arrangement of entries. For the Hindustan Antibiotics Library the speed in publication was offset by the delay in transit, so their evaluation was of the criterion of arrangement only. Problems in control of terminology, alphabetic scatter of related subjects, synonyms, misleading titles, and the need for "preparation" in using the index (i.e., determining all possible keywords that must be searched for a given request) were examined. In a test case, 100 documents dealing with antitumor agents were selected at random from periodicals indexed in Chemical Titles. Terms (about 17 of them) likely to be in the titles of documents on the subject were checked, but under no one term could 25 percent of the documents be found. Thirty-three titles on the list did not contain any of the "most likely" terms and therefore were not retrieved. For a more restricted question, antibiotics with antitumor activity, involving 25 of the 100 test documents, 18 documents (72 percent) were found. For the personnel at this Library, depth classification of the subject of the document would serve their needs better than the permuted title index.

II-39. Searching Titles by Man, Machine, and Chance, by Vaun A. Newill and William Goffman, in Parameters of Information Science (Proceedings of the American Documentation Institute, 1964 Annual Meeting, Vol. 1, October 5-8, 1964, Philadelphia, Pa.), Washington, D.C., Spartan Books and London, Cleaver-Hume Press, 1964, p. 421-423.

This experiment is part of a study to determine the range of searching tasks which a machine can perform as effectively as a human. A set of 12 documents was selected from the literature which appeared in Diabetes during the year 1963. The compilation of the bibliographic citations to these 12 articles constituted the file of titles (210) which was to be searched. Each of the 12 titles constituted a query to which its associated set of bibliographic citations constituted an answer. An expert in the diabetic field selected the relevant set of documents to each of the 12 queries by judging the relevancy of each member of the file on a four-point scale with respect to each query. A computer also selected the relevant sets of titles by means of a keyword from title search procedure. Chance outputs were obtained based on a random search. Effectiveness was measured in accordance with the methodology developed previously (cf. III-9). The actual and mean values for specificity, sensitivity, and effectiveness were calculated. Results indicate that the expert is most effective; the computer is more effective than chance; titles are not very sensitive, since a random selection seems to provide as sensitive a response as the expert; the computer seems to be at least as specific as the expert, although not as sensitive; and the effectiveness measure seems to behave as predicted, since the mean effectiveness score for chance is close to zero.

No extensive interpretation of these results can be given because these are only the preliminary observations of the study.

II-40. Test Program for Evaluating Procedures for the Exploitation of Literature of Interest to Metallurgists, by LaVahn Overmyer, American Documentation 13, 210-222 (April 1962).

A detailed breakdown has been made of actual costs incurred during 1960 to provide the input for the mechanical searching system for metallurgical literature established at the Center for Documentation and Communication Research, Western Reserve University. Conventional and telegraphic abstracts were prepared for about 39,000 articles, of which 34,000 were completely processed and put on tape. Costs for acquiring material and readying it for abstracters were $0.64 per abstract; abstracting by 2 full-time and 50 part-time staff members cost an average of $1.86 per abstract, varying according to whether a full article or an abstract was examined. Editing and quality control cost $0.90 per abstract; clerical work of typing and filing cost $0.59 per abstract; and preparation of new codes for the semantic code dictionary cost $0.42 per abstract. The automatic encoding of the telegraphic abstracts, including keypunching, sorting, and card-to-tape conversion, cost $1.75 per abstract. Adding general managerial costs of about $0.34 per abstract, brought the total to $6.50 per abstract.

II-41. An Analysis of Output Costs and Procedures for an Operational Searching Service, by LaVahn Overmyer, American Documentation 14, 123-142 (April 1963).

At the American Society for Metals-Western Reserve University Center for Documentation and Communication Research mechanized searching service, data have been accumulated for the five output steps of analyzing and interpreting questions, structuring them into Boolean algebraic form, automatic encoding and searching on the computer, reviewing output, and transmitting answers. Combined costs (entirely personnel costs) for analyzing and structuring a question depend on its complexity; the range is $2.00 to $46.88 per question. Currently there are 40 questions searched for the biweekly current awareness service at a cost of $1.00 per question. Retrospective searches are handled upon receipt and costs depend on how many can be processed at once. The range is $160.00 to $276.67 if one question only is searched; the figures drop to $32.00 to $55.33 if five questions are handled simultaneously. Abstracts answering a question are scanned for their pertinency, and the pertinent or peripheral answers are further reviewed. Assuming a "not-sent" average of 25 percent, the cost of review (again entirely personnel costs) averages $7.81 per 100 abstracts. In summary, it is not possible to produce a single figure as the cost for a search by the ASM service; ranges of $7.90 to $53.07 for biweekly and $105.31 to $150.48 for retrospective searches can serve only as examples.

II-42. The Dollars and Cents of Basic Operations in Information Retrieval, by LaVahn Overmyer, in Information Retrieval in Action (papers presented at 1962 Conference at Center for Documentation and Communication Research), Cleveland, Ohio, The Press of Western Reserve University, 1963, p. 199-211.

Figures for input and output operation costs, reported in previous papers (cf. II-40 and II-41 respectively), have been brought up to date; comparison is made between the old and new figures. Operating steps remained essentially the same so changes reflect increased efficiencies and increased overhead, fringe benefits or salary charges. Where the input cost per abstract was $6.50 it is here computed at $6.49. Output costs are affected by a number of variables, including the degree of complexity of the question. Where ranges were given of $7.90 to $53.07 for biweekly searches and $105.31 to $150.48 for retrospective searches, the spread is now $7.90 to $52.52 and $105.31 to $149.93

respectively. Efforts are being made to reduce costs by running trial searches to test the structure of a question, and by reducing the need for review of the machine output.

II-43. Relative Effectiveness of Document Titles and Abstracts for Determining Relevance of Documents, by A. Resnick, Science 134, 1004-1005 (October 6, 1961).

The Selective Dissemination of Information system in operation at IBM's Advanced Systems Development Division notifies users of the availability of documents selected to be of possible interest to them; the notifications usually consist of title of article, name of author(s), abstract, source of document, and number of pages. For the test reported here, 400 documents, consisting of IBM reports and patent disclosures and Journal of Applied Physics articles, were separated into two groups on a random basis; notifications for one group contained title and abstract as usual, while those for the other group contained title only (no abstract). Results showed no significant differences, at the .05 level, in rate of ordering copies of desired documents or in rate of acceptance as relevant after reading the documents so ordered, when the titles or titles plus abstracts are used for notifications.

II-44. The Use of Diary and Interview Techniques in Evaluating a System for Disseminating Technical Information, by A. Resnick and C. B. Hensley, American Documentation 14, 109-116 (April 1963).

Two tools, the diary and the interview, were used to study aspects of the SDI 1 system and the environment in which it operated. The primary goal of this study was to determine the major effects of the system on the selected user population; a secondary goal was the gathering of information on the subjects and their general information environment; the third goal was the exploration of the usefulness of diary and interview techniques in this context.

In September 1959 the pilot system operation was started. Two experimental groups of 15 persons received two different kinds of notification of incoming documents. One group received the SDI notification card which contains the title, author, source, and an abstract of the document. The second group received a hard copy of the document in addition to the card. Both groups were asked to respond to the notices. Prior to their introduction to the SDI system the subjects were asked to keep diaries for a 2-week period of all documents they read of a professional nature. After the diary period, each of the subjects was interviewed regarding his reading and information-gathering habits. After completion of the diaries and interviews, the subjects were exposed to the system for a 10-week period. On the last 14 days of this period, the subjects were again asked to keep diaries of their professional readings. They were then interviewed again at the end of the second diary period.

From the amount of data collected it is possible to formulate descriptive (after-the-fact) hypotheses to explain apparent patterns in the data. The descriptive hypotheses presented are arranged according to the goals listed above; those of interest to this bibliography include: Hypothesis H1, that SDI, at an average processing rate of 23 documents per week, does not affect the amount of time users read or the number of reading acts; Hypothesis H2, that SDI reallocates reading time toward SDI documents and away from other sources of information; Hypothesis H3, that it is not the lack of time to read that hinders users from keeping abreast of their area of work, but rather the lack of time to find relevant documents; Hypothesis H4, that users prefer a two-stage system as opposed to a single-stage system; Hypothesis H5, that users do not want to place a limit on the number

of notifications received; and Hypothesis H6, that users want to be able to obtain personal hard copies of documents.

This study points to the difficulty of obtaining valid measurements in the areas investigated with the techniques presently available, due either to lack of understanding of how to use the techniques in this area or to some deficiencies in the techniques themselves. This study has led to the conclusion that it is the combination of the two.

II-45. The Consistency of Human Judgments of Relevance, by A. Resnick and T. R. Savage, American Documentation 15, 93-95 (April 1964).

Attempts are being made to evaluate information retrieval systems on the basis of subjective human judgment of the relevance of documents either to a specific task or to general interest. The consistency of these judgments was tested by having representative personnel evaluate for relevancy a random sample of documents, citations to the documents, abstracts of the documents, and a set of index terms for the documents; the subjects were asked to repeat their evaluations one month after the initial experiment. Results showed humans are consistent in their judgments, independent of the material on which the judgment is based, except in the case of abstracts; this anamoly remains unexplained. The test leaves unexplored the problem of a procedure for the quantification of the results of human judgment, i.e., for the provision of a scale on which such evaluations can be measured.

II-46. Final Report on the ROUT Document Retrieval System, by J. F. Rial, Tech. Doc. Report No. ESD-TDR-64-96, Report to U.S. Air Force, Air Force Systems Command, Bedford, Mass., The Mitre Corp., May 1964, 66 p.

ROUT (Retrieval of Unformatted Text) is a document retrieval system based on coordinate indexing and programmed for the IBM 1410. The document file consists of NORAD intelligence messages automatically indexed against a thesaurus of key words and phrases grouped by synonym and subordinate relationships. Three query processing methods may be used: the query may be processed without synonyms or subordinates (no expansion), the query may be expanded by adding to each key word its synonyms (synonym expansion), and the query may be expanded by adding both synonyms and subordinates (complete expansion). The querist may modify his query by deleting any or all of the synonyms and subordinates (delete process) and can request a printout of the texts or indexing keywords of the messages selected by the search (bibliography process). A document association technique can be used to extend any of the expansion processes, in which a query is treated by some combination of those processes to yield a set of messages. The five closest messages to each of these retrieved messages are obtained (metric search process) and the process repeated on selected messages retrieved by an extension of the process (metric search 2). ROUT can thus process any given query in $3(2^4) = 48$ distinct ways.

One of the challenges in designing a retrieval experiment is to choose a class of questions which represents the kind of question typically put to the system. In the case of ROUT there was no criterion for choosing representative questions so all kinds were selected. The work of making up 90 questions was divided among three people, who rated the messages as relevant or irrelevant to the questions. Each of the questions was translated into two queries by the person originating the question. The queries were treated by different combinations of the processes described above, yielding for a given query and combination a set of messages called a computed reply. Evaluation data were

collected under three conditions: (1) querying three samples of 100 messages each, one drawn at random and the other two chosen by picking a message at random and taking it plus the following 99 messages, in order to discover the effect of the uniqueness of the file structure; (2) splitting the message file into eight mutually exclusive sets each dealing with a particular area of the world, thus creating eight independent samples which could be combined in many ways; and (3) rating each message in the modified message file against a selected subset of the queries. The results of the three kinds of testing were compared and a judgment made on how best to evaluate a retrieval system operating on a larger document file.

Two basic types of errors generated by a document retrieval system are that the system can retrieve irrelevant documents (E_1) and that it can fail to retrieve relevant documents (E_2). The performance measures chosen for studying the errors are (1) the number of irrelevant messages retrieved per query = ω (direct measure of E_1 describing efficiency), (2) the fraction per query of relevant messages retrieved = ρ (the relevance ratio measuring utility and efficiency and an indirect measure of E_2), and (3) the fraction per query of retrieved messages that are relevant = τ (another measure of efficiency indirectly linking E_1 and E_2). Errors E_1 and E_2 were broken down for detailed study as: irrelevant messages retrieved because of bad synonym or subordinate relations (Class A), proper combination of query words in single topic message but message not relevant (Class B), or combination of query words from different parts of multitopic message (Class C); and relevant messages missed because of absence of one or more query words in message (Class D), misspelled query word(s) in message (Class E), or incomplete synonym group(s) (Class F). The primary reasons for retrieving irrelevant messages are in Class B and C; the primary reason for missing relevant messages in Class D. The errors are measured by simply counting, for each query with respect to a fixed query-processing method, the number of irrelevant messages retrieved or relevant messages missed that fall into each class. The means of the query-processing times for the three expansion-type operations were acquired and rough estimates given for additional processing time involved in each of the other operations. Two auxiliary performance measures derived were the probabilities of ω and ρ exceeding or not exceeding certain values.

The basic tools used to study ROUT were statistical. Analysis of variance was applied to some phases of the experiment. Means of performance measures were arranged in rectangular tables and the hypothesis that there is no significant difference between row (or column) means was examined at both the 1- and 5-percent significance levels. The Mann-Whitney test was used to re-examine some of the inconclusive evidence derived from analyses of variance and to study the probabilities that ω and ρ exceed or fail to exceed certain values. The categories were combined in many ways to test for anomalies and conclusions were drawn at the 1- and 5-percent significance levels. The Chi-square test was used to test for differences between the pairs of people who made up the queries and evaluated the messages. Frequency of occurrence histograms were created for some of the processes with respect to the three basic performance measures and some success was achieved in obtaining generalized curves for the measures ω and ρ. Fisher's Z-test for determining significant difference between variances was applied to the data acquired by random sampling. A formula was derived for computing the probability that a random selection of messages will yield a relevance ratio equal to, or greater than, a given value, and the performance of several query processing methods are compared to the performance of random selection as exhibited in plots of probabilities.

In discussing the results, Error A is seen to be less significant than the more inherent Error B, and Error D, along with D and F, is overwhelmingly more important than Error F by itself. Improvements in the thesaurus will have less influence on increasing the relevance ratio than on reducing the number of irrelevant messages retrieved. The undesirable effect of having multitopic messages in the file is worth emphasizing. The

multiplication of query words has the effect of reducing the number of irrelevant messages generated because of any specific expansion term. The use of single-term queries considerably increases ω for the no expansion process. The most important single fact that can be derived here is that there is substantial improvement in the relevance ratio as query expansion capability is added. The no expansion process appears to be hopelessly inadequate; on the other hand, the high value of ω for the complete expansion process dictates need for means to reduce the bulk of irrelevant material retrieved. Some caution is needed in constructing a category-type evaluation of a retrieval system. Categorization of a message file is a valid technique for studying relevance ratio provided that the number of queries in each query factor group is more or less uniformly divided among the categories used. It is indicated by the data that ω cannot be measured by categorization. There is some doubt cast on the value of ω as a performance measure, although it appears to be a desirable one to use in describing the characteristics of a retrieval system. The measure τ is better behaved and is, perhaps, just as useful in indicating the bulk of irrelevant material retrieved. The frequency histograms exhibit much regularity in the improvement of ρ as expansion capability is added, but no such regularity in the change of the distribution of the values of ω with the addition of expansion capability. The conclusion is drawn that the subordinate expansion has a much greater effect on ω than on ρ. The importance of studying the distributions of the quantities used to measure the performance of a retrieval system can be seen. With respect to the complete expansion process, for example, the probability of achieving a relevance ratio exceeding 0.9 for a given query is greater than the probability of obtaining a relevance ratio of less than 0.5 for the same query. This is a significant fact about the performance of the system.

Several different effects were studied in the sampling phase of the evaluation. The power of the bibliography process in reducing ω was established, along with the relative weakness of the delete process. The most interesting aspect of the sampling study is the similarity of many of the results to those of the category testing. The variations due to sampling in different ways is similar to the variations due to categorization, except that the reasons for the categorization effects are known and the reasons for the sampling effects are not. This is a matter which should be studied before the validity of sampling as applied to the evaluation of retrieval systems is assumed.

The results of the last phase of the evaluation can be accepted with the most confidence. The only new process studied in this phase are the metric search and metric search ^{2} processes. It is seen that the metric search is a fair substitute for the subordinate expansion. Metric search ^{2} provides no significant increase in capability. The metric process boosted the relevance ratio considerably and also increased ω considerably, but there was much repeating of the messages retrieved. Metric searching must be studied under more reasonable conditions before statements can be made about its effect on retrieving irrelevant messages.

Basic findings of the ROUT evaluation can be summarized as: (1) categorization is superior to sampling in the measurement of both retrieval errors and performance measures; (2) the number of irrelevant messages retrieved is not a satisfactory performance measure, and the bulk of irrelevant material retrieved is best measured by the fraction of retrieved material that is relevant; (3) the impact of the inherent syntactical defects of coordinate indexing, which produce the most significant errors, is mitigated by a thesaurus of synonyms and subordinates which provides an increase in basic performance; (4) use of subordinate groupings which increase performance can be replaced by a generalized mathematical association technique without loss of performance; (5) deleting key words from query word thesaurus groupings is an inefficient and undesirable way of reducing the bulk of irrelevant material retrieved; (6) queries with a few multiplicative factors should be used in preference to complex queries; and (7) mathematical association techniques provide a sound starting point for constructing an improved replacement for the whole coordinate

indexing process. The chief conclusion of the report is stated as: a generally acceptable solution to the document retrieval problem must be achieved by either vastly extending the power of the basic coordinate indexing process or by replacing this process by an altogether different one.

II-47. Performance Indices for Document Retrieval Systems, by Joseph Rocchio, in Information Storage and Retrieval, Scientific Report No. ISR-8 to National Science Foundation, Cambridge, Mass., Harvard University, Computation Laboratory, December 1964, p. III-1 - III-18.

A generalized model of SMART, a document retrieval system employing fully automatic indexing, may be characterized by a set of reference documents in the natural language (D); a set of retrieval requests in the natural language (Q); an index language (L); transformations T and T′ (possibly the same) from the natural language to the index language, which operate on D and Q respectively; and a search function S, basically characterized by a correlation process involving a request image and the set of reference document images. A retrieval operation consists in applying the search function S to the cardinal product $T(D) \times T(q)$, $q \in Q$. (q is a member of the set Q.)

To evaluate the effectiveness of a retrieval operation, let us assign to each request q a subset D_q of D which is the set of documents "relevant" to q. Alternatively one may consider that corresponding to each request q, an ordering of D is defined which reflects the "degree of relevance" of each document to the request. The object of the system is to produce a subset D′ which is identical to D_q. Evaluation then requires determination of how each of the system elements affects the degree to which this objective is met for all members of Q. The most commonly used performance indices of document retrieval systems are the recall and relevance ratios introduced by the ASLIB-Cranfield project (cf. I-2). It should be noted that the joint behavior of these parameters is required to judge performance intuitively. It must also be noted that the decision function C of the model is required to specify the retrieved subset D_a. This is undesirable because it introduces an additional variable into the system. This suggests that one should deal directly with the search function S. Another justification for so doing is the fact that the decision function is determined subjectively, in the sense that the needs of a particular user dictate its characteristics. In practice, the search function can be used to induce a partial ordering on D; a user could then request that the results of the retrieval be presented in this induced order. In view of these considerations a set of performance indices has been developed which may be applied directly to the ordering induced on D by a retrieval search S.

The objective of a retrieval operation may be recast in the following form: A retrieval operation with respect to a request q is expected to produce an ordering on the reference collection D, such that every member of the set D_q is ranked above all members of the complement of D_q with respect to $D(\bar{D}_q)$. (No emphasis is placed on any relative order among the members of D_q.) Two functions of the ordering induced on D may be defined which are related to the recall and relevance (precision) of Cranfield. The function r(i) is viewed as the number of relevant documents having rank order less than or equal to i (the rank index induced on D) divided by the total number of relevant documents; it is Cleverdon's recall as a function of the order induced on D by a retrieval operation. The function p(i) is the number of relevant documents having rank order less than or equal to i divided by i. For each query, $r^*_q(i)$ and $p^*_q(i)$ define a desired recall function and a desired precision function. These functions are strictly defined only for discrete values of the rank index i. The extension to functions of a real variable is accomplished by defining two functions r*(x) and p*(x) such that r*(x) = r*(i) and p*(x) = p*(i) for x = i, i = 1, 2, ..., N.

At this point recall and precision functions may be defined for the results of a retrieval operation with respect to a particular query. A recall error and a precision error may be defined and their integrals computed. Then recall error = $\bar{O} - \frac{n_o + 1}{2}$, i.e., the integral of the difference between the recall figure for perfect retrieval and the recall function which results for an actual retrieval is the difference between the average rank $\bar{O}$ induced on the members of the set of relevant documents D_q by the retrieval operation, and the mean of the ranks which a perfect retrieval would induce. Maximum recall error = $N - n_o$; therefore, $\frac{\bar{O} - \frac{(n_o + 1)}{2}}{N - n_o}$ is a normalized index of over-all recall error.

If we reverse it, then $1 - \left[\frac{\bar{O} - \frac{(n_o + 1)}{2}}{N - n_o} \right]$ is the index of recall performance.

Again, an index of over-all precision performance is obtained by considering the reverse of the normalized index of precision error, i.e., $1 - \frac{\ln \prod_{i=1}^{n_o} O(i) - \ln n_o !}{\ln \left(\frac{N}{n_o} \right)}$

The difference between these two over-all measures lies in the fact that the recall index weights rank order uniformly, while the precision index weights initial ranks more strongly.

A method for defining document rank is required in the case where a partial order rather than a full order is induced on D by S. The objective of a retrieval operation would need to be redefined to take account of this partial ordering. The development of the indices for this case is more cumbersome than the case previously considered and is not presented here.

These indices have been used to evaluate the results of a variety of experiments conducted with the SMART system. In practice one is forced to expand the scale of the recall index; it has been defined as $1.0 - .5(1.0-x)$, where x = the normalized recall index, producing a range for the recall index similar to that of the precision index. Two related performance indices may be derived from the two considered above, which are useful in the case where a particular query is subjected to a set of retrieval operations to be compared. Their advantage lies in the fact that they are simpler and easier to compute, and the rank recall takes on a wider range for the results which have been observed. Their disadvantage is their dependence on n_o, the number of relevant documents for the query in question. This dependence limits the usefulness of these indices.

II-48. Test Design and Detailed Retrieval Results, by Joseph Rocchio and Margaret Engel, in Information Storage and Retrieval, Scientific Report No. ISR-8 to National Science Foundation, Cambridge, Mass., Harvard University, Computation Laboratory, December 1964, p. X-1 - X-25.

The SMART system operates as a document retrieval system in which requests are matched against a reference collection of 405 abstracts taken from the IRE Transactions on Electronic Computers (March, June, Sept. 1959). The semantic dictionaries are in part tailored to this collection so that most content-bearing words in the abstracts are included in the dictionary. Several important features of the thesaurus transformation, originally constructed empirically, are summarized: The mapping is many-to-many from word stems to concepts; variants of a given stem may be entered into the stem dictionary, if warranted by semantic considerations; the thesaurus is designed so that the sum of the weights of the images of an input term is constant, so as to give more weight to unambiguous concepts; and concepts may be images of phrases as well as of terms, the phrases being defined by statistical or syntactic phrase-detection techniques.

An initial set of retrieval requests, not tied to any particular reference documents, was generated by members of the research group, who were to some extent familiar with the collection and the system. Relevance judgments were provided by the originators and resulted from a full search of the total collection. The tests to date center around an investigation of the influence of the thesaurus, of phrase identification, and of the structure of the text image as well as the effect of modifying search requests with the concept hierarchy. A group of 17 retrieval requests was used under a total of 10 conditions, including concept classes derived from word stems on a one-to-one basis, thesaurus derived concept classes alone and with phrase detection for queries only and for both requests and documents, and thesaurus derived concept classes with request "altered" by terms from the concept hierarchy. Query or document images are numerical (weighted) concept vectors unless otherwise noted. A detailed summary of the results for each query and the processing schemes is given, as are the normalized evaluation parameters over specific requests, general requests, and all requests.

To investigate the performance which might be obtained by using pairs or triples of different analysis techniques for retrieval purposes, the retrieval lists for a given request over a number of processing methods are merged and this treated as a standard retrieval list. All of the evaluation measures developed for individual methods may be applied to the merged list.

II-49. The Evaluation of Automatic Retrieval Procedures-Selected Test Results Using the SMART System, by Gerard Salton, in Information Storage and Retrieval, Scientific Report No. ISR-8 to National Science Foundation, Cambridge, Mass., Harvard University, Computation Laboratory, December 1964, p. IV-1 - IV-36.

(The material in this report, and in the report by Joseph Rocchio (cf. II-47) is included, in less detailed form, in "The SMART Automatic Document Retrieval System-An Illustration", by Gerard Salton and M. E. Lesk, Communications of the ACM 8, 391-398 (June 1965).)

The evaluation of information retrieval systems has become of increasing importance in recent years, because more retrieval systems are being designed and because evaluation methods are of interest in themselves. This study differs from other reports on systems evaluation in that it deals with the evaluation of automatic information retrieval. SMART is a fully automatic system operating on the IBM 7094. Documents as well as requests are processed without any prior manual analysis. Facilities incorporated into

SMART include a system for separating words into stems and affixes, a thesaurus lookup system, a hierarchical arrangement of concepts in the thesaurus, statistical procedures to compute similarity coefficients based on co-occurrences of concepts, syntactic matching and statistical phrase matching methods, and a dictionary updating system. At the present time about 35 different processing options are available. The SMART system's organization makes it possible to evaluate the effectiveness of the various methods by comparing the outputs from a variety of different processing runs.

The data collection used consists of a set of 405 abstracts of documents in the computer literature published during 1959 in IRE Transactions on Electronic Computers. The results are based on about 20 search requests, each analyzed by 15 different indexing procedures. The search requests are separated into "general" (at least 10 relevant documents) and "specific" (less than 10 relevant documents). The user population consists of 10 technical people with background in the computer field. Documents found in answer to a request are listed in decreasing order of the correlation coefficient with the request. The average number of index terms used to identify each document is difficult to present and is at best an indication.

A large number of measures have been proposed for the evaluation of retrieval performance. Perhaps the best known of these are recall and precision. Measures such as these are particularly attractive for evaluating automatic retrieval procedures. Characteristics of this system important in this connection include that input errors due to faulty indexing or encoding are eliminated, conventional search errors are also excluded, errors cannot be introduced between original search request and final machine query, inconsistencies introduced by a large number of indexers cannot arise, and the role of human memory as a disturbance is eliminated.

In order to calculate the standard recall and precision measures, relevance judgments must be made by hand in order to decide for each document and each request whether the given document is relevant to the given request, the judgments are usually made so that a document is assumed either wholly relevant or wholly irrelevant, and a cutoff in the correlation between documents and requests is chosen so that documents whose correlation exceeds the cutoff value are retrieved while the others are not retrieved. The first task may require the performance of hundreds of thousands of human relevance judgments for document collections of reasonable size. Two solutions have been suggested: using sample techniques to isolate a suitable document subset and formulating search requests based on specific source documents included in the collection. Since the document collection used in these experiments is small enough to permit exhaustive determination of relevance, possible pitfalls were avoided to a great extent. The cutoff problem introduces a new variable extraneous to the task of measuring retrieval performance. In the SMART system a different cutoff would have to be picked for each of the many processing methods, if it were desired to retrieve about the same number of documents in each case. It was felt that the standard recall and precision measures should be redefined. A suitable criterion for these measures would be the set of rank-orders of the relevant documents, when these documents are arranged in decreasing correlation order.

The derivative for the normalized recall measure is based on the area difference (the integral) between an assumed ideal recall curve and the actual recall curve obtained by plotting the standard recall against the document rank. This area difference is not normalized and its maximum value may increase indefinitely with increasing size N of the document collection. To generate a normalized measure it is necessary to divide by N-n (where n = number of relevant documents) and subtract from 1 to furnish a measure ranging from 1 for perfect recall to 0 for the worst possible case:

$$R_{norm} = 1 - \frac{\sum_{i=1}^{n} r_i - \sum_{i=1}^{n} i}{n(N-n)} .$$

A similar derivation for the normalized precision measure results in the formula:

$$P_{norm} = 1 - \frac{\sum_{i=1}^{n} \ln r_i - \sum_{i=1}^{n} \ln i}{\ln \frac{N!}{n!\ (N-n)!}} .$$

If these measures are to be evaluated automatically as part of the retrieval process, it is necessary to introduce for each search request processed a list of the corresponding relevant document identifications. The requestor is asked to list those documents which he believes should be considered relevant to his request. Document ranks are used by the program to produce a variety of measures reflecting recall and precision, including simplified expressions called rank recall and log precision, defined respectively as:

$$\frac{\sum_{i=1}^{n} i}{\sum_{i=1}^{n} r_i} \quad \text{and} \quad \frac{\sum_{i=1}^{n} \ln i}{\sum_{i=1}^{n} \ln r_i} .$$

Two composite measures are also produced, the first being the sum of rank recall plus log precision, the other a weighted sum of the normalized measures, $1-5(R_{norm}) + P_{norm}$. The factor of 5 is so chosen as to give equal weight to the two component measures.

A number of conclusions become apparent when the recall and precision values are averaged over many different search requests: the measures for the various methods exhibit substantial differences, in proceeding from one method to another both measures vary in the same direction, all the measures are larger for the specific requests than for the general requests, use of the thesaurus seems more effective than use of the original words in document and requests, and the most effective procedures seem to be those which use combinations of concepts (phrases).

If full advantage is to be taken of the organization of the SMART system, search requests are best processed by several different methods and their outputs combined. Normalized recall and precision measures for combined methods are computed by automatically generating a combined rank list from the rank lists for the individual methods alone and averaging the resulting measures over several search requests.

In order to furnish some indication of systems performance comparable with previously published data, the standard recall and precision measures are computed. To generate these functions, threshold values which separate retrieved from not retrieved information are chosen, to produce a curve in the form introduced by Cleverdon (cf. I-2).

The data confirm that the statistical phrase run seems to give the best performance, and that both recall and precision measures are higher for specific requests than for general requests.

Since the study is based on manipulation of a small collection of documents and on few search requests and different processing methods, it is not possible to make claims of general validity. But it is believed that the data can be used as indications of the kind of performance to be expected of automatic retrieval systems. Of special interest are the facts that certain processing methods exhibit both high recall and high precision and the juxtaposition of a variety of processing methods provides improved performance over individual methods.

II-50. A Computer Analysis of the Merck Sharp and Dohme Research Laboratories Indexing System, by Claire K. Schultz and Clayton A. Shepherd, American Documentation 12, 83-92 (April 1961).

A sample of 9,868 references from the Merck Sharp and Dohme punched card system, consisting of references related to pharmaceutical research, was analyzed to determine size of the indexing vocabulary (the system's dictionary contained 644 accepted descriptors and 354 with notes indicating additional action; the sample used 629 unique descriptors); number of cross-references (567); average number of descriptors used per document (10); frequency of use of particular descriptors (if all had been used with equal frequency, they would have been used 153 times; actually 77 percent were used less than 150 times and 23 percent used very frequently); and frequency of combinations of descriptors (given as number of documents in which singles, pairs, and triplets occurred). These data are informative but not sufficient to evaluate fully an information system; more must be known of users' needs, questions asked, etc.

II-51. An Analysis of Factors Causing Irrelevant Answers to Machine Literature Searches and Proposed Solutions to the Problem, by Barbara B. Shaffer, in Parameters of Information Science (Proceedings of the American Documentation Institute, 1964 Annual Meeting, Vol. 1, October 5-8, 1964, Philadelphia, Pa.), Washington, D.C., Spartan Books and London, Cleaver-Hume Press, 1964, p. 433-436.

For every retrospective search conducted for the American Society for Metals Documentation Service during the period from September 1961 to January 1964 the date, number of responses, number of responses sent to subscribers and number rejected were noted. In this test the notation of objective relevance is ignored, and all answers sent to subscribers are regarded as good, all rejects as bad. Percent of total hits sent and percent rejected were compiled for each search; total number of hits for all searches and percent of responses sent were figured by the month and quarterly; searches were arranged in 10 percentiles according to percent of responses sent; a histogram was constructed showing number of answers sent or rejected quarterly; and percentiles for number rejected were shown in graph form. The searches to be analyzed in the final study were selected from the total population by the use of the Rand table of random numbers and were studied in random order. The number of responses sent and responses rejected to be studied for each search chosen for the sample was determined.

Observations regarding rejected search answers were recorded for each search. First, factors causing rejected answers were divided broadly into input, output, and system errors. Input errors fall into two groups, those associated with coding and those caused by machine indexing. A code may be too broad and will retrieve much unwanted

material. Improper use of levels, improper use of role indicators, omission of terms and ambiguities are error sources in the machine indexing area. Ouptut errors have two main subdivisions: logical errors (incorrect expression of relationships) and choice of unnecessarily broad search concepts. The third area causing rejected responses is the system itself: some concepts cannot be translated into system language on a one-to-one basis with the user request; abstracts are retrieved which contain the proper terms but do not match the interest of the subscriber; and the fact that all terms have equal weight causes retrieval of documents which mention the subject but do not discuss it.

The findings from the 28 searches studied have been tabulated three ways. The percentage of the total rejected responses caused by input procedures, output practice and system features are calculated. Then input and output are combined and a joint figure for the percentage of errors caused is calculated. In this second form, there is shown how much error is attributable to improper use of the system and how much to the system itself.

For all of the searches in the time period studied, the average percentage of answers sent out of total hits was only 23 percent. A great saving in man-hours and money will result if the number of responses can be lowered and the percentage sent can be boosted. Some practical solutions are: attack the unnecessarily broad search structure by using the system correctly on the input level and instilling confidence that retrieval terms on a level with user language will not miss good answers; solve difficulty in lack of one-to-one correlation between user and system expression of concepts by insertion of generic links; weighting of terms; and a thorough check of thesaural relationships. It is evident that poor use of the system has accounted for the great majority of rejected answers. Major effort needs to be expended in operation of the system, including education, rules to improve input consistency, and increased quality control.

II-52. Training a Computer to Assign Descriptors to Documents: Experiments in Automatic Indexing, by Mary Elizabeth Stevens and Genevie H. Urban, in AFIPS Conference Proceedings, Volume 25, 1964 Spring Joint Computer Conference, Baltimore, Md., Spartan Books, 1964, p. 563-575.

For small-scale experiments in automatic assignment indexing at the National Bureau of Standards, a sample was taken of approximately 100 items from a collection of about 10,000 reports in the fields of information retrieval and potentially related research, previously indexed by analysts at the Defense Documentation Center (formerly ASTIA) in 1960. By the method programmed for trial on the IBM 7090-7094 computers, each word from document title and abstract was associated with each of the descriptors previously assigned to that document, and the word-descriptor association lists merged into a master vocabulary list. Input material was processed against the master list to give for each word a "descriptor-selection-score" value for each of the descriptors previously associated with that word; the descriptor scores were summed and those with the highest scores were assigned to the new item.

In advance of evaluations based on retrieval tests, comparisons were made with the prior human indexing. Consistency-of-indexing criteria needed for assessment of machine as compared to manual indexing of the same items are difficult to establish. In one example, machine-assigned descriptors were compared with descriptors assigned to the same document by two different DDC analysts. The machine assigned 12 descriptors, none of which were entirely irrelevant; Indexer A assigned four descriptors; Indexer B assigned six, two of which were inappropriate. The machine method "hit" score was 50 percent with respect to Indexer A and only 33.3 percent with respect to Indexer B; at the same time, A's agreement with B was 33.3 percent, and B's with A was 50 percent. In the light of what little data are available on inter- or intra-indexer consistency, the automatic assignment indexing techniques appear promising.

In addition, investigation was made of relevance judgments by subject matter specialists reviewing the full text of the test items. Twenty-five of the test items were submitted to one or more NBS staff members, who were asked to choose 12 descriptors for each item from the list of descriptors available to the machine. The percent of machine-assigned descriptors judged by human analysts to be relevant, and machine-indexer and inter-indexer agreement when several people made independent evaluations, were computed. Taking both the "hit" accuracy and the human agreement data into account, the results compare favorably with those reported by Maron as 51.8 percent (see Maron, M. E., Automatic Indexing: An Experimental Inquiry, in Machine Indexing, Progress and Problems, Washington, D.C., American University, 1961, p. 236-265) and by Borko as 46.5 percent (see Borko, Harold, The Construction of an Empirically Based Mathematically Derived Classification System, in AFIPS Conference Proceedings, Volume 21, 1962 Spring Joint Computer Conference, Palo Alto, Calif., National Press, 1962, p. 279-289).

II-53. Searching Natural Language Text by Computer, by Don R. Swanson, Science 132, 1099-1104 (October 21, 1960).

(The material presented in this paper appears in fuller detail in "Word Correlation and Automatic Indexing. Phase I Final Report, An Experiment in Automatic Text Searching", Report No. C82-OU4, Canoga Park, Calif., Thompson Ramo Wooldridge Inc., April 30, 1960, 36 p. plus appendix.)

Formulating criteria for judging the relevance of any document to the need underlying a request for information lies at the heart of the problem of evaluating information systems. Since it is impractical for the requester to read all the literature and ascertain its relevance, the historical answer has been to organize and index the library, then search only a small amount of information. But a condensed representation is necessarily imperfect; in addition, it is impossible to catalog in advance to serve all needs. The work reported here involved the development of techniques for automatic full-text search and retrieval; such techniques should serve as a prerequisite for automatic indexing since techniques for the latter can be derived directly from a text-searching technique.

A collection of 100 articles in nuclear physics was selected from Physical Review covering a 10-year period. Physicists with experience in the subject specialties of the articles studied each one in the light of its possible relevance to each of 50 questions inspired by the articles. Degrees of relevance were represented by weighting factors, and all relevant responses to any subsequent retrieval question were thereby determined.

Two separate teams of physicists helped with the test: Group 1 determined the degrees of relevance of each article, as described above; Group 2, a total of 10 persons, transformed each question into search instructions for the computer. Three methods of organization (see below) were used, so about 1,000 retrieval experiments were carried out. The full text of each article was recorded on tape; the computer program tested each article in the collection to determine the presence of a word or phrase, or of a group of words or phrases. The collection was also cataloged by a subject heading index designed for the field of nuclear physics; test questions were transformed into appropriate index entries to search the catalog. Further, machine search was carried out with and without the aid of a thesaurus compiled for the experiment from a library that did not contain any of the search library documents. Thus the three methods referred to above were (1) conventional retrieval based on a subject-heading index; (2) retrieval based on specification of words or phrases in combinations (no aids); and (3) requests formulated as for method (2) but with a thesaurus-like list and its index as aids.

The retrieval score, which takes into account both the fraction of relevant material retrieved and the amount of accompanying irrelevant material, is given as R-pI, where R=sum of relevance weights of the retrieved documents divided by the total sum of relevance weights (for the given question) of all documents in the library; I = effective amount of irrelevant material (given by N-LR, where N = total number of documents retrieved and L = total number of relevant documents in the library); and p is the irrelevance penalty and may take on arbitrarily assigned values. A conspicuous implication of the results is that the proportion of relevant information retrieved under any circumstances is rather low. For no question did the average amount of relevant material retrieved exceed 42 percent of that which was judged to be present in the library. Another implication is the apparent superiority of machine retrieval over conventional retrieval in this experiment. In terms of "percent of relevant material retrieved averaged over all requesters and all questions", the conventional method showed 38 percent; unaided machine text search, 68 percent; and thesaurus-aided machine text search, 86 percent.

Much detailed study must be carried out to determine why relevant information was missed or irrelevant material retrieved; partial results show that (1) many failures were caused by "engineering" problems easily corrected; (2) some information marked as irrelevant turned out not to be irrelevant at all; (3) there was some co-occurrence of words not related in the manner intended by the requester (the substitution of proximity as a requirement rather than syntax is suggested); (4) problems arose of indirect implication or analogy, where a relevant article contained no words or phrases which would be designated relevant on the basis of the question alone; (5) there were unforseen contexts of single words (expansion of the thesaurus would probably help in this situation); and (6) near synonyms need to be listed in the thesaurus as an aid to retrieval of relevant material.

In summary, the effectiveness of all techniques tested was found to be low; text search by computer was better than the conventional method.

II-54. Effectiveness of a Pilot Information Service for Educational Research Materials, by Jean Tague, Cooperative Research Project No. 1743, Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, 1963, 56 p.

(Some of the material reported here also appeared in "Matching of Question and Answer Terminology in an Educational Research File", by Jean Tague, American Documentation 16, 26-32 (January 1965).)

This study investigates the effectiveness of a pilot information service for educational research in satisfying the needs of a particular user group and suggests procedures for system improvement. The file consisted of 4,837 documents published in the years 1952 to 1962, for which two types of abstracts were written. The conventional abstract gives bibliographic data and the significant contents of the article and is designed to serve as a guide to the questioner in deciding whether or not he wishes to consult the original article. The telegraphic abstract analyzes the subject content of a document in a form amenable to machine searching. Documents are also indexed by a faceted classification schedule. Both indexes may be searched by the computer for answers to an information request.

Two aspects of information retrieval were considered to be of greatest significance in evaluating the pilot system: determination of user needs and determination of the optimum method within the present system of satisfying these needs. User requirements were investigated by isolating the characteristics of documents judged relevant by the questioner, and retrieval strategies were compared on the basis of relevance and recall factors for each strategy.

A set of 24 questions was selected from the more than 400 submitted during the initial year of the project by educators and researchers. For comparative purposes the strategies used in programming each question were narrow and broad semantic code programs and faceted classification programs. The narrow semantic code programs (N programs) incorporated codes for the words of the question and words considered by the programmer to be related to the search. Codes were organized by role indicators, levels, and the connectives indicated by the logic of the question. Broader semantic code programs (B programs), employing fewer of the discriminatory features of the system, were written by using one or more of the following devices: eliminating role indicators, raising the level of required combinations, omitting one or more of the conjuncts of a conjunction, adding alternate codes as disjuncts, and omitting infixes. Programs using appropriate combinations from the faceted classification schedule were also written (C programs), using the concepts rather than the words of the question and being generally quite broad. Narrow and broad semantic code programs were run against the entire test file; faceted classification programs were searched against that portion of the file (543 documents) to which the faceted classifications had been assigned.

Answers retrieved by these programs were evaluated internally by staff members as relevant (has direct bearing on the question), peripheral (of possible or partial bearing on the question) or nonrelevant (has no bearing on the question). It was noticed that some of the relevant answers retrieved by the C programs had been missed even by the B programs. The latter were revised so that the missed answers would be retrieved, and these revised programs (BR programs) run against the total file. Answers rated relevant or peripheral by staff members were sent to the original questioners for evaluation, again as relevant, peripheral or nonrelevant. Evaluations were received from 14 of the 24 questioners. Their evaluations verified the conclusion that the C programs were retrieving a number of relevant answers missed by the semantic code programs. The faceted classification codes were applied to all the documents in the file, the entire file searched with the C programs, and those results evaluated. Relevancy and recall factors for all the programs were computed for each question, and the significance of the difference between programs was tested by the Wilcoxon matched-pairs test. Relevant answers not retrieved by the original N programs were examined to determine causes of non-retrieval, and non-relevant answers retrieved were examined to determine causes of retrieval. These evaluations were used in comparing the alternate retrieval strategies and in investigating the characteristics of relevant as opposed to peripheral and non-relevant documents.

It appears that the narrow semantic code program is most efficient in producing a high proportion of relevant documents among those which it retrieves. If one were guided by relevancy considerations, such a program would always be preferable. However, in considering the recall factor the situation is almost completely reversed. A less discriminatory system, the broad semantic code programs, will have a higher recall figure but at the cost of a decrease in relevance. If a question has been phrased in very specific terms, it appears that there are better results when the indexing vocabulary has been very strictly controlled as in the faceted classification. The mean relevance and recall factors are: N program, .35 mean relevance, .48 mean recall; B, .23 and .73; BR, .12 and .83; and C, .26 and .54 respectively. The C programs appear to lie between the high relevance and low recall of the N programs and the low relevance with high recall of the B programs. As to whether the differences between relevance and recall factors of the programs may be considered significant on the basis of a sample of 14 questions, the Wilcoxon matched-pairs signed rank test was applied to the difference scores between each set of relevancy and recall factors. The results are, for relevancy factors: N and B programs, no significant difference at .01 level; N and C programs, and B and C programs, no significant difference at .05 level. For recall factors, results are: N and B programs, significant difference at .01 level, i.e., B has greater recall factor; N and C programs, and B and C programs, no significant difference at .05 level. Apparently a larger sample is needed.

In an attempt to raise the recall factor for the N programs, the relevant answers not retrieved by the N programs were examined to determine how either the program or the abstract might have been altered so that these answers would have been retrieved. Some general causes of non-retrieval were (1) terminology: words in the telegraphic abstract relating to the inquiry were not predicted by the programmer from the question words, semantic code relationships, or intuitive knowledge of the file. The problem is more acute in the social sciences than in the physical sciences because of "soft" terminology. Modification of the semantic code to achieve more consistent terminological control is recommended; (2) role indicators: those specified in the N program are not the ones which appear in relevant telegraphic abstracts. The strictest control must be exercised in their application in the abstracts; (3) combinations: some relevant answers are missed because too many conjunctions of semantic code and role indicators are required. Many of these problems can be eliminated by a closer working relationship between questioner and programmer; (4) levels: answers may be missed because a combination of terms occurs in the telegraphic abstract at a level higher than that requested in the program. Increased control over terminology and role indicators would go far in eliminating this program; (5) abstracting errors; (6) infixes: a related term is coded with an infix different from that specified in the program. Terminological control appears to be the answer here, as well as care when writing programs that all infixes contained in relevant codes are specified; (7) programming errors; and (8) obscurity: since no contact was made with the questioner, some misunderstandings arose concerning the scope and meaning of the inquiry. The same procedure was followed in determining the causes of retrieval of nonrelevant answers by the N programs. General causes of retrieval of nonrelevant answers were (1) omission of information in the conventional abstract which the questioner used for evaluation: with the additional information in the telegraphic abstract the questioner might have considered the answer relevant. User tests should aim at this point and the extent to which the conventional abstract should be expanded to include all information indexed in the telegraphic abstract should be tested and considered; (2) point of view: abstract considers the question in a context not of interest to the questioner. In some cases, greater precision in defining the question would have eliminated much of the non-relevant material; (3) role indicators: not sufficiently selective to distinguish relevant and nonrelevant answers; (4) infixes: a universal rather than a specific infix was used with some semantic factors. Result of infix specification might be to eliminate one relevant answer while selecting many nonrelevant answers. Use of infixes depends to some extent on user needs, i.e., completeness vs. relevancy; (5) false combinations of terms: generally reduced by use of levels and role indicators; (6) semantic codes: when assigned to a group of terms are not sufficiently selective to make distinctions of interest to the questioner; (7) additional terminology added by the programmer: in some cases proved nonrelevant to the questioner; (8) minor reference to subject; (9) coding errors; and (10) abstracting errors.

It is recommended that the investigations be extended on a more discriminatory basis to a larger set of questions and to a user group with whom there is a maximum of personal contact and cooperation.

Objective criteria of relevance are usually based on a special situation, such as questions with predetermined answers or an intimate and detailed knowledge of user needs. It seemed preferable in this study to obtain subjective assignments of relevance by the questioner and then attempt to determine the objective properties of these relevant documents. In determining these objective criteria two questions were asked: (1) if answers are evaluated by two evaluators, how much will assignments of relevance vary; and (2) do the properties which characterize relevant documents lie in the language of the document as opposed to extra-linguistic factors and, if such linguistic properties can be isolated, to what extent are they reflected in the indexing-searching apparatus?

The sample contained a total of 1,214 abstracts, answers to the 14 questions previously evaluated at the Center as relevant or peripheral, evaluated by the original questioners as relevant, peripheral, or nonrelevant. Four questions of the 14 with 320 answers were given two outside evaluations. In deciding whether to send out peripheral answers to questioners, there are two possibilities for making a wrong decision: if the questioner considers the peripheral answer sent as nonrelevant, or if the questioner would have considered the peripheral answer not sent as relevant. One determination made in this study was whether the number of wrong decisions would have been greater if answers considered peripheral had not been sent.

By sending relevant and peripheral answers a wrong decision was made with 37 percent of the answers: 13 percent of the relevant and 24 percent of the peripheral were called nonrelevant by the questioner; if only the relevant had been sent, a wrong decision would have been made 19 percent of the time: the 13 percent of the relevant judged nonrelevant plus 6 percent of the peripheral considered relevant. With the four questions given two evaluations, by sending both types of answers a wrong decision was made in 36 percent of the answers: 8 percent of the relevant and 28 percent of the peripheral were rated nonrelevant by both questioners; had the peripheral answers not been sent, a wrong decision would have been made in 20 percent of the answers: 8 percent rated nonrelevant as above plus 12 percent of the peripheral rated relevant by at least one questioner. Applying the Wilcoxon matched-pair signed ranks test to the difference scores for the sample of 14 questions, there is no difference at the .05 level in percent wrong decisions made by sending vs. those made by not sending peripheral answers.

Results indicate that fewer wrong decisions are made when peripheral are not sent (19 percent as compared with 37 percent), but these figures must be interpreted in terms of relative cost to the customer of missing a relevant answer and receiving a nonrelevant answer. The evaluations by two external evaluators would indicate that there is disagreement, about the relevance of answers, amounting to 18 percent of the answers. In general, these results indicate that one cannot say an answer is relevant or nonrelevant in any absolute sense.

As a first hypothesis it was postulated that some linguistic property distinguishes the relevant answers, and that it is the relative frequency of the question words and words related to the question words in title, conventional abstract, and index (telegraphic abstract) to a document which characterize relevant answers. The frequency of the three types of words of search programs --- question words, semantic code references, and program additions --- in relevant, peripheral, and nonrelevant answers was computed and used to determine the effectiveness of each in locating relevant answers (i.e., as indexing terms) and in indicating the relevance of answers (i.e., as evaluating terms). The effectiveness of the types of words as indexing terms is rated by the ratio of number of relevant answers indexed by the types to total number of relevant answers, not a single ratio but a frequency function plotting number of one type against the observed probability of the number of that type occurring in relevant answers. If the mean number of words of one type is greater than the mean for another, then the former are said to be more effective as indexing terms. The relevancy factor indicates the evaluating effectiveness of a single term; if the graph for words of one type is everywhere above that for another type, then the former are more effective in evaluating.

The three types of words were matched against the words in the telegraphic abstracts, conventional abstracts, and titles of the 1,214 answers (474 relevant, 285 peripheral and 455 nonrelevant as determined by the questioners). The ratios or probabilities were computed for matches of each type of word in each type of answer, for the indexing effectiveness, and for evaluation effectiveness. Functions indicating indexing effectiveness and evaluating effectiveness of each type of word were obtained. The significance of

differences in distribution of the types of words in relevant vs. peripheral and peripheral vs. nonrelevant answers, and of each type of word in the different answers, was tested by Chi square.

The mean number of matches for each type indicates that question words are most effective as indexing terms, followed by program additions. The probability of occurrence of words of all types is greater for relevant answers than for peripheral or nonrelevant answers. At the .01 level the frequency of the three types of words was found to be greater in relevant than in peripheral or nonrelevant answers. It appears that relevant answers differ the most from peripheral and nonrelevant answers in total number of word matches rather than particular type of word match.

Some general conclusions drawn from the data include that question words though most useful in locating relevant answers cannot be employed exclusively if completeness is desired, since 14 percent of the relevant questions contained no question word; the greatest drop in percent relevant answers retrieved by words of all types is between combinations of two and three words, since 87 percent of the relevant answers are retrieved when at least two words are required and 60 percent when at least three words are required; however, the greatest drop in percent relevant answers retrieved by question words occurs between combinations of one and two words, since 86 percent of the relevant answers are retrieved if at least one word is required and 46 percent if two or more words are required. It is apparent that words in the conventional abstract are the best indexes to relevant answers and that title words have much lower indexing effectiveness. On the other hand, title words are most effective as evaluating terms. A Chi square test of the difference in distribution of words in relevant vs. peripheral and peripheral vs. nonrelevant at .01 level shows that for all three indexes --- titles, telegraphic abstracts, and conventional abstracts --- the probability of question words is greater in relevant than in peripheral or nonrelevant answers. It may be concluded that frequency of question words in titles, conventional abstracts, and telegraphic abstracts distinguishes relevant from nonrelevant answers.

It is suggested that the method of quantitative assessment of indexing and evaluating effectiveness of terminology be utilized in further studies.

II-55. Evaluation, in Automatic Abstracting, Report to Rome Air Development Center, Report No. C107-3U1, Canoga Park, Calif., Thompson Ramo Wooldridge, Inc., February 2, 1963, p. 47-51.

The purpose of this investigation and study was to develop techniques for the automatic abstracting of textual information, including development of external objective criteria to evaluate the different abstracting routines relative to each other. The assumption was made that as a minimum abstract content would to a large degree satisfy the function of indicative-type abstracts in the screening of documents. Computer routines for automatic abstracting were tested in a series of experimental cycles consisting of computer run, analysis of output and program correction. Conclusions of the study include that a definition of an abstract has been developed which leads to a more uniform target abstract prepared by humans and permits the creation of automatic abstracts based on machine-recognizable clues, and that a method of evaluation by judges' ratings of similarity was developed showing that the machine abstracts have a 66 percent degree of similarity with the target abstracts. Recommendation is made that further research and experimentation be conducted on evaluation methods in order to increase their speed, simplicity and discrimination.

A sample of 40 documents of the test library, not used in the abstracting experiments, was selected for evaluation. (The corpus consisted of 200 documents dealing with the chemistry of exotic fuels obtained from ASTIA by choosing documents whose length did not exceed 4,096 words, a condition imposed because of limited storage capacity in the computer memory.) The evaluation procedure was designed based on (1) the definition of target abstract, and (2) the rating of the degree of similarity between two abstracts. Random extracts were used as controls. Instructions had been developed to create abstracts satisfying the definition that a sequence of sentences of a document selected in text order is said to be an abstract of the document if it contains only eligible, non-redundant sentences and is coherent. Abstracts generated in accordance with these instructions serve as the target abstract, containing a fixed (25) percent of the sentences of the original document. During the preparation of these abstracts means was provided to record a random selection of sentences, and this randomly generated abstract, so defined, serves as the control in the evaluation procedure.

The machine abstract, target abstract, and random extract were typed in the same format. The raters judged the similarity between the target abstract and each of the other two, giving 4 points for complete similarity, 3 points for considerable similarity, down to 0 points for no similarity. By dividing the total score by the maximum possible a normalized score was then obtained. This was interpreted as the degree of similarity between the abstracts and was recorded as a percentage. Two raters were used, one familiar with the 40 documents and the other not associated with the project except for the evaluation test. The high consistency of the raters is shown by the small variation of the mean scores in three samples of 10 documents each. The scores of the two raters were averaged for a single similarity score. Thus for each document two calculations were made: the percent similarity between target and machine, and percent similarity between target and random abstracts. The documents were divided into four samples of 10 each, and for each sample the mean similarity rating as well as the standard deviation of the mean were computed. Overlap data (i.e., percent agreement between target and machine) for cumulative samples 1, 2, and 3 gives 44 percent agreement between them. The mean similarity rating between the target and machine abstracts is 66 percent; the standard deviation of the mean is 3 percent. The mean similarity between the target and the randomly generated abstract is 34 percent; the standard deviation of the mean is 3 percent.

Because no attempt was made to evaluate the utility of the target abstracts, the above findings do not evaluate the utility of the machine abstracts. A sentence-by-sentence analysis is valuable as a supplement to the similarity rating. There can also be represented the degree to which the machine process included the information of the target abstract. In addition the total of abstract-worthy sentences may be taken to represent that part of the machine abstract which conforms to the content aspect of the working definition of "abstract". The average of 84 percent seems to represent a highly promising achievement in the automation of the abstracting process.

II-56. A Survey of Users of the American Society for Metals-Western Reserve University Searching Service, by Ivor Wayne, Report No. BSSR: 352 to National Science Foundation, Washington, D.C., Bureau of Social Science Research, Inc., July 1962, 27 p. plus appendix.

A questionnaire-based survey of 111 users of the metallurgical literature machine searching operation revealed that 42 percent of the users rated the service satisfactory while the rest were nearly equally divided between excellent, adequate, fair and poor ratings; that coverage of material was ranked as the most important of performance characteristics, with speed second in importance; that most (7 out of 8) persons said the

coverage of sources in metallurgy was satisfactory or better and 3/5 of the subscribers found the service as fast as they had anticipated; that system characteristics of intermediate importance to users include ease of requesting a search, general quality of the abstracts, and the proportion of relevant abstracts over total received; that characteristics of minor importance include the number of abstracts supplied, their physical makeup, and factors which reduce the utility of relevant abstracts; that 3/5 of retrospective searchers and 46 percent of current awareness subscribers rated the cost of the service appropriate to its value; and that almost all subscribers to both types of searches were planning to use the service again.

SECTION THREE

DISCUSSIONS

The reports in this section of the bibliography are concerned with the problems and potential of testing and evaluation procedures rather than with the development of application of such procedures. The subject matter of the reports is, in one sense, less concrete than that of items in the previous sections. On the other hand, the questions raised and the points made here are fundamental to the programs and studies covered before.

The discussion papers included here can be divided into three types: those which review progress in the field, those which seek to clarify or define the concepts of the field, and those which offer suggestions as to studies or programs needed for the field. These three types are not, of course, mutually exclusive, for elements of review, definition and suggestion intermingle in all. But the primary emphasis in each case can be seen to fall into one of the three categories.

The review papers contribute to knowledge of the field by summarizing extensive studies and programs, and by pointing out similarities or differences in the approaches taken to the subject of evaluation of information systems. Often these reviews also point out problems still to be faced (as for example in III-2 and -4), including problems in specific areas such as in evaluation of automatic indexing methods (III-25). Other reviews of work will be found in several of the reports in Section IV, preceding and leading up to proposals for further work or differing approaches to the problems.

Papers which attempt to clarify or define concepts in the field are stimulated by ambiguities in usage of such basic terms as relevance (III-6 and -8, for example), models (III-3), and measurement (III-28). The need to define the problem first and to design tests very carefully, with appropriate controls, is pointed out in the papers here. Sometimes the authors have been actively working in the field; other times they are observers and students of the active programs.

The suggestions offered in the third group of papers deal, on the one hand, with general problems of design and conduct of evaluation studies; on the other hand, some papers discuss specific problems associated with testing or offer specific lists of criteria to be applied in testing. In the former category, of general discussions, appear five papers covering the purpose and scope of the Comparative Systems Laboratory at Western Reserve University (III-9, -10, -11, -19, and -24). In the other category, of specific papers, are included lists of criteria to be considered (III-15 and -20) and such specific problems as those of automatic indexing (III-21, -22, and -31, for example).

III-1. Evaluating the Effectiveness of Information Retrieval Systems, by Harold Borko, SDC Report No. SP-909/000/00, Santa Monica, Calif., System Development Corp., August 2, 1962, 7 p.

Suggested methods for evaluating the performance of a given retrieval system include determination of user satisfaction; comparison of one system with another, on the basis of number and relevancy of their responses to the same inquiries; and comparison with an ideal system. A system can be viewed as aproblem in sampling statistics: one must select from a store of documents that set which provides information relevant to a given search question, maximizing the amount of relevant information and minimizing the irrelevant. Cleverdon (cf. I-2), Swanson (cf. II-53), and Borko (cf. IV-9) have compared the retrieval effectiveness of different indexing systems. Hertz (cf. IV-2) and Bourne (cf. IV-11) have discussed criteria and procedures for evaluation and comparison of entire scientific information retrieval systems.

III-2. A Review of the Methodology of Information System Design, by Charles P. Bourne, in Information Systems Workshop: The Designer's Responsibility and His Methodology, (A publication of the American Documentation Institute), Washington, D.C., Spartan Books, 1962, p. 11-35.

One of the steps in the process of designing an information system is evaluation, which can be separated into performance evaluation and economic evaluation. The simplest performance measurements are single-criterion evaluations of specified parameters; the one used most commonly involves asking sample questions of one or more systems, then observing the volume of references furnished. A more effective test is to compare the performance of a system against some standard, such as a perfect or ideal system. Cleverdon (cf. I-2) has compared four indexing systems under controlled conditions; the basic performance measure was the fraction of searches that were successful in locating documents which had generated test questions. A secondary performance measure was the relevance of the documents produced by the search.

Swanson (cf. II-53) has compared machine and manual searching. Determination was made of all relevant responses, and their degree of relevance, in a collection of articles and search questions; the retrieval results were scored as to the fraction of relevant material retrieved and the amount of irrelevant material retrieved.

Borko (cf. IV-9) used similar criteria in studying the performance of machine indexing techniques. Mooers (cf. IV-27) has suggested an evaluation technique similar to Swanson's which purports to yield an absolute rather than a relative measure of performance. The technique involves studying a sample of file items from which queries are formulated, posing the queries to the file, and examining the results to see which of the sample items were retrieved.

Composite figures of merit evaluations can be derived by describing both the requirements and performance of a system quantitatively, measuring numerically the agreement between requirements and performance, and weighing this agreement according to the importance of the requirement. Summing these figures gives a single merit figure for the system's performance. Alternatively, a time-cost model may be used to describe the system's performance in terms of time or cost penalties to the user.

There is practical value in determining the economic benefits of information systems, even though it is a difficult task. Good cost accounting procedures may be used to make accurate estimates for different parts of a system, which could then be compared to costs for other types of systems or to standard costs for similar operations. Simulated operation

of a proposed system can be done on a computer to estimate the system's operating costs over wide variations of operating parameters. In comparing alternative systems, each can be analyzed and simulated to find areas where one system is economically more attractive than the others.

A simple cost-performance index, e.g., productivity and effectiveness ratios, may also be developed to provide evaluation on a cost basis.

In summary, it seems that only some crude measures are available for the practical evaluation of system performance. Much work needs to be done in this area to develop better tools. (There are 54 references, of which 20 are pertinent to the discussion of evaluation and are included in this bibliography.)

III-3. Measurement, Modeling, Experimentation, and Evaluation in Information Retrieval Systems, by Edward C. Bryant, in Information Retrieval Among Examining Patent Offices (Proceedings of Third Annual Meeting of the Committee for International Cooperation in Information Retrieval Among Examining Patent Offices, Vienna, September 1963), Ed. by H. Pfeffer, Baltimore, Md., Spartan Books, 1964, p. 198-207.

(The elements of the discussion in this paper, plus an extensive discussion of the statistical process, can be found in "Evaluation of Information Retrieval Systems in Patent Office Environments, Part I, Statistical Concepts", by Edward C. Bryant, Denver, Colo., Westat Research Analysts, Inc., February 1965, 37 p.)

One must attempt to measure the characteristics of an existing system before trying to change it, in order to have some criteria against which to measure the effect of changes. It is convenient to think of three areas of measurement: cost, quantity, and quality. The measurement of operating costs alone raises no major problems, but in determining "costs" in the larger, or decision-making, sense we would like to assign a cost to each type of incorrect decision which the system can produce. All errors tend to undermine confidence, and it seems reasonable to suppose that there is some minimum level of quality below which the system will become ineffective because of lack of confidence on the part of the public. Quantity of output is clearly related to both costs and quality.

Measurement of the existing system implies judgment concerning what characteristics to measure and their interrelationships. One approach to measurement being experimented with in the Patent Office is to describe the operations of the Office in terms of a mathematical model, and then to measure the parameters necessary to describe the system. The contemplated model views the examining system as a stochastic (probability) process in which there is a network of operations (or work stations) and probability laws which govern the flow of patent applications through the network. The model defines the system and specifies the information to be gathered to describe the characteristics of the particular system under study. The problem of what to measure is then determined by the model.

Experimentation contains two components: the generation of ideas which may prove fruitful and the careful testing of those ideas. In terms of the Patent Office environment, it is important to conduct experimentation under conditions as nearly like those in the real world of patent examination as possible. To reduce the personality effects of the participants in experiments, the statistical techniques used include randomization (use of a random process to assign experimental units to tests) and replication (repetition of the same test on different experimental units). The statistician assists with the solution of some of these experimental design problems.

In considering evaluation of the experimental results, we assume that the subject of the experimentation is some proposed modification of the system. We wish to draw conclusions concerning the entire patent office under normal operating conditions. Experimental results should either confirm or discredit the hypothesis under test and should lead either to implementation of changes or to further experimentation or both.

In conclusion, preliminary models may be constructed on the basis of initial observation and measurements; study of the models will reveal need for experimentation; evaluation of the experiments will lead to further experimentation or to modifications to the system; modifications to the system lead to further observations and possible changes in the model, and the cycle begins to repeat itself. The four operations are all components of the systems development process, and none can be ignored without serious damage to the entire effort.

III-4. Progress Toward Evaluation of Information Retrieval Systems, by Edward C. Bryant, in Information Retrieval Among Examining Patent Offices (Proceedings of Fourth Annual Meeting of the Committee for International Cooperation in Information Retrieval Among Examining Patent Offices, Washington, D.C., October 1964), Ed. by H. Pfeffer, Washington, D.C., Spartan Books, 1966, p. 362-377.

The persistence of problems concerning evaluation of information retrieval systems seems to be related to such difficulties as lack of well-defined objectives and of meaningful models, uncertainty as to measurement criteria and relevancy, problems of scale and of design, and the like. There has been little progress toward development of retrieval models upon which experimentation can be based, models which define the relationships among system components. Giuliano and Jones have constructed a class of models for associative retrieval broad enough to encompass traditional matching operations as well as ranking by "relevance indexes" (see Vincent E. Giuliano and Paul E. Jones, "Linear Associative Information Retrieval", in Vistas in Information Handling, Vol. 1, Ed. by Paul Howerton and David Weeks, Washington, D.C., Spartan Books, 1963, p. 30-54). The class of models also lends itself to a study of the effects of errors. In the Cranfield ASLIB study (cf. I-2) the synthetic nature of the searches must be recognized as one of the principal weaknesses of the experiment. The analysis of failures to retrieve yielded the most useful results. The failure to index a concept was a more frequent error than overindexing. These results agree with experiments at the Patent Office (cf. II-24). Cleverdon proposed the "relevance ratio" and "recall ratio" as the two criteria by which to compare systems. Much has been made of the "inevitability" of the inverse relationship between these ratios. For some purposes one can tolerate a small relevance ratio if the total retrieval is small (less than 50) and contains one or more good references. Excessive concern about the inverse relationship has obscured some of the real issues, including some important definitions. Thus, relevance is not an inherent property of a document, but must be established as to what it is that documents are to be judged relevant to, and by whom.

Nearly all evaluations of retrieval systems run a series of searches based upon a set of sample questions. We have difficulty in defining what is a question and the population of users. Cleverdon (cf. I-2) had questions prepared from documents known to be in the file. Mooers (cf. IV-27) suggested selecting a random subset of documents in the file to be used as a test file; users prepare three questions for each document, one question for which the document is crucial, one relevant, and one irrelevant. Fels (cf. II-28), in searching legal text, used a modified Mooers plan. Bornstein (cf. IV-10) suggested four classes of relevance and construction of questions by group interviews, without regard to

what is known to be in the file. Rees (cf. I-6) pointed out the importance of question analysis. Borko (cf. IV-9) also suggested four classes of relevancy and groups of questions from the user population. The Patent Office (cf. II-24) indexed a sample of chemical patents in various ways to determine the effect of indexing errors on retrieval. Results with real questions searched by examiners and synthetic questions composed by chemical analysts were compared.

What to measure and how to measure it are ever present problems. Operating cost is relatively easy to measure or estimate for operating systems, time comparisons can be made, timeliness is also an important item, and quality must be assessed in some manner. Some judgment must be made concerning the applicability of the references retrieved under alternative systems. One of the areas which needs work has to do with the problem of scale. A great deal also needs to be done with the problem of evaluating file subdivision, whether by classification, clustering, or more conventional techniques.

It is suggested that model building and experimentation be considered together. Most experiments should be designed to examine components of the model rather than the entire model. We must study the failures of systems and judge whether the component that failed can be improved. Probably no other issues are as controversial as the selection of test questions and determination of documents relative to them. The basic problem of evaluation is one of matching the output of the system against the need, the request (expression of the need), the query (request in index vocabulary terms), or the encoded form of the query. The proper comparison depends on the inadequacies of the system that are to be revealed: the vocabulary, input errors, or errors in the search. If we compare output against the query, we are checking on input errors or deficiencies. Failure to retrieve and excess output should be analyzed. Judgment as to whether the output matches the query can be made by an impartial observer.

If we compare output with request we are confounding errors of input with errors in stating the query. The translation of request into query must be studied; if the request cannot be encompassed by the query except at the cost of high unwanted retrieval, the vocabulary is defective. Recall is related to errors in the system and relevance to inherent deficiencies in the system.

No one but the user can compare output with need. This comparison confounds input errors, errors of translating need into request, and errors in preparation of the query.

One thing we would like to measure is the proportion of time the system satisfies the user's need. One technique is augmentation of the output with a small random sample of the remaining documents in the file. The proportion of time the user satisfies his need by a document in the random component, when averaged over a long sequence of trials, indicates the adequacy of the system. The reasons for failure to provide these documents should be analyzed.

There is no simple solution to the problem of obtaining a test corpus nor of defining a test run. It is desirable to draw a random sample of documents from the file about which generalizations are to be made. In the initial stages of development of a test file, artificial queries may be used to test the system; as quickly as possible, queries framed by real users should be used. The point at which to start "real life" testing depends to some extent on the results of preliminary testing, possibly when about one-quarter of the file or 1,000 documents (whichever occurs first) have been encoded. It is important here to analyze failures in view of restraints on user time, and to determine whether failures are inherent to the system or can be eliminated by controlling indexing and/or searching errors. The concept of "relevance" may be modified as the depth of testing changes; in real life testing, the criteria are the number of documents looked at in order to reach a conclusion and the probability of missing a better reference.

III-5. The In-House Testing and Evaluation of the Operating Efficiency of the Intellectual Stages of Information Retrieval Systems, by Cyril W. Cleverdon, Cranfield, England, College of Aeronautics, September 1964, 13 p.

(This is a slightly amended version of the paper prepared for the F. I. D. Conference on Classification Research and appears in Classification Research (Proceedings of the Second International Study Conference, Elsinore, Denmark, September 14-18, 1964), Copenhagen, Munksgaard, 1965.)

Terms used in the title are defined as follows: "intellectual stages of information retrieval systems" refers to the decisions made in the indexing of a document as to concepts recognized, the decisions made at the search stage as to programs used, and the decisions involved in the design and maintenance of an index language (the set of indexing terms used in the system, the rules for their use, and aids such as cross-references or classification schedules); "operating efficiency" refers to the ability of the system to retrieve references to all relevant documents and not retrieve non-relevant documents (recall ratio and relevance ratio, respectively); "testing" provides data for "evaluation", which covers the analysis of test results and decision as to actions for improving the operating efficiency or, while maintaining the operating efficiency, improving the economic efficiency; "in-house" implies the testing of an actual operating system rather than one specially prepared for testing purposes.

There are a number of basic statements to be made: (1) there is an inevitable inverse relationship between recall and relevance; (2) the optimum performance curve of an information retrieval system depends on the set of documents included in the system (the more precise their language, the nearer will the curve come to the ideal of 100 percent recall and 100 percent relevance), the set of questions put to the system, and the level of relevance demanded (these two are interrelated, because the generality or specificity of the question can be expressed as a variation in the level of relevance demanded); (3) the maximum possible recall figure is determined by the level of exhaustivity of indexing (the highest level of exhaustivity implies the recognition and inclusion in the index of every concept); and (4) the maximum possible relevance ratio is dependent on the degree of specificity of the index language (the index term is co-extensive with the concept and covers nothing else besides the concept).

The factors to be investigated in an information retrieval system include the normal operating performance: One needs first to know the general indexing strategy and the particular indexing decisions for each document; a set of search questions is required, and a user group to analyze the search outputs for relevance; the actual programs used in the searches, with complete records of all documents retrieved, must be available; finally, the test must be designed. An ideal test situation hypothesizes a system where indexing is done at two levels of exhaustivity --- at a low level for production of a printed index and at a higher level for entry into a computerized index. The index language is a strictly controlled list of descriptors. The entries are arranged serially in the computer, and the system covers some 100, 000 documents indexed over the past three years. A supply of actual search questions is available, and the subject field of the collection is not too diffuse.

Having the results of actual searches analyzed by the user group will give the relevance ratio, probably in the range of 10 percent to 20 percent. To establish the recall figure by finding how many relevant documents were not retrieved, three alternatives are possible: one can add to the actual output for each search a number of documents selected at random from the whole collection (but this method will not provide sufficient data without giving the questioners an inordinate number of documents to assess); one can take a sample block of the collection and have each document in the block assessed against each question (but this method would involve a very large number of individual assessments, too time-consuming

a task to do efficiently); or one can make searches using questions that have been based on documents known to be in the collection (prepared questions based on source documents, which give reliable results).

A major criticism, the assessment of relevance, asks how one can validly produce precise performance figures based on imprecise subjective assessment. It must always be assessment by the person asking the question, for systems have to meet the requirements of individual users. It is desirable to equate the different overall judgments which individuals might make in relation to different questions. To do this we must specify the environment of the system being tested, in relation to the number of documents in the collection and the number that are relevant to a question. It would seem appropriate to express this factor as 1,000 C/N, where C equals the number of documents of an agreed standard of relevance to a question and N equals the number of documents in the collection. This is called the generality ratio and evidence shows that at a given recall ratio there is nearly a straight correlation between the relevance ratio and the generality ratio. The difficulty arises as to how one calculates the figure for N. It is tentatively suggested that N be taken as the number of documents in the smallest classes in which a collection can be divided whilst still being certain of obtaining 100 percent recall.

Testing is the first stage of evaluation, and provides the necessary data for analysis. The first job of analysis must be to discover why relevant documents have not been retrieved, and to decide whether the failures were due to bad indexing, bad search programming, or weakness in the index language. The next stage in analysis is to find why non-relevant documents are being retrieved; the same principles of analysis apply here and similar reasons will be shown. It was originally hypothesized that the test was being made of two indexes, permitting comparison of the results. From this it is possible to make an evaluation of the whole system and estimate the effect that changes would have on the operating efficinecy of the system.

To determine whether a system meets the requirements of the users would demand a "field" evaluation to show in what percentage of cases a quick but possibly incomplete search in a published index met the requirements, and how often some delay in having a complete search made by a computer would be acceptable.

The weaknesses of the test proposed here include the matter of relevance assessment discussed above and the difficulty of finding two separate indexes which can be strictly compared. The nearer one comes to research testing (as for example at Western Reserve University (cf. I-1), or the ASLIB Cranfield project (cf. I-2),the more certain one can be regarding the reliability of the results. We are left with a compromise between absolute reliability of results, amount of information required, and amount of time and money spent on the test. Experience suggests that the actual test results are relatively unimportant compared with the evaluation which the testing permits.

III-6. On the Utility of the Relevance Concept, by Carlos A. Cuadra, SDC Report No. SP-1595, Santa Monica, Calif., System Development Corp., March 18, 1964, 9 p.

A key concept in evaluation of information retrieval systems is "relevance", the number of relevant items retrieved in relation to total number of items retrieved, or to number of non-relevant items retrieved, or to number of relevant items not retrieved. Some argue that the concept is not adequate, that distinction must be made between relevance to a searcher's need and relevance to his request. However, we must distinguish between relevance as a "construct" (i.e., an intellectual synthesis) and the means (e.g., judges' ratings) which have been used to measure it. We also need to examine the results of using

certain measures of relevance before we decide to discard the concept or the measures. The only level of correspondence open to measurement is that between the expression of an information need (rather than the need itself) and the system's response. The question of choosing the best form of judgment, i.e., who should evaluate relevance at this level is still open, and needs research attention.

III-7. Is Relevance an Adequate Criterion in Retrieval System Evaluation?, by Lauren B. Doyle, in Automation and Scientific Communication, Short Papers, Part 2 (papers contributed to the 26th Annual Meeting of the American Documentation Institute, Chicago, Ill., October 1963), Ed. by H. P. Luhn, Washington, D.C., American Documentation Institute, 1963, p. 199-200.

In evaluation one compares system performance to some ideal or standard; in the case of document retrieval the ideal of performance has been to retrieve all and only those documents relevant to a particular need. For this purpose, judges inspect documents and agree as to which ones should be retrieved in response to a given request. But there may be a great difference between relevance to a given request statement and relevance to a real information need. If several persons ask the same question and receive the same documents in return, each person will pick out a different subset of the total package as being relevant to his question. This reflects the general inability of searchers to ask the right questions of the system. Feedback must be provided so that the searcher can redefine his need in a series of iterations. But exploratory capability is not always provided by modern machine literature searching systems. The concept of relevance allows suboptimization on the machine side of the man-machine interface. But the human side must also be studied, as well as both sides in interaction. The concept of relevance might be replaced by the concept of "sharpness of separation of the exploratory regions in which the searcher finds documents of interest from those in which he does not find such documents." The individual information need is so complex that it cannot be stated accurately in a simple request for which relevance of documents can be measured.

III-8. Implications of Test Procedures, by Robert A. Fairthorne, in Information Retrieval in Action (Papers presented at 1962 Conference at Center for Documentation and Communication Research, Western Reserve University), Cleveland, Ohio, The Press of Western Reserve University, 1963,

An information retrieval system can claim only to provide, at given cost and within given time and with given completeness, documents that assist the user more than other documents would, within the limitation of providing only what the user requests. If one can compare a given system with an ideal system ideally invoked, then one has an absolute test of efficiency under given circumstances. Mooers (cf. IV-27) envisages someone who knows the entire collection and can state what documents are relevant to what requests; he sets up procedures for estimating, in the statistical sense, the performance of such an ideal in retrieval and for comparing an actual system performance with this. With relatively small specialized collections this idealization corresponds to reality. However, the ideal retriever must be considered as knowing only the contents of the collection and the documentation habits of his colleagues. He must not make inferences about relevance of documents to requests, based on understanding of the subject matter. The only relevance that can be tested or measured is that based on correspondence between the words of the request and the words of the collection. A measure of relevance must apply only to what can be measured.

Artificial requests are put to the system not to retrieve information but to recall certain expressions known to be in the collection. Such requests could be achieved by actual users only by iteration of a request according to the documents retrieved by the previous request. Iteration is a procedure for improving the language, not the basic thinking of the request. Tests using requests based on known documents tend to favor systems based on the actual words used in a document; however, they are legitimate and useful inasmuch as they approximate the "ideal user." The utility of these recall tests might be increased if queries were expressed so as to be satisfied by more than one document. This would reduce the bias in favor of textword indexing. In whatever way "relevance" is defined to give a measure to some aspects of a retrieval system, it refers to a type of request rather than to a user; it can persist usefully only within the field of public agreement. If a homogeneous group of users examines the total response to all their individual questions, each user can find the response best for his own request. Group response to a group request shows how well the system supplies all pertinent documents, but not only such documents; it also shows how well the system supplies only pertinent documents, but not all such documents. The important point is that retrieval of any kind must be within the scope of public agreement; the individual user is too small a public.

III-9. Methodology for Test and Evaluation of Information Retrieval Systems, by William Goffman and Vaun A. Newill, Report No. CSL:TR-2, Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, July 1964, 19 p.

Experimental studies of the performance of information retrieval systems, with the exception of the work of Cleverdon (cf. I-2), have been rare. Few of the abstract theories proposed have been derived from experience with operational systems and fewer still have been tested against real user needs. The Comparative Systems Laboratory (CSL) has been established to obtain quantitative experimental results in this area. A system is defined as an integrated assembly of components that interact cooperatively to perform a predetermined function for a specific purpose. The information retrieval process requires matching a query against a collection of documents called a file. That subset of the file matched with the query is called the answer, and the property which assigns members of the file to the answer is relevance.

Performance of systems might be compared in two ways: a single system's performance of several different tasks may be compared, or several systems' performances of a single task may be compared. In the CSL components are defined and evaluated in a manner that permits their assembly into a system. Evaluation will be accomplished by measuring the variability in the system performance caused by alteration of the components contained therein. The major components of an information retrieval system related to function, the ones required to carry out the process of providing an answer to a query, are acquisitions, source of input, indexing language, coding, organization of the file, question analysis, searching strategy, and format of output. The main concern of the work in the CSL is with these eight major function components.

The performance of a system is measured in terms of a function of two variables, the effectiveness and efficiency. Effectiveness is defined as the measure of the system's ability to perform the task for which it was designed, while efficiency is a measure of the cost of performing this task. The effectiveness measure must be an evaluation function that behaves in accordance with requirements imposed on the system's performance. Effectiveness is defined as a function of sensitivity and specificity, where sensitivity measures the system's ability to provide the user with relevant members of the file and specificity measures the system's ability not to provide him with nonrelevant members. A uniquely defined subset of the file, which serves as a basis for measuring variations in

systems performance, is called the ideal set or relevant set, and will be represented by that subset of the file the user would have selected as a response to his query had he searched the file himself. Since the query is formal representation of some need, it follows that documents relevant to the query may not necessarily be appropriate to the need. A subset of the file appropriate to the need is called a pertinent set. Thus the difference between relevance and pertinence is that relevant documents answer the user's query and pertinent documents answer his need. Study of the relationship between the two is needed.

The measure of sensitivity is the conditional probability that a member of the file will be retrieved by the system given that it is relevant or pertinent. The measure of specificity is the conditional probability that a member of the file will not be retrieved by the system given that it is not relevant or pertinent. Effectiveness (u) as a function of these measures is expressed as their sum. The effectiveness will be 0 if the conditional probability of a document being retrieved given that it is relevant or pertinent is equal to the conditional probability of its being retrieved given that it is nonrelevant or nonpertinent. Positive values of u indicate that the chances of the system yielding useful members of the file are better than its chances of yielding non-useful members, while negative values of u indicate the opposite.

Sensivity and specificity are equally weighted in this effectiveness measure, but it would be more realistic to assign coefficients of value representing the worth of a successful retrieval and the nuisance of a spurious response. Such coefficients can only be determined in a specific situation and would vary with the situation.

The conditional probability, which is a proportion, evaluates the extent to which a given event is contained in some other event. We can approximate sensitivity by the intersection of the documents in the answer and those in the ideal set divided by the ideal set, and specificity by the intersection of the documents not in the answer and those not in the ideal set divided by the not ideal set. Evaluation of the system output, determination of the ideal set, can only be made by the user who must be presented with a set from which to make his evaluation (the universal set). Where the experimental file is very large, subsystems are defined and the union of their outputs is the universal set. If the measure of sensitivity or specificity of one subsystem is less than the measure of sensitivity or specificity in another subsystem relative to the union of their outputs, the inequalities also hold with respect to the total file. This indicates that if the evaluation of the output is based on an approximation of the file, the analysis cannot be based solely on the effectiveness measure but must be examined with respect to the sensitivity and specificity measures.

III-10. The Place of Indexing in the Design of Information Systems Tests, by Alvin J. Goldwyn, in Automation and Scientific Communication, Short Papers, Part 2 (Papers contributed to 26th Annual Meeting of the American Documentation Institute, Chicago, Ill., October 1963), Ed. by H. P. Luhn, Washington, D. C., American Documentation Institute, 1963, p. 321-322.

(This paper is also available as Report No. CSL:TR-3, Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, August 1964, 9 p.)

Because of difficulties in measuring the effectiveness of an information retrieval system, attention has been turned to testing one or another aspect of a total system. Indexing procedures have been tested, for example, but with only moderately interesting results in assigning ratings to indexes. In testing one aspect of a system it is difficult to neutralize all the other variables. The Center for Documentation and Communication Research is setting up a Comparative Systems Laboratory, with support from the National Institutes of Health. Resources include pilot files in several medical subject areas and an

accumulation of data and experience relating to the general theory of documentation. But an important question with regard to the comparison of separate information systems is how different must the systems be? A system covers the whole flow of information, including acquisition, storage, search and dissemination. It is not clear that varying the method of indexing creates another system. Five "systems" for handling a collection of documents in the communicable disease literature, for example, vary only in their indexing techniques and, as a result, in their methods of storage (conventional manual index, permuted keyword bibliography, telegraphic abstracts with semantic codes on tape, etc.). To answer the question of how different the systems should be, the Center will attempt to define the components of an information retrieval system, to identify the variables affecting the system's performance, to design techniques for testing systems, and to test the tests. At least effectiveness and efficiency need to be measured, as suggested by Swets (cf. IV-31). But only total systems can be compared by today's test methods; a single operation such as indexing has not been isolated for testing. Those who have had real experience with methodological issues and who can design experiments to test the variables that exist must design meaningful tests of information retrieval systems or of their parts.

III-11. Purpose and Objectives of the Comparative Systems Laboratory, by Alvin J. Goldwyn, Report No. CSL:TR-1, Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, August 1964, 21 p.

Documentalists today are giving their attention to a number of activities which include evaluating systems which they have built and studying their own techniques from the systems point of view. The Comparative Systems Laboratory (CSL) at the Center for Documentation and Communication Research (CDCR), which approximates in its design these various activities, has as its primary purpose: (1) to define the essential components of an information retrieval "system"; (2) to identify the variables affecting the performance of a "system"; (3) to design techniques for testing "systems"; and (4) to test the tests. A system may be defined as the structuring of the whole complex flow of information, including acquisition procedures, storage techniques, question analysis and searching techniques, and dissemination processes. The purpose of the CSL will permit the CDCR to exploit its files of biomedical literature, to enlarge the experience gained in the collection of these files, to explore the interaction between the files and the users for whom they were intended, and to provide for teaching and training of biomedical documentation.

Certain other documentation research is being coordinated within the framework of the CSL: fundamental study of the nature of relevance, study of medical vocabulary, coordinated action between a stored-program computer and a mechanized dictionary, and reprocessing of previously processed materials into machine-searchable form.

The principal tasks of the Laboratory are to determine at what point, and under what conditions, the performance of a retrieval system is optimized and to establish the comparative performance of a number of systems under controlled conditions. The influence of a number of variables is to be tested, including indexing quality, relevance, other factors affecting retrieval quality, effectiveness of the source, and comparative costs. The first set of experiments will test whether alterations in methods or subject make an important difference in the documents retrieved as evaluated by sensitivity (ability to provide members of the ideal set), specificity (ability not to provide members not belonging to the ideal set), and effectiveness (a function of the first two). Details of methodology are described by Goffman and Newill (cf. III-9).

Preparation of files and planning of experimental design have been the chief areas of current operation. Unwanted variables are to be reduced to the minimum so that testing would indicate significant results. Contact has been made and cooperation received from two user groups, to provide "real" questions and evaluation for the actual testing phase.

III-12. Measuring the Value of Information Services, by James Hillier, Journal of Chemical Documentation 2, 31-34 (January 1962).

(Essentially the same information in this paper is contained also in "Management's Evaluation of Information Services", by James Hillier, in Information Retrieval Management, Ed. by L. Hattery and E. McCormick, Detroit, Mich., American Data Processing, Inc., 1962, p. 54-60.)

An information service is considered to be an organization with a responsibility to take positive action toward satisfying the information needs of the technical staff it serves. The first task is to establish the nature of the needs. Effective information flow in an industrial research laboratory doing exploratory research in a number of different scientific disciplines must satisfy several types of needs for information: it must prevent excessive duplication of research, provide specific information needed by the technical staff, provide "catching-up" information for the individual who finds it necessary to become familiar with a new field, provide an efficient means for enabling the member of the technical staff to "keep current" in his project and his field, and provide an information flow which will stimulate creative thought in a way which will maximize the probability of occurrence of creative ideas that are valuable to the company. But the individual in a creative pursuit will have learned from experience the channels through which he has the highest probability of obtaining a useful piece of information and will tend to select those channels whenever a choice is involved. An information service will have to be able to demonstrate by actual performance that it provides a more profitable input channel to the creative individual. Moreover, it is the judgment of the individual rather than that of management that will prevail. Management's responsibility is to find means of increasing the probability of creative action by the technical staff by increasing the individual probabilities involved.

Management's considerations must go beyond the simple provisions of extra brains to perform a function that has grown to exceed the capability of the individual. The additional brainpower must operate in conjunction with the individual's experience-based optimization of his information input. A further consideration is the balance between a centralized information service and a dispersed system. Here too the effects on the probability of creative action of the two approaches must be a most important factor in the evaluation. A qualitative understanding of the role of an information service activity can be developed on a rather general fundamental basis. There is not yet nearly enough detailed and quantitative data to enable management to make a reliable evaluation of an information service activity in a specific laboratory.

III-13. The Notion of Relevance (1), by Donald J. Hillman, American Documentation 15, 26-34 (January 1964).

The importance of having a clear notion of the concept of relevance, or connectivity, cannot be doubted, for retrieval outputs consisting of citations to documents either unrelated to each other or irrelevant to the prescription are unserviceable for the user. It is imperative that the concept of relevance be adequately defined and a suitable place found for it in the theory of information storage and retrieval systems. The problem is to describe a concept of relevance independent of, and logically prior to, any notion of relevance as determined by, and thus restricted to, a particular system of storage and retrieval. More basically, the problem is one of characterizing the relatedness of concepts so as to make such characterization independent and logically prior to any notion of relevance determined by a given system. It is to be presumed that any characterization of the relatedness of concepts will, if adequate, serve as the basis for all relevance judgments expressed by explicit operations on symbols denoting concepts in any system whatsoever of information storage and retrieval.

Even though the problem of relevance is susceptible of a reasonably clear formulation, there exists an abundant variety of opinion concerning its method of solution, ranging from absolute pessimism to a cautious offer of promise. This paper examines the possibility of constructing a formal theory of relatedness or relevance satisfying the conditions of independence and logical priority described above. The working hypothesis is that the most desirable retrieval output for a prescription requesting all documents relevant to a given topic will consist of citations to all those items that are conceptually related to the topic in question. Consider the interpretation of "concept" as a property of, or relation between, objects of one kind or another. Properties may be regarded as properties in extension or properties in intension. It can be said that the extension of a dyadic predicator is the class of ordered pairs for which the predicate in question holds and that its intension is the relation which it expresses. The empirical data for which the theory of relevance must account take the form of similarity judgments which differ from one questioner to another, and are regarded as description pairs of like and unlike documents supplied by the questioner in any way he chooses. The function of the theory is then to provide the appropriate formal structure accommodating the questioner's similarity judgments. One and the same theory will suffice for all different user requirements; no matter what preferences the questioner lists, the theory will always be able to account for them.

In order to secure such generality, it is plain that the theory must always provide the means for constructing classes of conceptually similar documents on the basis of relevance-judgments supplied by any user. The theory must solve two major problems, that of defining classes corresponding to concepts in their extensional interpretation, given only a list of similarity pairs provided by some user, and that of the definition of the relatedness of concepts thus construed. In the theory concepts are regarded as classes. One of the advantages of theory for retrieval purposes is that it permits the association of a concept with a class of documents possessing the property expressed by the concept. Instead of considering relations between classes, formation of those (larger) classes whose members belong to the field of a partial similarity relation might be attempted. Classes thus formed will have as members all documents that are conceptually related as defined by some partial similarity relation.

Now the problem is one of defining concepts via properties on the basis of similarity judgments, and defining conceptual relatedness on the basis of pairs of partial similarity judgments. The first part is essentially the problem of abstraction, i.e., the separation of a property or quality from the document (object) to which it belongs. We deal only with certain lists of pairs belonging to the field of the relation "is similar to". A method

already exists for dealing with this problem of abstraction (cf. Carnap, R., Der Logische Aufbau der Welt, Berlin, 1928), applicable to the problem of defining the concepts contained in the documents of a collection, given only a questioner's list of pairs of document descriptions which he affirms to be similar. Yet this whole process of concept formation suffers from a weakness that limits the successful application of the method to those cases where concepts are independent of each other. Since there would be no a priori information concerning the independence of concepts expressed by the documents of a collection, there is no initial guarantee that the method of concept-formation will always work. Another difficulty is that where circumstances are such that every pair of a set of documents has a concept in common, yet no concept in common to all elements of the set. It is apparent that too many dubious assumptions of a preanalytic nature are needed to insure the applicability of this method of abstraction, or indeed of any method which purports to form concepts on the basis of recognizing "clumps" of similar objects.

It might be thought that once these purely accidental deficiencies of the definitions and measures are overcome, a successful method will be forthcoming. Unfortunately there are good reasons for attributing the failure of these methods to shortcomings in the general method of constructing similarity-classes to define concepts, independent of any particular way in which the classes in question are arrived at. It might be objected that failure to define concepts in the ways considered show neither that no method resembling Carnap's can be discovered, nor that the further project of defining conceptual similarity is impossible of fulfillment. Although we have no explicit proof of failure for all methods resembling those considered, we have good evidence for believing in their essential sterility. The point of the second objection is that it may be possible to define similarity of concepts without first having to define concept. If the definition we seek is of an extensional nature, it will obviously not be identical with that of class similarity, that two classes will be similar if they are equivalent classes. The similarity of concepts, even if the latter are to be regarded as classes of documents, is quite different from class equivalence.

Confronted by this overwhelming array of difficulties, it must now be admitted that the attempt to define document relatedness on the basis of some similarity relation has not been successful, nor will it be without recourse to highly dubious and quite unverifiable assumptions governing the patterns of concept co-occurrence. A more promising approach is to consider the grounds on which users themselves form similarity judgments, and perhaps profitably turn to an analysis of similarity judgments in order to determine what it is that forms their basis. This problem can be treated as one of providing a formal theory of conceptual similarity, except that in this case similarity judgments will not be taken as unanalyzed, primitive assertions. Instead there will be chosen a new primitive relation in terms of which similarity judgments will be desirable. This new primitive will not be entirely independent of human judgment, which will guarantee the primacy of human judgment in the assessment of relevance, although the theory based on the new primitive will be unaffected by any of its interpretations.

This problem will form the topic of Part II of this paper (to be published).

III-14. Techniques for Relating Personnel Performance to System Effectiveness Criteria: A Critical Review of the Literature, by A. J. Hoisman and A. M. Daitch, Report to Bureau of Naval Personnel, Santa Monica, Calif., Dunlap and Associates, Inc., September 1964, 45 p.

Of some 50 reports judged to have relevance to the subject (from over 400 reports reviewed), those of particular interest to this bibliography are the reports from Magnavox Research Laboratories and ARINC Research Corp., and the paper by White, Scott and

Schulz. The Magnavox report (see Magnavox Research Laboratories, "Mathematical Models for Information Systems Design and a Calculus of Operations", Report No. RADC TR-61-96, Final Report to Rome Air Development Center, Torrance, Calif., Magnavox Research Lab., October 27, 1961, 178 p.) is addressed to the problem of development of a mathematical model that could be used in the design of information storage and retrieval systems. Elements of a system are considered to be "jobs and components." Models are constructed using linear algebra (e.g., matrices and eigenvalues). Some specific models developed describe the efficiency with which a given component performs a given operation, and an efficiency matrix for a system composed of a specific set of components. The latter model represents a system by a matrix of efficiency values of the individual components. The efficiencies are functions of cost, time, and size of an operation. The result of mathematical manipulations is a deterministic efficiency matrix for any given system configuration. The efficiency matrix is predicted on an index of efficiency whose probability-distribution is unknown. As an analytic tool the model emphasizes the relations among cost per unit time, time, and size of the operation.

In ARINC's proposed method for system evaluation (see ARINC Research Corporation, "System Effectiveness Concepts and Analytical Techniques", ARINC Pub. No. 267-01-7-419, Washington, D.C., January 1964), the major purpose is to provide an initial step in categorizing and describing inputs and methods required for formal quantification of system effectiveness. The effectiveness model framework is introduced by discussing various figures of merit for system effectiveness; the probabilistic figure of merit is the probability that the system can meet an operational demand within a given time when operated under specified conditions. Even though the basic model of system effectiveness appears to have simplicity, it is stated that quantification of the various elements makes analysis a problem. Further analysis treats optimization of system effectiveness in terms of cost, and this requires representing effectiveness as some function of cost. Possible optimization methods include mathematical, statistical, programming, and operations research techniques. The model developed has potential ramifications in several areas of systems research; its principal design emphasis is as a method for performing evaluations.

The paper by White, Scott, and Schulz (see White, D.R.J., D.L. Scott and R.N. Schulz, "POED - A Method of Evaluating System Performance", IEEE Transactions on Engineering Management, EM:10, December 1963, p. 177-182) describes the development of POED, the Performance Organization for Evaluation and Decision. The POED method is a system performance evaluation technique using a figure of merit representing a performance of effectiveness assessment in terms of the user's requirements. An overall system figure of merit may be fractionated to lower echelons, e.g., mission, system, subsystem, and facet echelons in decreasing order. Lower order figures of merit are related to a higher order figure by means of weighting factors that correspond to relative importance and/or relative applicability for each lower order element. The POED technique contains three principal components: the figure of merit mathematical model, a sensitivity model, and a confidence factor. The sensitivity model defines a change in the overall figure of merit for a given change in a lower echelon figure of merit. The confidence factor is defined as the ratio of the mean to the standard deviation for a population of figure of merit data; the higher the value of the ratio, the higher the confidence in making a decision based on variability of the input data. This method should not be a candidate for general system analysis use; the figure of merit measure is rather weak, the sensitivity model is not unique, and the confidence factor is of questionable meaning.

III-15. Minimum Criteria for a Coordinated Information System, (Letter to the Editor) by Allen Kent, American Documentation 11, 84-87 (January 1960).

A suggested statement of criteria for a coordinated information service, prepared by the Documentation Committee of the American Society for Metals, includes factors such as scope of subject matter covered by the system, variety of services to be provided, expandability into other subject areas, timeliness, and costs of operating the system, including costs of inputting information, extracting information, and using the information so extracted. Technical criteria that must be considered in evaluating these factors include file size, amount of activity, penetration (depth of indexing, representation of relationships, etc.), ability to control (synonyms, special meanings, generic or specific aspects), and quality (including "noise" level --- the amount of irrelevant material selected or relevant material not selected). To evaluate a particular system in terms of particular requirements, these criteria should be considered and a set of standards should be established for each of them. A list is given of these various factors as they applied to the pilot machine searching system at Western Reserve University.

III-16. Methodology for the Comparative Analysis of Information Storage and Retrieval Systems: A Critical Review, by Irving M. Klempner, American Documentation 15, 210-216 (July 1964).

A comprehensive search of library and documentation literature shows how few studies seek to establish through comparative analysis the superiority of one system of information storage and retrieval over another. This paper examines these studies, the framework within which they were carried out, and particularly the methodologies employed.

Using historical analysis, Jonker attempted to synthesize a "generalized" theory of indexing and sought to discover differences and similarities among various indexing systems (see "The Descriptive Continuum: A 'Generalized' Theory of Indexing", by Frederick Jonker, in Proceedings of International Conference on Scientific Information, Washington, D.C., National Academy of Sciences-National Research Council, 1958, p. 1291). He traced the development of indexing methods from the earliest hierarchical classification to the present combinatory systems. The various indexing systems, rather than being contrasting systems, were found to be merely an outgrowth of one another and not at all in conflict.

With normative survey methodology, Jahoda attempted to find out whether correlative indexing had any advantages over traditional indexing (see Correlative Indexing Systems for the Control of Research Records, by Gerald Jahoda, unpublished DLS Dissertation, New York, Columbia University School of Library Service, 1960). By analyzing replies to a questionnaire and information gained from interviews he compiled the fundamental ingredients of an index to research records. He sought to establish reasons for traditional indexing failures and to analyze reference questions which allegedly could not be handled with traditional systems. Jahoda concluded that such systems were as effective as correlative indexes, and that correlative indexes were interchangeable with one another. No experiments were undertaken to determine effectiveness of the respective systems or superiority of one over another.

In finding an efficient system for the storage and retrieval of Defense Documentation Center (formerly ASTIA) documents, Documentation, Inc., developed the Uniterm system of coordinate indexing. A reference test was carried out (cf. I-14) in which 15,000 ASTIA documents were indexed by the Uniterm system and by the ASTIA conventional system with descriptive cataloging and alphabetical subject analysis. Approximately 100 reference questions were selected from routine requests received by ASTIA, and were searched

under both systems by, respectively, Documentation, Inc., and ASTIA personnel. Relevance or non-relevance of documents was determined for each group by its respective personnel. Since relevancy was interpreted differently by each group, and since other variables were not tested under comparable conditions, the results of the test were inconclusive.

More refined methods for testing were used by Stevens in his investigation of three systems --- a punched card file, a handbook reproduced from that file, and a conventional technical library catalog (cf. I-30). The investigation included study of the background of the three systems; of the use made of them; and of their operation in terms of time, cost, and comparative retrieval efficiency. Conclusions were that manual searching of the handbook was more efficient than machine searching of the punched cards, and that for the general question the reference library approach proved most satisfactory. The methodology employed by Stevens takes into account both internal factors of the system itself and external factors such as consumer satisfaction.

The ASLIB Cranfield Project represents yet a more sophisticated approach to the comparative analysis of information systems (cf. I-2). Primary variables tested were indexers with different educational and experience backgrounds, different systems, and different indexing times. The four systems selected for comparison were the Universal Decimal Classification, an alphabetical subject catalog, a faceted classification scheme, and the Uniterm system of coordinate indexing. To test retrieval efficiency, 1, 600 questions were selected from documents known to be in the collection and to provide a satisfactory answer; this method tackled the relevancy problem. One realization emerges from the investigation: the study represents significant studies in the application of experimental methodology for the evaluation of information retrieval systems. However, the factors investigated relate to internal efficiencies of the systems: external criteria are yet to be analyzed. Other methodology would have to be applied before a composite and representative constant could be obtained as a measure of the overall efficiency of the systems.

III-17. Toward Document Retrieval Theory: Relevance-Recall Ratio for Text Containing One Specified Query Term, by Manfred Kochen, in Automation and Scientific Communication, Proceedings, Part 3 (papers presented at the 26th Annual Meeting of American Documentation Institute, Chicago, Ill., October 1963), Ed. by Paul C. Janaske, Washington, D.C., American Documentation Institute, 1963, p. 439-442.

The performance of a document retrieval system is sometimes measured by the extent to which it attains the desired goals of missing only very few documents highly relevant to a given query, and retrieving only few documents of no or only moderate relevance. These variables are called "hit rate" and "acceptance rate" here. A trade-off relation between the two is derived, based on certain assumptions about the likelihood with which authors use words when they are relevant to a topic and when they are not. That is, mathematical explanation is given for an increase of acceptance rate involving a decrease of precision in document retrieval. To this end three assumptions are made: (1) about the use and frequency of words in requests, (2) about the use and frequency of words in the text (a Zipf - Mandelbrot relation), and (3) that texts are retrieved only if they contain words used in the request. When the acceptance rate is plotted against the hit rate for various values of the parameters, certain properties of the curve can be discussed: when the hit rate is maximum, the greatest attainable acceptance rate is the same as the a priori probability that a randomly chosen document will be judged relevant. When the hit rate is minimal, the acceptance rate is high; as the hit rate increases slightly, the acceptance rate drops rapidly to a minimum. A topic with many different keywords and a limited vocabulary could give the largest hit rate but the associated acceptance rate will be small. The shape and location of the curves suggest that retrieval based on a small sample of the text and title is perhaps almost as effective as retrieval based on the entire text.

The two measures of performance of a document retrieval system bear a significant analogy to concepts in statistical-decision theory: errors of type I (analogous to hit rate) and errors of type II (analogous to acceptance rate). The mathematical analysis reported here should be regarded as the first in a series of steps toward document retrieval theory.

III-18. On Relevance, Probabilistic Indexing and Information Retrieval, by M. E. Maron and J. L. Kuhns, Journal of the Association for Computing Machinery 7, 216-244 (July 1960).

The major difficulties associated with the problem of information search and retrieval are the identification of content, determination of which of two items of data is "closer" in meaning to a third item, and determination of whether or not (or to what degree) some document is relevant to a given request. The fundamental notion here is that of the relevance number, a measure of the probable relevance of a document for a hypothetical requestor, which provides a means of ranking documents according to their probable relevance to a particular request. For proper selection of those documents which are to be ranked, it is necessary to establish various measures of closeness of probable meaning, via statistics. The notion of weighting the index terms that are assigned to documents and using these weights to compute relevance numbers is basic to the technique called Probabilistic Indexing. If a strictly quantitative measure for relevance is not possible, a comparative measure would make ranking of documents by relevance possible.

By P(A, B) is meant the (conditional) probability of an event of Class B occurring with reference to an event of Class A. Thus if $P(A \cdot I_j, D_i)$ = the probability that if a library user requests information on I_j, he will be satisfied with document D_i, then by the use of Bayes' theorem one can treat $P(A \cdot I_j, D_i)$ as proportional to the product $P(A, D_i) \cdot P(A \cdot D_i, I_j)$, where $P(A \cdot D_i)$ is the a priori probability of document D_i satisfying a request, and $P(A \cdot D_i, I_j)$ is the probability that if a user wants information of the kind contained in document D_i he will formulate a request by using I_j.

Library statistics provide the $P(A, D_i)$; the weights coordinated with the index terms, when properly scaled, give estimates of $P(A \cdot D_i, I_j)$; consequently the value of the relevance number $P(A \cdot I_j, D_i)$ can be computed, by means of which documents can be ranked according to their probable relevance to the requestor.

In the technique of Probabilistic Indexing, the result of a search is an ordered list of those documents which satisfy the request, ranked according to their relevance numbers. The request is considered as a clue which the user gives to the library to indicate the nature of his information needs, and which may be used by the library system to generate a best class of documents (to be ranked subsequently by their relevance numbers).

A machine strategy can elaborate upon the basic selection process in order to improve the search in one of two different ways. The first is to establish a measure for "closeness" in request space so as to formulate R', a request similar in meaning to the given request R. The other way is to use the class of documents C, obtained by the initial request R, to define a new class C".

The cross-indexing ("see/see also") aspect of a library indicates some of the relationships that index terms have for one another; i.e., indicates some of the relationships between points in index space. Numerical evaluation of relationships between index terms can be made explicit by formulating probabilistic weighting factors between them. Once numerical weighting factors are coordinated with the distances, the cross-indexing aspect of a library can be mechanized so that, given a request involving one (or many) index terms, a machine could compute other terms for which searches should be made. The elaboration of a request on the basis of a probabilistic "association of ideas" could be executed automatically.

Once the various "connections" between the points of index space have been established, rules must be formulated which describe how one should move in the maze of connected points. Such rules, called "heuristics", would enable a machine to decide, for a given set of request terms, which index terms to "see" and "see also", and how deep this search should be and when to stop, etc.

Besides the two possible measures of closeness, the conditional probability and the inverse condition probability, a third statistical measure, which appears to be the most promising of the three, is one of several possible coefficients of association between predicates.

By "extension heuristics" an initially retrieved class of documents is extended by considerations concerning this class itself. It is preferable not only to extend this class by measures of distance between documents but also to introduce such measures into a "generalized" relevance number computation. In addition the requestor is permitted to assign numerical weights to index terms according to how important a role he wishes them to play in the processing of his request.

In an over-all search strategy the variables involved are: (1) Input: the request R and the request weights; (2) the Probabilistic Matrix $[w_{ij}]$: dissimilarity measures between documents, significance measures for index terms related to the number of documents tagged with the term (the smaller this number the greater the significance of the index term), and closeness measures between index terms; (3) the a priori Probability Distribution; (4) Output: the class of retrieved documents (C), the number of documents in C (n), and relevance numbers; (5) Control Numbers: the maximum number of documents to be retrieved (n_o), relevance number control, generalized relevance number control, request weight control, and significance number of index term control; and (6) Operations: basic selection process, elaboration of the request by using "closeness" in the request space, adjoining new documents to the class of retrieved documents by using "distance" in the document space, and merge.

Two basic hypotheses to be verified are, first, that the relevance number computed for each document, given a request, is in fact a measure of the probable relevance of the document; and second, that the automatic elaboration of a search does, in fact, produce relevant documents which are not retrieved by the original request. A collection of 110 articles from Science News Letter formed the experimental library. Keywords were sorted into categories on the basis of their meanings. Documents coordinated with each category were indexed by assigning to each the names of the corresponding categories, and the degree with which each tag holds for the document was indicated by assigning weights to the index terms.

The problem now considered is: How well does the relevance number perform in ranking documents according to relevance? To relate documents to information need, a bridge is provided between request language and information need through statistical data relating library users with the utility they derive from documents. Such statistical data is given by the a priori probability distribution. It is a probabilistic estimate of the relevance of a document for the information need of the requestor, and is called the "relevance to information need" or "probable relevance".

The goal may be formulated as that of attempting to confirm that the value $w_i(R)$ computed for each document selected by a given request is, in fact, a measure of relevance with respect to the request formulation. If the basic notion is correct, it implies the following hypothesis, H_1: If a document is relevant to a request, then a high number $w_i(R)$ will be derived for it. A number of documents from the experimental library were selected at random and for each document a question was formulated which could be answered by reading the corresponding document. Several persons who acted as test subjects were asked to formulate a library request for information on the basis of which, hopefully, relevant documents would be retrieved. The accession numbers of those documents satisfying the logic of the request were searched and selected. Documents in the lists were ranked according to the number $w_i(R)$ that was computed for each. Each list was examined to determine whether or not the so-called "answer" document was on the list, and if it was its relative position on the list was recorded. The number of times that the correct answer document retrieved was associated with a high number $w_i(R)$ was determined. In summary, 40 library requests were made, and in 27 cases the answer document was retrieved. The number of documents on the output lists ranged from a minimum of one (in four cases) to a maximum of 41. In the majority of the 23 cases which contained more than a single document, the answer document appeared towards the top of the list. Thus, the question arises: "If a document has a high number $w_i(R)$, is it relevant to the request?" This represents the converse of hypothesis H_1 and is formed as H_2: If a document has a high number $w_i(R)$, then it is relevant to the corresponding request.

If H_2 as well as H_1 can be confirmed, an hypothesis H^* which is stronger than each will, in fact, have been confirmed: The methods of Probabilistic Indexing will derive a high number $w_i(R)$ for an arbitrary document if and only if the document in question is relevant to the request. Four test subjects were given the actual documents corresponding to the retrieval lists, and they were asked to read each document and decide whether they considered it to be Very Relevant, Relevant, Somewhat Relevant, Only Slightly Relevant or Irrelevant. In turn their judgments were compared with the numbers $w_i(R)$ which had been computed for each document. The results show quite definitely that if a document had a high number $w_i(R)$ that document was judged by the evaluator as Very Relevant or Relevant, in most cases.

The relevance number, as is seen, provides a means of ranking documents according to their probable relevance. However, the solution to the problem of retrieval effectiveness involves more than ranking by relevance --- it involves the proper selection of those documents which are to be ranked. Two methods have been described for automatically elaborating upon the selection process which is involved in information searching. One method establishes a measure of distance in document space and the other method involves measures of closeness in request space. For closeness in request space there were described three different statistical measures, viz., forward conditional probabilities, inverse conditional probabilities, and coefficients of association. It would be desirable to be able to establish the following: that the elaborated request catches relevant documents which are not selected by the original (unelaborated) request, and that, although the elaborated request catches more documents, the relevance number can be used as a guide for eliminating the ones with low probable relevance.

A substitute type of evaluation consists in using the answer documents as a measure of retrieval effectiveness. Those original requests which did not catch the answer document can be automatically elaborated upon in order to see whether the elaborated request succeeds in retrieving it, thereby establishing some measure of the retrieval effectiveness of the automatic elaboration procedures. Of the 40 requests that were made, the document was retrieved in 27 cases and it was not retrieved in 13 cases. Using the method of request elaboration via forward conditional probabilities between index tags, the correct answer document was retrieved in 32 cases out of the 40. Elaborating the requests via the inverse conditional probability heuristic, the correct document was retrieved in 33 of the 40 cases. Using the coefficient of association to obtain the elaborated request, success was obtained in 33 cases of the 40.

Could the number of answer documents have been improved? In looking at the seven cases for which the answer document was not retrieved when elaborating via the coefficient of association, in three cases the indexing was at fault, that is, the answer document was poorly indexed. In one case the request formulation was very poor and no reasonable elaboration would help. In one case the answer document was caught by a different heuristic and in the remaining two cases, again, the requests suffered by being poorly formulated. Although the automatic elaboration of a request does catch relevant documents that would not otherwise have been selected, it also increases the total number of retrieved documents. The relevance numbers then are used to separate out the highly-relevant from the less relevant documents.

In conclusion it can be observed that to a very large degree the procedures for automatically elaborating upon a request are empirical; i.e., their development and refinement must rest on further empirical testing and experimentation. Hopefully the results of further tests will shed light on and provide new insights into the difficult and intriguing problems of information identification, search and retrieval.

III-19. A Use for the Techniques of Structural Linguistics in Documentation Research, by Jessica S. Melton, Report No. CSL:TR-4, Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, September 1964, 20 p.

(This paper was presented at the F.I.D. Conference on Classification Research and appears in Classification Research (Proceedings of the Second International Study Conference, Elsinore, Denmark, September 14-18, 1964), Copenhagen, Munksgaard, 1965, p. 466-480.)

In order to test the classification or index scheme as a separate component and as a variable, a method must be first found to describe the scheme in isolation from other closely related variables. This paper discusses the feasibility and possible advantages of analyzing and describing various classification and index schemes along the lines of structural linguistics for application to research situations in which the formal characteristics of the schemes must be reckoned with. The term index language is used here to refer to the language of classification and indexing schemes, i.e., their systems of signs and symbols and the rules governing their organization.

Documentalists describe index language either in terms of its logic, or in terms of its performance. Neither of these modes of description, however, focuses on the formal characteristics of the language as language. It would not seem unreasonable that linguistic techniques of analysis and description might yield meaningful insight into the form and structure of index language.

All index languages which combine terms to express various aspects of subject content involve a dictionary and a grammar, i.e., selection of terms from a lexicon and combination of the selected terms according to rules. On this level of description various index languages are structured quite differently. This difference is referred to as pre-coordination or post-coordination of terms. The dictionary of an index language usually orders relationships, makes only certain possible relations explicit, and in most instances does not permit "unauthorized" associations. The present discussion is restricted to the purely structural relations of index language, continuing on to a more detailed description of index language according to the precepts of structural linguistics. A structural model of language appropriate for illustration and consideration of the feasibility of such a description is Garvin's "Definitional Model of Language", in Natural Language and the Computer, Ed. by Paul L. Garvin, New York, 1963, p. 3-22. He defines language as a "system of signs, the structure of which is specified in terms of three sets of levels --- namely two levels of structuring, two levels of organization, and more than one level of integration". The two lower levels, those of structuring, are concerned with elements that constitute the signs of language, and as such have both form and meaning, i.e., the graphic and the morphemic level. The set of levels of integration is concerned with the order of complexity of a linguistic unit. Recognition of the levels of integration would appear to be very important in describing index languages because of their seemingly wide variation on this level. The highest levels, those of organization, are explained in terms of selection and arrangement of units. The relation to analytico-synthetic indexing has been mentioned in terms of pre-coordination and post-coordination of index terms.

A full description of the structure of an index language according to this or any other linguistic model would involve detailed analysis by proper linguistic techniques. If such analyses could be made, the structure of the language system of each index language could be evaluated and various index languages could be compared on a common, well-defined basis. Also, data resulting from such analyses would be useful in studies of inter-convertibility or automation. Index languages suggest a significant difference from natural language on the phonemic and graphemic level, using as they do mixtures of alphabetic, numeric, and special characters many of which have no ready phonemic basis. Accepting such a statement as a linguistic expression amenable to structural analysis would permit its formal structure to be described and compared to that of other index languages. Index languages operate on the remaining levels of structure, integration and organization. Their manner of operation on these levels, however, differs markedly, and this variation may significantly affect the performance of the information retrieval systems of which each may form a part. On the morphemic level an index language which uses a semantic code exhibits a vastly different structure from one which uses numeric hierarchical notation. The elements of a semantic code would seem to measure up to the linguistic criteria for morphemes (i.e., minimal units which convey meaning). But if one looks at a hierarchical classification language in this formal way, an entire notation may be a single morpheme, inseparable and minimal.

A few comments on two index languages on levels higher than the morphemic will illustrate how this type of analysis could indicate similarities and differences in structure. Two index languages have been applied to a portion of an information system for researchers in the field of education, namely, a faceted classification scheme and a telegraphic-abstracting scheme. Concerning the structural effect of order and arrangement in the faceted classification, changing the order of the whole terms changes neither their meaning nor their relationships. Therefore the "grammatical" relationships of this expression are not dependent on the order in which the terms are presented. Changing the characters within the terms, however, would change the meaning of the terms, hence the order of characters representing a particular term must remain fixed. On the other hand, in the telegraphic abstract, changing the order of terms within each interfix would not change the meaning of the expression. Interfixes operate on the level of integration, creating fused-

units of various degrees of complexity. The "grammatical" class of these fused-units is indicated by the role indicator within the unit. In the telegraphic abstract there are analogies also to the grammatical concept, of agreement and dependency. For example, the role indicator-term combination set off by a lower order interfix agrees "grammatically" with the role indicator-term combination set off by a higher order interfix. In the telegraphic abstract the higher order interfix and the role indicator following it designate a "facet" which then relates to other such facets in the same manner as do the facets in faceted classification.

The important point to make here is that description of these index language expressions according to their linguistic structure revealed a similarity which is difficult to describe on another basis. At a certain level of integration the telegraphic abstract becomes a "true" faceted classification. The difference between the two occurs at a lower level where the traditional faceted classification is entirely fixed, or pre-coordinated; the telegraphic abstract is not fixed, not pre-coordinated.

The preceding comments are, of course, not intended to reflect the results of thorough analysis of any of these index languages, but are intended only to demonstrate the types of structure that might be revealed by such analysis. Describing index language from the point of view of structural linguistics would, for one thing, permit description of index language on a common basis regardless of the manner of physical storage. Such description would be restricted to the formal characteristics of index language. Because of these limitations, a structural description of index language would not of itself solve the communication problems between the stored file and the user in an information system, but does seem to offer a technique whereby the entire range of index languages could be described and compared on a definable, consistent basis. A formal self-contained description of index language which would permit this component of an information system to be given a precise numerical evaluation as a variable in an information system is presently being attempted at the Center for Documentation and Communication Research.

III-20. Indexing Criteria, by L. L. Mitnick, Technical Memo No. 64-75, Washington, D.C., General Electric Co., August 12, 1964, 14 p.

There does not seem to be major concern with the criterion problem in the information storage and retrieval area. The lack of explicit criteria by which research can be evaluated may hinder attempts to perform such research. It is time that specialists in storage and retrieval include in their system design activities a comprehensive evaluation program. Criterion as used in this memorandum will signify the variable by which performance may be measured, the dependent variable which is used to demonstrate the effects of manipulating the independent variables.

The ultimate criterion for an information storage and retrieval system is to produce the best possible system. But such criteria cannot be obtained or measured, so related criteria must be specified. Intermediate criteria may include such variables as cost of retrieval, speed of retrieval, etc. Immediate criteria are easier to obtain; they are criteria associated with particular subsystems such as inter- and intra-indexer reliability, indexing rate, etc., for the indexing subsystem of an information storage and retrieval system. One method of initiating research in this area is the identification of the immediate and intermediate criteria mentioned in the literature. For control purposes, attention was focused on a single subsystem, indexing, and, it was decided to search for immediate criteria of indexing and intermediate criteria identified in documents concerned with indexing. The purpose of this preliminary study was to characterize a small sample of published articles in terms of the type of criteria utilized.

From the 10 most recent issues of American Documentation 22 articles on indexing were obtained, and an additional 18 documents from the Indexing Aids Procedures and Devices Contract file were selected. For each article, the criteria were identified without regard to classification in terms and context used by the author. Of the 40 articles examined, only 8 could be considered experimental, and contained some kind of quantitative data. These articles had an average of 2 criteria each. The 40 articles were divided into three groups: 8 articles concerned with immediate criteria only, 16 using intermediate criteria only, and 16 using both criteria. Articles which used immediate criteria only specified fewer criteria than those using intermediate only; articles which mentioned both types of criteria contained the largest number of criteria per paper.

Intermediate criteria include retrieval effectiveness, relating to the ability of the system to perform retrieval in some manner and suggesting that indexing performance may be assessed in terms of how well a system recovers the documents that have been indexed; relevance, referring to the value of the documents retrieved for the particular request and to "false drops" or documents retrieved that have no relevance to the search question;user needs and evaluation, suggesting that a system can be evaluated in terms of ability to satisfy the users or searchers and including searchers' ratings of the system in terms of the index, degree of specificity of access points, convenience of use for searching and number of useless entries; speed of retrieval, the time necessary to perform whatever authors consider retrieval to be, including time spent in formulating a request, in the retrieval process, and in identifying documents with or without delivering the material to the user; cost, as the total system cost; number of steps to locate documents, which may be highly correlated with both cost criterion and user needs criterion, but which is objective and not dependent upon the opinion of the users; searching flexibility, the various ways in which a search may be performed; browsability, permitting searchers to look through the index; and accessibility of documents, the manner in which documents are stored physically.

Immediate criteria include cost, the cost of indexing itself, in terms of a specific portion of the indexing process or the total subsystem; redundancy with a standard, comparing a new classification with the one set up as the standard, measured in terms of numbers of subject headings omitted or percent match between title words and indexing headings or similar methods; indexing rate, documents indexed per unit time; reliability of indexers, consistency of indexers in assigning terms to documents, including both inter-indexer reliability and intra-indexer reliability; ratio of indexing terms and documents, the number of terms associated with each document or the depth of indexing, and the number of entries (documents) related to each indexing term; size of index or dictionary, the number of words produced by an indexing approach; scattering of like information, whereby different entries may carry the same information or documents; and opinion of indexers, which may be used to evaluate an indexing approach.

If frequency of use of criteria is in itself a criterion then retrieval effectiveness mentioned in 42 percent of the articles, user needs and evaluation in 30 percent of the articles, and relevance mentioned in 30 percent are the most important intermediate criteria. The most important immediate criteria and percent of time used are cost, 18 percent; redundancy with a standard, 15 percent; and indexing rate, 15 percent.

None of the authors of the 40 indexing articles appeared to exhibit concern with the criterion problem. While many of the criteria used appear to measure what they are supposed to measure, it is necessary to develop criteria which possess true validity.

III-21. Mechanized Indexing Methods and Their Testing, by John O'Connor, Journal of Association for Computing Machinery 11, 437-449 (October 1964).

Mechanized indexing methods which have been proposed and are being studied require some document preparation, followed by application of indexing rules. Comprehensive text preparation includes input of full text marked for document place of words and symbols, sentence and paragraph divisions, replacement of pronouns with antecedents, and syntactic information for text words; kernelization of sentences; and addition of thesaurus headings, position numbering for words and symbols, frequencies of expressions, closely associated expressions, importance measures of expressions, and reference information. Indexing rules may call for selection of expressions from the original text or added during document preparation, or assignment of terms from a standard indexing list.

The quality of such indexing can seemingly be determined by measuring the quality of retrieval it permits, although retrieval quality is also affected by such factors as cross-references provided, user-searcher consultation, searcher's background, and the like. However, one way to measure retrieval quality in order to determine indexing quality is to have the user decide whether documents presented to him in answer to his query are what he wants and as many as he wants, or are related to his needs. Problems in this connection include the observation that a user cannot always judge immediately the value of a retrieved document. In any case, a retrieval system can be compared with other systems or against a standard. But always the key question is how should the document set retrieved for a question be graded and how should the non-retrieval of the rest of the collection be scored?

An alternative way of investigating mechanized indexing methods might be to duplicate by computer rules the human subject indexing done in a well-reputed retrieval system (cf. III-22). This however is a kind of empirical study rather than a test of mechanized indexing methods.

III-22. Some Suggested Mechanized Indexing Investigations Which Require No Machines, by John O'Connor, American Documentation 12, 198-203 (July 1961).

(Another discussion of the work reported here is contained in "Some Remarks on Mechanized Indexing and Some Small-Scale Empirical Results", by John O'Connor, in Machine Indexing, Progress and Problems, Washington, D.C., American University, February 1961, p. 266-277.)

This brief paper describes some simple explorations concerning the possibilities of mechanized indexing and makes a few comments on their possible outcome. The "equipment" required for the investigations suggested here is a vocabulary of indexing terms, and a document collection to which they have been applied and are being used for retrieval.

To begin the investigations, select a term T from the indexing vocabulary of the retrieval system to be used. Try to think of a simple rule to guide a computer in assigning T to documents. A very simple rule is that the computer should assign T as an indexing term to just those documents in which the word T occurs. The difference is between a term's meaning occurring in a document and a word's physical presence. Distinguish the two cases by speaking of "the term T" when concerned with term T's meaning occurring in a document and speaking of "the word T" when concerned with term T's physical form, its sequence of letters, occurring in the document. The simple computer indexing rule suggested above can now be expressed in the following way: Assign term T to just those documents which contain word T. This rule might over-assign term T. Over-assigning can increase input costs and storage but mechanizing indexing might be worth the added

cost. Over-assigning might also increase the number of irrelevant documents retrieved, but the increase might be insignificant. By such considerations try to make a rough decision about how much over-assigning is acceptable from a mechanized indexing rule for assigning a term.

The computer rule might under-assign, but perhaps the same degree of under-assigning as people indexing documents might be acceptable from a computer rule. If the alternative to mechanized indexing is no indexing at all then a considerable amount of under-assigning might be acceptable. By such considerations try to make a rough decision about how much under-assigning is acceptable from a computer rule for assigning terms to documents. Specifications of limits on acceptable over-assigning and under-assigning may be called the "precision requirements" for a mechanized indexing rule.

Now ask of the simple indexing rule suggested earlier, does it meet the precision requirements? Consider over-assigning. Imagine kinds of documents which might exist in the collection which contain the word T but should not be indexed by term T: the paper might use the word T only to say something like "T will not be considered in this paper"; a paper may contain only conjectures about the subject T; a paper may contain no new information about the subject T; the subject T may only occur in the paper in a subsidiary or incidental way; or aspects of subject T dealt with in the paper are not important for the retrieval system. The number of over-assigned documents found will indicate a lower limit on the amount of over-assigning resulting from the rule.

The frequent word approach suggests a change of the rule to the following: Assign term T to a document if the document contains word T at least f times. Examine the over-assigned documents found in testing the first rule, and see how many remain over-assigned under the new rule for various values of f. This will roughly indicate lower bounds on over-assigning for various values of f.

By making f large enough, one can eliminate all over-assigning. But then to anticipate a bit, the rule might under-assign too much. Thus try to think of a rule of the form, "Term T when word T and X." The question here is, what should X be? In trying to think of modified rules, see if there are any words, or patterns of words, or frequencies of particular patterns of words, which distinguish the documents which have been over-assigned and those which have been correctly indexed. Any rules which look plausible should be tested on the over-assigned documents found earlier to see if they appear to decrease over-assigning sufficiently.

This paper suggests the trying out of relatively simple possible rules for mechanized indexing. It is possible to get useful information about the effectiveness of such rules of the kind described in this paper. It is well to remember that we have been talking about mechanizing the assignment of terms of an existing indexing system, not the development of a new indexing system which would both permit mechanized assignment of terms and would work well for retrieval.

Now consider under-assigning. For the study of under-assigning think of some kinds of T-documents which might not contain the word T: a paper contains a synonym or near-synonym of word T; a document is about a scientific phase of the subject T; the paper is about a subject related to subject T, though not a specific phase of it. Try to find among the T-documents some which do not contain the word T, the "under-assigned documents" for the rule "Term T when word T." Think of some possible modifications of the rule which will decrease the under-assigning sufficiently. One familiar idea which seems natural is to make a list of synonyms and near synonyms of word T. Then a natural rule to consider is, "Term T when any T-synonym." It will not decrease the number of over-assigned documents, and, more important, it may increase them. If the T-synonym rule

does not under-assign too much but does over-assign too much, rules to be considered will be of the forms, "Term T when T-synonyms at least f times," and "Term T when T-synonyms at least f times and X." Suppose the T-synonym rule does not over-assign excessively, but does under-assign too much. Then examine the T-documents which are not assigned term T by the rule, and try to modify the rule to allow for them. With a list of words and phrases, "term-T-indicators," which indicate specific phases of T or subjects related to T, consider the rule, "Term T when term-T indicators at least f times." The form of the term-T-indicator rule is that of rules for mechanized indexing by "thesaurus notions."

Suppose a rule being considered is found to both over-assign and under-assign too much. Then it may have to be modified in several different ways, for a single change cannot decrease both under-assigning and over-assigning. Either several such changes together are needed or more complex word patterns must be tried. If no satisfactory rules are found for assigning term T mechanically, then a few other terms might be tried.

Various useful compromises between human and mechanized indexing might be suggested by the results of the investigations. Such results and knowledge of the retrieval system might suggest ways in which the vocabulary, or the procedures for using it, can be changed to make mechanized indexing possible without making retrieval poorer.

III-23. Defining the Query Spectrum --- The Basis for Developing and Evaluating Information Retrieval Methods, by James W. Perry, IEEE Transactions on Engineering Writing and Speech EWS-6, 20-27 (September 1963).

Methods of information retrieval (IR), regardless both of their design and of types of equipment used, serve the general purpose of directing attention to certain documents in extensive files and in libraries. Human effort or machine operations or various combinations of the two may be used to bring about selective interaction of some query (which may be denoted by Q) with the total file of documents (denoted by $[D]^N$ for a file of N documents) or with some set of representations (denoted generically by [S]) for individual documents or for groups of documents. The IR system can be considered to respond to a given query, Q, by selecting a subset, T, of documents from the total file $[D]^N$. The degree of capability of an IR system to provide satisfactory service for the queries that constitute a given query spectrum (denoted by $[S]^E$) is, of course, essential in evaluating benefits. It is essential that the query spectrum, $[S]^E$, shall be defined in terms of practical information requirements rather than derived from logical principles. Accordingly, before initiating the design of IR methods for application in a given practical situation, the latter should be studied carefully for the purpose of understanding what kind of query spectrum is to be serviced. This paper discusses the role of recorded factual information in the general area of scientific research, and concludes that there are a number of important factors whose qualitative importance is undeniable but whose precise definition and quantitative measurement present severe difficulties. Corresponding factors are identified in Patent Office operations which offer the possibility of more explicit statement and more precise measurement of such factors. Study of IR methods in these operations could be helpful not only in optimizing Patent Office novelty searches but also in optimizing IR methods for technological development and for scientific research in general.

This paper is concerned with the question of arriving at a set of queries which would generate responses of optimum value assuming, of course, that it is possible for some IR system to provide such responses at reasonable cost. An IR system can provide maximum value of responses only if two conditions are met. The queries must be such that responses provide information of maximum usefulness in dealing with practical problems and

situations, and the IR system must be able to respond to such queries by appropriate selection of information. These two conditions suggest kinds of value rating of the different costs incurred when different information systems are used to select information from a given file. Under the cost ceiling imposed by the value of information provided, the best system is the one that provides as much useful information as possible at minimum cost. In considering the specialized knowledge that a professional man applies to his problems, let us note that different sources may be involved. In competitive situations, the extra effort required for better --- or optimum --- solutions may result in gratifying rewards. Information retrieval methods serve the purpose of making information available to professional mem when they need it. Research assignments of broad scope involve a sequence of phases: exploratory research, preliminary investigation of different possible solutions, detailed investigation of the approach selected as most promising, transition from laboratory stage to pilot plant to full scale manufacturing, customer service, product and process improvement. In each of these stages, diversity of queries may be generated. This diversity is of prime importance to the designer of an IR system, whose ability to respond to a broad spectrum of queries is decisive in determining its practical utility. A professional man generates a query to obtain additional information which he brings into interaction with his previous knowledge to arrive at a decision. The problem posed to the professional man is not of itself the source of the query. Rather, it is the need for information on the part of the professional man, and his previous knowledge is a most important factor in determining this need.

An IR system that functions with a library or with a file of documents is not the only possible source of needed information. It is a problem of research management to make the best use of these various sources of information, in such combination and coordination that the most useful solutions to research assignments may be achieved with minimum expenditure of funds budgeted for research. In dealing with an assignment, information may be needed from some specialized field concerning which a given professional man may be informed to such a slight degree that he scarcely knows what information to ask for. His first query may have the purpose of acquiring general background information. Eventual queries will, of course, resemble those that an expert in the given specialty would have asked in the first place. There is a further factor that is of importance in determining what may be an optimum query to address to a given file of documents. It is often the case that directly pertinent information is not contained in the document file or library to which queries may be addressed. The best available information may be of less direct nature. Considerable imagination and intuition may be required to work out useful queries based on analogies. These considerations tend to broaden the spectrum of queries with which an IR system may have to deal.

The total spectrum of useful queries --- actual and also potential --- should be taken into account when designing and evaluating IR methods and systems. The novelty searches as made in the Patent Office generate a wide variety of queries that are similar, in a number of important respects, to those of practical importance in Research and Development activities. The provisions of patent law permit the scope of claims to vary tremendously, with the consequence that the investigation of the novelty of the subject content of patent claims generates queries of corresponding variation in scope. The Patent Office IR system must respond by directing attention to a certain set of documents. An IR system, set up to serve Patent Office requirements, may be highly useful in a practical sense even if the subsets of documents as provided in response to queries include some documents that are of no interest to the examiner in arriving at a decision as to the patentable novelty of the various claims of an application. The question: "Just what documents should be provided in responding to a given query so that the examiner will spend no time at all in reviewing documents of no interest with respect to a given application or --- more specifically --- a given claim in some application?" may be approached from two different points of view: (1) What was the reliability with which pertinent documents were included

among the actually selected subset T? and (2) What proportion of nonpertinent documents were included in the subset T? A key consideration is the distinction between pertinent and nonpertinent, here virtual synonyms to "useful" and "not useful" in resolving a given problem. The subject of allowable patent claims must not only be novel, in the strict sense of the term, but must also constitute a significant advance above the level of previously known, related developments. In order to decide whether the subject matter of a patent application meets this "inventive level" requirement, a patent examiner must take into account additional subject matter that is outside the strict scope of the claims yet sufficiently closely related to require consideration in deciding on the inventive level --- and hence the patentability --- of one or another of the claims in an application. There is considerable similarity between the query --- or queries --- that a patent examiner must formulate to retrieve information relevant to novelty and inventive level and the query --- or queries --- that would be optimum to obtain information useful when carrying through a technical development.

It is not intended to imply that the Patent Office IR systems are ideally perfect sources of experience and data for developing a comprehensive mathematical model for IR methods in general. Perhaps the single most serious defect is a consequence of the fact that scientific theories are not patentable. It would be necessary to extend the Patent Office range of queries in order to arrive at a spectrum covering both the purer forms of research as well as technological development and applied research. Another limitation inherent in Patent Office queries is that the application describing an invention as well as documents that are of interest in deciding questions of patentable novelty and inventive level will --- indeed must --- provide substantially complete descriptions of the subject matter. This limits the extent to which the Patent Office IR system --- as a system --- may serve as a prototype for systems that must deal with fragmentary information.

III-24. The Evaluation of Retrieval Systems, by Alan M. Rees, Report No. CSL:TR-5, Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, July 1965, 21 p.

The rapidly growing interest in testing and evaluation is doubtless inspired in part by the belief that the provision of data concerning the efficiency and effectiveness of retrieval systems is essential to progress in system design and operation. The subject holds considerable appeal to the many persons charged with the management of operational systems, for it is still most difficult to select the optimum system for any situation or to know how to improve an existing system. Despite the amount of literature on the subject published the significance of work accomplished to date is far from clear. The essential issues continue to worry us and cannot be adequately answered at the present time. Some of them will be discussed in this paper.

Effectiveness of systems, which we may define as the system's ability to perform the task for which it was designed, is usually measured in terms of relevance. Yet the subjective assessment of relevance on the part of individual users may have nothing to do with the stated objectives or purpose of the system. Relevance is an unstable and dynamic phenomenon which reflects subjective responses on the part of individuals. We are faced with a low level of consistency in assessments among and between relevance judges. The use of relevance as a criterion of the effectiveness of systems is open to considerable criticism. For example, it has been stressed that there is a distinction between relevance to a request statement and relevance to an individual's information need. However, research to develop a methodology for assessing the impact of a retrieval system upon the cognitive processes of users has yet to be performed. No practical alternative to the use of relevance exists at the present time. Various measures of effectiveness with relevance as the criterion

have been proposed. Recall ratio, precision ratio, sensitivity, specificity are examples of measures. The best known are recall ratio and precision ratio used by Cleverdon in the Cranfield Project (cf. I-2). Alternative measures proposed by Goffman and Newill and being applied in the Western Reserve University Comparative Systems Laboratory (cf. III-9) are sensitivity (equivalent to Cleverdon's recall ratio) and specificity. Specificity takes into account one of the vital parameters in a retrieval system --- size of file. Effectiveness is a combined measure of sensitivity and specificity.

It is to be noted that all of these measures presuppose that a dichotomy between relevant and non-relevant can be adequately achieved. A distinction must therefore be made between the process of relevance assessment and the measures employed to compute such assessments.

The experimental design of tests constitutes another area of considerable difficulty. There has been a deplorable lack of definition accompanied by deficiencies in experimental design and control. It is often difficult to determine what has in fact been tested and under what conditions. Experimental design is of great significance considering that an information retrieval system is a complex assembly of interacting components (sub-systems) each of which in turn contains a set of interacting variables. Sloppiness in experimental design to the extent that variables are not identified and controlled vitiates the value of tests. In the same manner that variables must be controlled within a sub-system, the variables operating within the total retrieval system must be identified and held constant. Any of the variables identified in any of the sub-systems may be subjected to experimental observation and measurement. In this case all other variables must be held constant. Meaningful comparison between systems or sub-systems or a test of a particular system or sub-system necessitates similar control to isolate the sources of variation in performance.

This discussion is intended to emphasize the considerable difficulties involved in providing valid quantification. Experimentation in information science is becoming a specialty in itself. Researchers work with models, construct hypotheses, design and execute experiments, refine models, formulate further hypotheses and so on. Such research will eventually furnish useful tools for those providing information services. Application of the findings of retrieval systems tests, up to the present time based almost exclusively upon simulation under research conditions, is limited. More consideration should be given to perfecting a methodology for testing and less to the attempt to generalize performance data beyond simulated experimental contexts. We do know that there are no intrinsic properties of systems which determine their performance in all situations at all times for all users. Their performance must be determined within a total system in relation to a specified purpose, with an identified group of users with known information needs in a given environment at any one time. Those entrusted with the management of operating systems must assess the effectiveness and efficiency of systems as ingeniously as they can. The use of a variety of testing techniques is advocated. There is no short-cut to the achievement of adequate evaluation.

In the next few years research will become deeply concerned with experimentation with the result that experimental tools and methods will be created. The research in testing and evaluation will steer away from investigation of the operating mechanics of retrieval systems and concentrate more on the psychological and sociological factors involved in the communication process. Relevance will be studied and measured in relation to these tasks along a continuum instead of as a dichotomous decision. Such research will hopefully provide data on the impact of information services on the research productivity of scientists and engineers. This insight will in turn provide superior criteria and measurements than those based on present conceptions of relevance.

III-25. Problems of Evaluation, in Automatic Indexing: A State-of-the-Art Report, by Mary Elizabeth Stevens, NBS Monograph 91, Washington, D.C., U.S. Department of Commerce, National Bureau of Standards, March 30, 1965, p. 143-163.

If automatic indexing procedures are to be based upon previous human indexing or if their results are to be compared with human results, then the questions of the quality, the reliability and the consistency of human indexing are crucial ones indeed. The most generally accepted criterion for appraising the effectiveness of indexing is that of retrieval effectiveness. But, in general, this is merely the substitution of one intangible for another. We shall try to distinguish here between the core problems that make the evaluation of indexing as such an extremely difficult task, the available data on human indexer reliability, and the possible advantages and disadvantages of automatic indexing techniques.

First and foremost of the core problems implicit in the question of evaluation of any indexing scheme are those of interpersonal communication itself: first, the problems of language as a means of communicating and secondly, the question of language representations of real transactions and events. A second core problem is the heterogeneous and somewhat arbitrary development of natural languages themselves. The problems are aggravated if men themselves must know enough about language and its conveyances of message content to specify precisely to a machine what is is to look for and to use. A third core problem is the proper choice of appropriate selection criteria if condensed representations of document content must be used for scanning, search, and relevance decisions. The substitution of machine-compiled or machine-produced condensation alternatives may make the problem of how adequate the user judges the selection and condensation to be that much worse. A fourth problem in evaluation, therefore, is the question of whether or not the benefit to users is worth the cost. At least some data on the use made by scientists of various sources of information on material which might be of interest to them suggests that subject indexes are not the most important source, nor even a major source. These data suggest that KWIC type indexes may be adequate for many purposes. The problem here involves the lack of information on indexing costs, user needs, and the matter of defining "interest" for different users with differing purposes and requirements. A final core problem, then, is that of the question of relevancy itself. The problems of how to measure relevancy remain largely unresolved.

Since the evaluation of quality of indexing per se raises such fundamental and elusive questions, obvious bases for evaluation are those of time, cost, availability of alternative possibilities, and customer acceptance. The Cranfield project (cf. I-2) has attempted to compare different indexing systems against proposed measures of "retrieval effectiveness" and the relevance ratio. In spite of the fact that the Cranfield tests have so far been directed principally to indexing systems applied manually, certain findings and conclusions reached by Cleverdon and his associates are pertinent to the questions of evaluating automatic indexing procedures. O'Connor has cogently observed that the problem is whether or not indexing by machine is capable of producing results that are "good enough" for retrieval purposes, raising in its turn the still more basic question of how "good retrieval" can be evaluated (cf. III-21 and III-22). Typical points made by O'Connor include the possibilities that the use of automatic indexing techniques might free trained technical people for other work, that it might permit more indexing than is now possible with available resources, that it might cost less, and that it might produce a better or more consistent indexing product. The paucity of objective data on the effectiveness of indexing systems generally extends to even such obvious questions as costs of indexing and time required to index. Insofar as such meagre data is indicative, there does not appear to be any particular cost-advantage for machine-compiled and machine-generated indexing other than the title-only KWIC indexes.

In view of the difficulties the question of evaluation of automatic indexing procedures largely reduces to the weighing of potential advantages and disadvantages. Suggested bases for evaluation may be summarized as speed and timeliness, relative economy, consistency and reliability, elimination of the need for further human intellectual effort after initial planning and programming have been done, providing a product that could not otherwise be obtained, ease of updating and revision of indexes so produced. Thus we may determine that an automatic indexing procedure produces a product at least as rapidly, at least as inexpensively, at least as consistently as human indexing operations would, and with substantially less investment of manpower resources. However, will this product be as useful or as "good" as the human product? It is important to recognize that we should compare the products of automatic indexing methods with the average, the routine output of any other indexing system, including the critical question of how well and how consistently the system, whatever it is, is applied in practice by the human analysts.

Very few objective studies have as yet been made of inter-indexer and intra-indexer consistency. There can be little doubt that the quality and consistency of most human indexing is not good. On the other hand, today's indexing, whether accomplished by man or machine, is probably no better and no worse than any other classificatory or indexing procedures. The only excuse, therefore, for choice between man and machine is the cost/benefit ratio.

The difficulties and problems of evaluation so far considered are generally applicable to any indexing system, whether manual or automatic. Certain special factors arise, however, when we consider some of the proposed automatic assignment and automatic classification techniques. These factors include the question of the amount of computation required in the inversion and other manipulations of large matrices and the problems of how large a vocabulary of clue words can be used effectively and of whether some documents cannot be indexed at all because they contain none of these words. There are serious questions of whether groupings can be conveniently named or displayed for the benefit of the user.

Suggested bases for evaluation made possible by machine processing include proposals (cf. III-7) that substitution for the elusive concept of "relevance" of criteria be based on "sharpness of separation of exploratory regions." Edmundson, however, points out (cf. II-27) that while there is in general only one translation of a document, there may be as many abstracts as there are users. Thus we are back again at the questions of purpose and relevance.

III-26. Interrogating a Computer in Natural Language, by Don R. Swanson, in Information Processing 1962, Ed. by Cicely M. Popplewell, Amsterdam, North-Holland Publishing Co., 1963, p. 288-293.

The purpose of this study is to investigate the problems of fully automatic indexing and retrieval of information. Two essential aspects, full text input and evaluation of retrieval effectiveness, are incorporated into the experiments. Success in retrieving information depends on the human ingenuity exercised in formulating a search instruction. It was concluded, after studying such search instructions to determine their intellectual content, that the process of translating the original question into an instruction could have been done better by a machine following a systematic and consistent procedure and taking into account synonyms and approximate equivalents. If the task of transforming the question into an instruction could be mechanized successfully, one could interrogate a computer in natural language and cause the computer to search natural-language text for information. The significance of this study lies in the insight into fundamental problems of indexing and information retrieval which one can gain.

In an extremely simple version of the process of natural-language interrogation, the computer is instructed to search for those words in the text of a document identical to the words of a question; the search can be carried out by a simple matching procedure. The technique could be improved by providing the machine with a synonym dictionary, or thesaurus, and with some kind of capability for distinguishing between words which are potentially important for retrieval purposes and those which are not. Unfortunately these techniques involve hidden problems difficult in concept. Relationship of synonymity between two given words cannot necessarily be specified in the absolute sense independently of context. Similarly words can be important for retrieval purposes in some contexts and unimportant in others. Quite distinct are problems of syntax. Co-occurrence is the weakest form of syntactic specification and can lead to irrelevant retrieval. Earlier studies led to interesting empirical results on the usefulness of proximity as a stronger substitute for syntax than simple co-occurrence within an entire article.

The experiments which were carried out tested the following hypotheses: for a particular type of collection, a thesaurus of nearly equivalent words can be constructed such that equivalence is largely independent of context; these thesaurus groups can be coded to reflect varying degrees of importance independently of context; and syntactic relationships can be approximated by specifying proximity of terms within certain spans. The thesaurus for machine use was so constructed. In the machine procedure for processing the question and search instruction, each document searched was assigned a relevance score dependent on the number of terms present and on their weights. In addition, the computer was programmed to assign a premium to the final relevance score of a document for each pair of terms in the search instruction that occurred contiguously in the document. The output from the search procedure consisted of a list of article numbers sequenced according to relevance score; with each number was a list of those words in the search instruction which were found within that article. A second list of article numbers was also printed out to indicate the relevance scores based on human judgment. The experimental library of articles on nuclear physics was searched for each of 60 questions, of which 50 had been used in earlier experiments (cf. II-53). For the same amount of irrelevant data, the percentage of relevant information retrieved in the fully mechanized procedure was somewhat higher than with the manually formulated search instructions. The results were also compared with those using a revised subject index for conventional retrieval; the conventional subject heading method was less successful than fully automatic text interrogation.

III-27. Cost as the Measure of Efficiency of Storage and Retrieval Systems, by Mortimer Taube and Associates, in Studies in Coordinate Indexing, Vol. II, Washington, D.C., Documentation, Inc., 1956, p. 18-33.

The physical form of diverse systems does not affect their basic potentialities for handling different types of searches. If all systems have comparable indexing possibilities, then all systems can exhibit the same degree of reference adequacy. Criteria which then distinguish one system from another are input costs and output costs measured in labor, time, and equipment. This is a basic notion that cost factors must enter into the measurement of reference adequacy. Thorne (cf. I-33) measures the efficiency of an information system by the cost of handsorting an entire collection when the system fails to disclose an answer to any given question. His formula assumes that input and searching costs are identical or nearly so in all systems. If however we assume that systems will respond equally well to any reference question but that input and search costs are higher in one than in the others, then we could measure efficiency in unambiguous dollars rather than in subjective assessment of reference adequacy. The measurement of the efficiency of a system in terms of cost requires the assumption that all systems can be made equal in reference adequacy. This assumption is in accordance with the general statement that all physical

systems are alike with respect to indexing possibilities, searching possibilities, or reference adequacy. It is simpler to compare the costs of indexing each document by an equivalent number of desirable access points in various systems. We must deal with input costs and output costs, fairly definite operations which can be reduced to dollars per unit of time or equipment. We can go on to measure hypothetical costs, the cost of answering any hypothetical question put to the system. The only hypothetical questions we should consider are those in the mind of the indexer as he indexes an item. But if the indexer is not constrained, by economic considerations or limitations in coding space or such, then the cost of answering hypothetical questions is not a basic factor in evaluating different storage and retrieval systems.

III-28. A Note on the Pseudo-Mathematics of Relevance, by Mortimer Taube, American Documentation 16, 69-72 (April 1965).

There seems to be little agreement on what relevance means. There is, nevertheless, a growing agreement that a fixed and formal relationship exists between the "relevance" and the "recall" performance of any system. We will find in the literature the phenomenon of shifting back and forth from an admittedly subjective and non-mathematical term to equations in which the same term is given a mathematical value or a mathematical definition. Use of a single term in the same document to cover two or more distinct meanings represents a more serious situation than merely careless ambiguity.

In this paper the demonstration of the shift from subjective relevance as a reaction of a user to mathematical relevance as a property of systems will be restricted mainly to the work of the Cranfield Studies (cf. I-2) and the Arthur D. Little "Report on Centralization and Documentation" (cf. IV-3). Criticism of the former may be directed to the point that relevance as defined by the studies is not measurable. The Cranfield studies not only recognized the difficulty of measuring relevance, but also indicated that there was a very real question concerning the ability to recognize a relevant document. The studies nevertheless concluded that it was possible to test each index language device to get precise figures for their effect on recall and relevance. A vague, hardly recognizable and admittedly difficult notion turned out to be precisely measurable.

In the Arthur D. Little work discussion of relevance and recall is introduced by a statement to the effect that "... 'relevance' is at best a hazy notion." But this hazy notion turns up later in the report carried out to the third decimal place.

Having established that the shift does take place in the Cranfield work and the Arthur D. Little report, the next task is to analyze how it takes place. The axioms of set theory are falsely applied to subjective judgments of relevance. The Cranfield studies used the term "relevance". This usage implies that there is in a particular collection, a single relevant document so characterized in advance of the search. Such a measure has no relationship to the presumed desires of a searcher for items other than the initially characterized document; nor is there any question of the percentage of relevant documents in the collection since, by definition, there is precisely only one relevant document for each question. Discussion now concerns not a known relevant document but any relevant document. This being the case, the expression $\frac{R}{C} \times 100$ does not have a fixed mathematical result because the new meaning of relevance, R, has not been defined. Further, if R has no mathematical value in the expression which gives the recall ratio, then it has no mathematical value in the expression $\frac{R}{L} \times 100$ (where L is the number retrieved in a search) which gives the relevance ratio.

If an attempt is made to determine exactly what is characterized by the term "relevance", it becomes very clear that at least two meanings of relevance are being used and that only one of them has been defined. When relevance is used to characterize a document or characterize the relation between a document and the subjective requirements of a searcher, we may accept the supposition that the searcher uses the term relevance with fair consistency in referring to his own responses. When it comes to asserting a consistency of usage for all searchers, we are on much shakier ground.

When, on the other hand, discussion shifts to the "relevance-recall" ratio, relevance is assuming to characterize or describe systems as a whole. It is supposedly an indexing or I-R system as a total complex which is characterized or ranked according to its relevance and recall performance. A subjective response may be trusted to determine that a certain percentage of items retrieved are not "relevant", but such a response cannot establish a percentage of recall because the subject does not know what other documents there are in the collection which might be relevant to his question. Hence "recall" performance of a system must be defined without reference to the subject, as a characteristic of the system itself. Psychological variation of subjective responses is not regarded as important because of a presumed relevance-recall ratio which is presented as an objective property of systems and not an empirical measure of subjective response.

This phenomena of the almost universal acceptance of an undefined ratio and the conclusion that there has finally been achieved a measurable objective property which enables the precise comparative evaluation of systems, as we have said, arises from a confusion between certain axioms of set theory with these demonstrably vague concepts of recall and relevance. In general, the larger the set retrieved the more likely it is that the answer will contain irrelevant material, and the smaller the set retrieved the more likely it is that relevant material will not be recalled. The art of searching is to formulate a specific question which isn't too specific, and the art of indexing is to use class designations consistently so that the searcher can have confidence in his art.

The mathematical results of the Cranfield studies and the Little report must be considered as pseudo-mathematics. The conclusions of these reports, namely, the claim of the Cranfield studies to have discovered a mathematical ratio which will permit the precise evaluation of systems, and the Little conclusion that large coordinate indexing systems must inevitably break down, are alike without substance or merit.

III-29. On Criteria for Evaluating Information Retrieval Systems, by Bjorn V. Tell, et al., Trend Report (submitted by Swedish Classification Research Group to the 29th FID Conference in Stockholm, 1963), Bromma, Sweden, Tekniska Litteratursallskapet, 1963, p. T1-T8.

Evaluation criteria need to be established for total information retrieval systems rather than for components of systems. Study of evaluation criteria can be restricted to an optimization of the performance of a system between the two bounds of retrieving all pertinent but some irrelevant documents and retrieving only but not all pertinent documents. Time and costs alone do not constitute the decisive criteria for evaluating a system. A statistical description of a system performance may be obtained by subjecting to statistical analysis the reliability of performance of some components of the system. However, it is difficult to specify the components which can be so treated. Alternatively, a useful approach would be to establish a comprehensive list of parameters of an information retrieval system; Mooers (cf. IV-27) and the Andersen Co. (cf. IV-2) suggest some. The parameters could be fixed in a matrix and allowed to vary with the aid of a computer. Or, using game theory, chance events can be allowed to creep in to the matrix, as suggested by

Hillier (cf. III-12); the aim is to raise the probability that the user will get the right pieces of information. Supplying right information to the user suggests a criterion of user satisfaction. Models have been used for cost analysis programs as developed by Bourne (cf. IV-11). An approach to evaluation has also been made in the Cranfield study comparing four different systems (cf. I-2). A sound basis for judgment has to be provided in the form of a set of criteria for evaluation of information retrieval systems.

III-30. Inefficiency of the Use of Boolean Functions for Information Retrieval Systems, by Jacobus Verhoeff, William Goffman and Jack Belzer, Communications of the ACM 4, 557-558, 594 (December 1961).

A retrieval system's problem of optimizing its average performance is difficult, because to give the same answer to different people as a response to the same request provokes different reactions; the system cannot make all its customers happy, but must try to do as well as possible. Consider one customer asking for a reference list of documents, P_1, which he considers relevant to his question. Let A be the list actually given by the system as its answer. The intersection of A and P_1 is the relevant part of A. If different customers submit the same question, they consider lists $P_2 \ldots P_k$ to be relevant. Then if these lists are not identical, the system will yield different values for a measure of merit $M(A, P_i)$ on the different occasions. The problem is to select A in such a way that $M(AP_1) + M(AP_2) + \ldots + M(AP_k)$ is maximal. This can be done by defining a critical probability for each document relevant to each question. All documents with a probability above the critical one should be included in the answer. If documents with probabilities below the critical one are given, they will on the average be more of a nuisance than a help.

But if the system responds to a request for "a" and "b" by giving the intersection of responses it would have given to requests for "a" and "b" separately, it risks giving too much irrelevant material. If in response to a request for "a" or "b", it gives the union of the responses to requests for "a" and "b" separately, it risks leaving out relevant material. In both cases a decrease in efficiency results.

III-31. A Discriminant Method for Automatically Classifying Documents, by John H. Williams, Jr., in AFIPS Conference Proceedings, Volume 24, 1963 Fall Joint Computer Conference, Baltimore, Md., Spartan Books, 1963, p. 161-166.

The objective of a classification technique is to classify a document into one or more subject headings, to connect related subject headings into a tree structure. Classification of incoming documents is then performed at each level of the structure. The purpose of the present study has been to examine alternative techniques and accumulate empirical evidence with which the techniques can be evaluated. The study has revealed many parameters affecting the automatic classification of documents. A few of the problems that arise immediately are: the non-normal distribution of word frequencies, the fact that many documents cannot be categorized into mutually exclusive classes, the varying depth of detail in documents, the homogeneity of subject headings, and the imprecise definition of word populations. The discriminant method consists of starting with a hierarchical classification structure and a small set of reference documents previously classified into each category by human indexers. All the words are counted in each of the reference documents, and theoretical frequencies for each word type are computed for each category. The most significant words in each group of categories are selected statistically. Each subject is represented by the theoretical frequencies of its most significant words. An incoming document can then be classified by comparing the observed frequencies of each word type in the document with the set of frequencies representing each subject. This statistic, the Relevance Value, is computed for the document with respect to each category.

For the experiment 400 computer abstracts, classified by professional indexers for the purpose of actual retrieval, were selected. The classification structure consisted of fifteen categories at the major level. Each of these major categories was divided into 10 subcategories. For the actual experiment a truncated structure was generated based on the criterion that each subcategory must comprise at least 20 documents. Each of four major categories was subdivided into five minor categories. Then 300 of the 400 abstracts were used as reference documents, and were equally divided among the twenty subcategories. The remaining 100 were used as the test documents. The objective of the experiment was to classify the 100 test documents into their correct categories. The key problems in all statistical approaches to classification and indexing center on the selection of the set of most significant words. In the present approach a statistical technique is developed for identifying the insignificant words. Because words change in their discrimination power with the context, it is necessary to compute a discriminant coefficient at each level and within each group. Once the discriminant coefficient has been computed, it is used to set up discriminant thresholds determining which words will be used in the classification equation and to assign a weighting factor to the word itself. The computer program classifies documents by comparing the observed with the theoretical word frequencies and computing a Relevance Value (RV) for each document with respect to each category.

Of the 100 test documents initially selected, 17 were not completely indexed within the experimental structure. Therefore, complete results are available on only 83 of the original 100 documents. Type I documents were classified into only one category at both the major and minor levels. Type II documents were classified into one category at the major level, and two categories at the minor level.

An additional series of experiments were conducted to investigate the value of the discriminant coefficient in the classification process. Different uses of the discriminant coefficient may facilitate the classification decision by causing a greater dispersion of RV's. In all these experiments essentially the same number of correctly classified documents was maintained while the confidence with which the assignment can be made was increased. Two of the major reasons for misclassification were heterogeneous categories and the small sample sizes.

The present approach considers the problem of definition of a category. In an operational situation the categories are determined by the total subject content of the documents. A group of documents is selected by the user as representative of what should be contained in a category. Each document would be entered into the classification program and a Relevance Value (RV) computed for it. Those having an RV outside the standard deviation limits would be returned to the user for a re-evaluation. The user also specifies the relationship in which the subjects are to be connected, eliminating the fixed viewpoint of certain classification schemes. A classification structure in connection with the discriminant method can also be used to advantage in the query phase of an information system. The user has the option of requesting all documents within a subject heading or writing a narrative description of his request, which would be input as though it were an incoming document to be classified. Relevance Values would then be output with respect to each category in the system. The user can choose the categories in which a search is to be performed on the basis of the highest Relevance Values.

SECTION FOUR

PROPOSALS

Many suggestions for methods and techniques of evaluation were made in the papers included in Sections 1 and 2. Such suggestions were applied in testing programs, and results reported, or were developed as a consequence of testing programs, in which methodological needs became apparent. In Section 3 there appeared discussions of criteria that should be applied in testing and of system parameters that should be studied for evaluation purposes. Such discussions were mainly abstract and did not include test details or results.

In this section further suggestions are made for evaluation methods and techniques, more concrete than those found in Section 3 but usually not applied as fully as in Sections 1 and 2. That is, proposals are made for methods of evaluation which have been developed in detail and which are suggested as broadly applicable to operating or experimental systems. These proposals range from relatively simple processes to highly complex models of performance. In some cases the methods have been applied, via computer programs, to simulations of operating systems.

In most cases the proposals are for techniques for the evaluation of total system performance. However, some suggestions apply to subsystems or components. Thus, methods for organizing files to improve retrieval performance are proposed and tested (IV-7 and -30, for example). Languages used in framing requests to information systems are the subject of other proposals (IV-17 and -18, and -24). Again, aspects of the indexing process are considered in IV-21.

Many of the techniques for assessing total performance include equations by which to calculate a measure of effectiveness. Such quantitative measures are proposed as means for evaluating one system's performance or comparing the effectiveness of several systems.

IV-1. Proposed Scope of Area 5, in Proceedings of International Conference on Scientific Information, Vol. II, Washington, D.C., National Academy of Sciences-National Research Council, 1959, p. 817-821.

This area of the Conference was concerned with the design of effective systems, with the problems encountered in processing recorded information for subsequent search, and with the possibilities of using machines or devices in the processing, storage and search operations. The processing of information for retrospective searching must take into account the purpose to be served, the capabilities of equipment, and the costs involved in applying the various methods of processing. Not only the effectiveness of traditional procedures but also of recent experimental work directed toward the development of new or modified methods of organizing material should be considered. The principal subjects for discussion in the area might include comparative evaluation of experimental systems. Persons and groups engaged in the development of systems embodying new principles for organizing subject matter could compare their systems by applying them to a common set of documents. Use of the same sample would make comparison and evaluation easier. The systems should be discussed in terms of size of the file to be covered, rate of growth of file, purposes to be served, range of subject matter, kinds of concepts to be represented, specificity and type of analysis, personnel required, cost of processing information and conducting searches, reliability of results, and form of system.

IV-2. Research Study of Criteria and Procedures for Evaluating Scientific Information Retrieval Systems, by Arthur Andersen and Company, Report to the National Science Foundation, New York, Arthur Andersen and Co., March 1962, 103 p., plus appendices.

(The National Academy of Sciences-National Research Council ad hoc Committee, charged with evaluating the American Society for Metals-Western Reserve University Metallurgical Searching Service, recommended that operations research studies be made on criteria for evaluating information retrieval systems. Two limited, exploratory studies directed toward the development of measures of effectiveness were prepared by Arthur Andersen and Company, reported here, and Stanford Research Institute (cf. IV-11).)

Evaluation of scientific information retrieval systems depends upon knowledge of the objectives of such systems, identification of the mechanics of the systems, and a method of making common performance measurements on various systems. All scientific information retrieval systems consist of four stages: coding and indexing objects for storage, coding inquiries in the same language as that applied to objects, matching inquiries with objects, and appraising retrieved objects. Criteria for evaluation should be objective and should include cost, time to process an inquiry, and volume of usage; in addition means to simulate performance should permit critical examination of competing systems.

To these ends, a cost-time-volume model was developed and tested. Two distinct operations were represented: the input cycle and the inquiry cycle. An operating statement of a system records cost, time, and volume data associated with each element of the system, under nine headings: encoding objects, inserting them in storage, encoding the inquiry, preparing it for search, searching the store, identifying retrieved objects, appraisal of search results, obtaining source objects, reformulation or withdrawal of the inquiry. Data were shown on a monthly basis although they can be shown on a per inquiry basis. The operating statement can be useful for comparing a system against an absolute standard or against another system. The conclusion is drawn that it is a fundamental requirement that cost, time, and volume criteria be included in measures of effectiveness of information retrieval systems; the model proposed here seems practical for a rough evaluation at the present time.

In addition to this model, a mathematical performance simulation model was developed; it demonstrated that performance simulation is a potentially important tool for learning about information retrieval systems. The objective of the simulation is to reproduce the input, search, and output characteristics of a system, so that the system's performance can be monitored without having to use the actual system. All objects in a system may be coded through a binary description, which becomes a row in a storage matrix. A column of this matrix identifies all stored descriptions on objects with a common feature; search process differences are reduced to questions of how rows or columns are scanned sequentially or in parallel. The requirements of a set of users may be thought of either as a specific combination of column feature references or a set of specified objects or both. This performance simulation model was programmed for test on an IBM 1401 computer and a number of runs were made to test the feasibility of the program.

It is suggested that these two tools for evaluation have potential use, for example, for resolving questions in design, conducting competitive selection evaluations, familiarizing users with specific systems, and running error and reliability studies. Recommendations for further research include a detailed operating study of specific information retrieval systems, application of the performance simulation model to specific systems, and further development and elaboration of the simulation model.

IV-3. A Model for Study and Evaluation of Coordinate Retrieval Systems, in Centralization and Documentation, Second Edition, by Arthur D. Little, Inc., Report No. C-64469, Final Report to the National Science Foundation, Cambridge, Mass., Arthur D. Little, Inc., June 1964, p. 19-31.

An empirical description in mathematical terms has been made of the performance of coordinate-type systems, using searches made in an industrial engineering library containing 70,000 documents as the major source of data for the study. The first portion of the development predicts the number of documents that would be expected from an n-term search with and without an intermediary between user and system; the second portion examines the performance of precision and recall ratios with and without intermediaries.

The model for predicting the number of documents was derived from the pattern of frequency of use of index terms assigned to documents. In coordinate indexed collections approximately 10 percent of all index terms are responsible for 80 percent of all term usage. This pattern of usage conforms closely to the Zipf distribution for frequency of usage of words in natural language text. The expected number of documents to be retrieved in a single search depends on the following factors: (1) the number of index terms simultaneously searched on; (2) the number of documents in the collection; (3) the average number of terms used to index a document; (4) the distribution of usage of terms for indexing; (5) the pattern of usage of individual terms for search purposes (the assumption is made that the number of times an index term is used in search requests is roughly proportional to the term's frequency of usage for indexing purposes); (6) the correlation in usage among index terms in formulating search requests and in indexing documents (the computed probability that any given pair, triplet, quadruplet, etc., of index terms will be used in indexing a document is compared with the actual probabilities computed from the numbers of documents retrieved during searches for pairs, triplets, etc.; since about three times as many documents (per term used after the first) are produced as would be expected, this number is called the term usage correlation factor for the system); and (7) the effect of the intermediary in limiting or encouraging specific types of searches (he can estimate the number of documents from the frequencies of the terms used, and adjust the request to yield reasonable results; when the search involves more than three terms the intermediary does not significantly alter the number of documents to be retrieved and his influence can be ignored).

An equation providing a simple geometric relationship between expected number of documents and the seven factors indicates that the expected number of documents is reduced by 1/10 - 1/12 each time a new index term is added to those simultaneously sought (after the first two). The results of applying this model are compatible with those observed in practice, although they do not furnish strong proof of its validity. The model does provide two ingredients to the analysis: a method for estimating number of documents retrieved for various levels of depth of search requests, and a process by means of which the intermediary could control depth of search for one-, two-, and three-term requests.

To determine how much the recall ratio is decreased when depth of search is increased, the relative importance of nearly synonymous index terms to a specific request is considered. If the documents relevant to some given concept are inspected and the use of terms which best characterize the concept and also of near synonyms of those terms is examined, then the partially synonymous terms can be ranked according to how many of the documents they indexed and the fraction indexed can be plotted as a function of the rank of the synonym. Although only limited data were available, the analysis suggests that a geometric relationship is appropriate to describe the shape of the curve. In general, the term best representing a concept will contribute at least 1/3 but probably no more than 1/2 of the total indexing applicable to that concept. The next best term will probably contribute 1/3 - 1/2 of the remaining indexing for the concept, etc. So if the intermediary selects the best term for search, only 30-50 percent of the relevant documents will be retrieved; if he selects a term other than the best one, there will be an even greater loss. A relationship between precision and recall ratios can be established and shows that the precision ratio on the average tends to increase by a factor of about 2.5 when a new concept and index term are incorporated into the search, while the recall ratio simultaneously tends to drop by a factor of about 4.

One search strategy available for coordinate-type systems is to multiply the number of searches conducted each time the depth of search is increased, with each new search using a different combination of near-synonyms (Method I). Another technique consists of specifying index terms for k concepts but accepting documents which have k and k-1 terms (Method II). Improvements in recall can be effected by using Method I although greater effort is required in the design of the search, usually done by the intermediary. Method II yields higher recall and slightly lower precision than Method I, at a price of a large increase in the expected number of retrieved documents and increased effort in the searching process, in the machine processing stage rather than in the search formulation stage. Regardless of the exact search techniques employed, it seems unlikely that a drastic improvement over the general preformance reported herein can be expected within the basic framework of coordinate searching. It will be extremely difficult to obtain high precision and high recall ratios simultaneously whenever near-synonyms are employed in an indexing system.

This method for analyzing the performance of a coordinate retrieval system has been applied to one specific automated system now in operation. In three areas the analysis is dependent on assumptions: pattern of index term usage for search requests, relative importance of nearly synonymous index terms to different concepts, and effects of the intermediary. Exhaustive search is apt to be a most difficult operation even with a relatively small concentrated collection like the one studied here. If such a collection were to be drastically enlarged, in depth of coverage or breadth and scope, exhaustive searches would be even more difficult, and the cost and effort required to secure a given level of performance would be increased.

IV-4. Models of Performance of Coordinate Search Systems, Appendix II, in Centralization and Documentation, Appendices to a Final Report to the National Science Foundation, by Arthur D. Little, Inc., Cambridge, Mass., Arthur D. Little, Inc., July 1963, p. 29-69.

Details are given of the development of the mathematical description for performance of coordinate retrieval systems, which was presented in the main body of this report (cf. IV-3). The study is restricted to the number of documents retrieved without regard to relevance, and gives attention first to the use of index terms and their rank according to frequency of use. The expected number of documents retrieved in a singlet (single term) search is calculated, as are the numbers for doublet and triplet searches. These expected numbers are compared with the numbers retrieved by actual searches in the industrial library system studied. By assuming that the probability of any of these types of searches is proportional to the number of documents likely to be retrieved provided this number is less than 50 documents, reasonable agreement is obtained between computer and actual number of documents retrieved. The expected fraction of searches that will be singlet searches is also computed and compared with the observed value. Under the assumptions made 6.2 percent of all searches should be singlet; it was found that 10 percent actually were.

Attention is also given to the influence of relative importance of terms on performance of a coordinate retrieval system. If a request is considered as consisting of the intersection of some concepts, translated into several intersections of index terms, then searches are made on all possible n-tuplets formed by taking one term from each concept. The documents retrieved can be assigned relevance measures by the user. The hypothesis is made, and supported by the data available, that terms can be ranked in order of their importance to a concept. To determine how much more important is one term than another, the number of relevant documents indexed under a given n-tuplet is related quantitatively to the number of relevant documents retrieved by the most important n-tuplet (i.e., the one which yields the largest number of relevant documents). Comparison is made of computed values and those obtained from actual searches in the test collection; agreement is good.

The quantitative description of the relative importance of terms is related to the recall that can be achieved by a retrieval system, and to the precision ratio. System parameters of interest can be measured, except for the ratio of total number of relevant documents to total number of documents in the collection. But in general it can be said that for any given request this ratio is large for a collection specializing in the field of the request, of intermediate value for a general collection, and small for a collection specializing in a field outside that of the request.

IV-5. Evaluation of Performance of Large Information Retrieval Systems, Appendix III, in Centralization and Documentation, Appendices to a Final Report to the National Science Foundation, by Arthur D. Little, Inc., Cambridge, Mass., Arthur D. Little, Inc., July 1963, p. 70-109.

Techniques for measuring the actual performance of retrieval systems in their operation, and for comparing the performance of different retrieval systems, are suggested in this appendix to the main body of the report (cf. IV-3). Retrieval systems are understood to handle documents either indexed by a coordinate system or stored as entire text; to obtain listings of documents relevant to the purposes of an inquirer, exhaustive either to a degree of relevance, to a certain number of documents of highest relevance, or to all documents found with a certain amount of effort; to classify documents either as relevant or non-relevant in mutually exclusive relevance classifications, or according to a continuous scale of relevance; to serve inquirers who request documents which contain information

relevant to a given problem which they state, as well as inquirers who request a specific document whose contents they describe; and to be mechanized to the extent that their performance is reproducible.

Two major areas of characteristics of such systems are quality characteristics and "cost" to the user; these areas may be interdependent. Quality characteristics include: (1) completeness, measured absolutely or relatively; (2) precision, also measured either absolutely or relatively; (3) uniformity of these characteristics; (4) learning ability of the system, using the results of past inquiries to improve; (5) teaching ability of the machine, for man-machine or man-man exchange to sharpen the inquiry; (6) compatibility with other systems; and (7) variability of these characteristics, i.e., the distribution of the measure of performance about the mean. Cost characteristics include: (1) monetary cost by user, per search or as subscription to a service; (2) cost in terms of time lag, either absolute or relative; (3) accessibility, in terms of geography, time, or procedural effort; and (4) convenience of form of search product.

The objectives of the pragmatic evaluations developed here, which are considered to be tools for measuring and comparing systems in actual operation on an empirical basis, include testing performance or cost of a system relative to performance of another system; testing performance of another system; testing performance of a system relative to needs of its customers or determining degree to which a system satisfies its customers; and testing general design of a system or of modified versions of a system. Most measures of performance to be proposed can be based on objective quantities, but the notion of relevance of a document to an inquiry is subjective, and attempts to make the concept objective have not been very successful. What is needed perhaps is not so much an absolutely objective measure as one determined by the individual inquirers for the documents in the search product or in the collection. The measures for the various system characteristics listed above should have the following properties: expressible in quantitative terms (even if the measure cannot be expressed in quantitative terms, its effects may be); universality, applicable to all types of systems; reproducibility and objectivity, being measures either independent of the judgment of the measurer or dependent to such a small extent that different measurers will arrive at the same measure; and uniformity, with respect to all parts of the collection and to all inquiries.

The amount of effort required for obtaining these measures constitutes the greatest limitations in their construction and use. Thus, obtaining an absolute recall ratio requires examination of the entire collection or of a sizable portion of it by one or more individuals; it is impractical. Again, comparison of different coordinate indexing systems requires indexing the same collection by the different systems, a costly procedure. These difficulties may be partially overcome by replacing the measures by statistical estimates of them, based on samples from the total population of documents, users, and inquiries.

Proposed measures for the system characteristics listed above are: (1) completeness --- measured by recall ratios. Documents can be divided by individual or collective users of the system into three major classes of relevance --- non-relevant, relevant, and crucial. Absolute measures of completeness for an entire system can be determined by use of sampling techniques. Relative recall ratios for comparison of different systems operating on the same documents can be determined, as can relative recall ratios measuring the completeness of a search in relation to the emphasis placed on completeness by the users; (2) precision --- measured by precision ratios, in the case of absolute measures, for individual inquiries against all the documents in the collection or against the total search product. Relative precision may be determined for time spent to discard irrelevant items or for cost in effort to word inquiries more sharply; (3) uniformity --- of performance and in responses to queries, tested by comparing recall ratios and precision ratios for a subset of documents with those for the entire system. Lack of such uniformity may be caused by

differences in degrees of completeness or competency at indexing or by lack of up-dating in systems where indexing depends on time. If the lack of uniformity depends on formulation of the inquiry to the system, it may be measured by comparing recall and precision ratios produced by different formulations of the same questions; (4) learning ability --- to what extent the feedback mechanism actually improves the performance of the system, and how fast the improvement proceeds in time or in number of inquiries received, measured by recording questions and search products from the system and re-asking the questions later in order to compare the results with earlier results; and (5) teaching ability --- measured by comparing the search product of the system before any man-machine or man-man exchange with the search product in answer to the same question after such an exchange. Measures of cost, mostly straightforward, include the time cost to the user, measured by the average lead time. Lead time (lt) is the time elapsed between the original inquiry and obtaining the desired information. Average lt is the value of lt over the population of all inquirers and users, and may be determined from a representative sample of actual inquiries.

Detailed procedures for obtaining the measures defined above are proposed: (1) the user-inquiry samples (samples over the population of inquirer-inquiry pairs rather than only of inquiries because different types of users may place different emphasis on the same inquiries) may be obtained by selecting all inquiries requesting an exhaustive search received by the system during a fixed period; (2) samples from the collection of documents can be made by selecting a random sample from the collection of a size (5-10 documents) such that scrutiny to determine relevancy does not impose too great a burden on the user. The number of such samples had to be quite large, or the collection has to be divided into broad subject categories and the random sample drawn from the category appropriate to the inquiry. It is shown that if the number of users and number of documents is large enough so that the total number of relevant documents found by users exceeds 271, then overall ratio of the system can be estimated with 90 percent confidence within $\pm$ 5 percent; and (3) uniformity with respect to certain subsets of either the collection or the types of inquiry, testing whether the recall ratios of the subset differ from the recall ratio of the entire system, can be measured by forming a hypothesis about the relation between the numbers of actually retrieved documents and the expected number, and testing it for significance. A small probability indicates a lack of uniformity.

IV-6. Quantitative Relations Useful for the Evaluation of Information Retrieval Systems, in Appendix to Centralization and Documentation, Second Edition, by Arthur D. Little, Inc., Report No. C-64469, Final Report to the National Science Foundation, Cambridge, Mass., Arthur D. Little, Inc., June 1964, 18 p.

Algebraic expressions which describe the performance characteristics of three specific schemes of coordinate document information retrieval are formulated in terms of probabilities which can be measured through observation of a given retrieval system. The collection is assumed to consist of a large number of documents each containing a large number of concepts. We assume that the degree of correlation of concepts in documents and requests is the same; and that if k concepts are present, their joint probability is γ^{k-1} times the product of their individual probabilities; γ is approximately 3. Each concept in a document is assumed to be represented by an index term, and we shall assume

that there are no errors of commission. The user is assumed to want to retrieve all documents that have a specific set of concepts; a document is assumed to be pertinent only if it contains all the concepts he specifies. The fundamental retrieval measures are the recall and precision ratios; another convenient measure that summarizes the behavior of a system in one number is the correlation coefficient, ρ, between retrieval and pertinence. This measure has several interesting properties that may be readily verified. If a document is retrieved only if it is pertinent, $\rho = +1$. If no retrieved document is pertinent, $\rho = -1$. If retrieved + documents are selected by extracting documents from the collection at random, $\rho = 0$. Only systems with positive ρ's are interesting.

We shall consider three retrieval schemes: in Scheme I_1 a document is retrieved only if its index set contains acceptable synonyms for all concepts in the search sub-set; under Scheme I_2 a document is retrieved only if its index set contains acceptable synonyms for at least all but one of the concepts in the search sub-set; in Scheme I_3 a document is retrieved only if its index set contains acceptable synonyms for at least all but two of the concepts in the search sub-set.

We begin by computing the probability that a document chosen at random is retrieved by Scheme I_1, then compute that a document picked at random is pertinent, derive the joint probability of retrieval and pertinence, and thus derive the recall and precision measures for I_1. We compute the probability of retrieval for I_2; the probability of pertinence is the same as for I_1, because it is a property only of the collection and the definition of pertinence. Then we compute the joint probabilities of retrieval and pertinence and the recall and precision ratios. We follow the same computations for I_3.

An interesting measure of a retrieval scheme is the number of searches that must be made using the scheme. We compute this for each scheme under the assumption that there are r acceptable synonyms for each concept. For Scheme I_1, we must conduct a search for each synonym for every concept; for Scheme I_2, we must search each of the subsets generated by leaving out one concept; and Scheme I_3 requires searching each of the subsets generated by leaving out two concepts. In comparing the various measures, including correlation coefficient, for I_1 and I_2, the system I_1 is found to be preferable. The general results on measures specialize when concepts are independent; the recall ratio is higher for I_3 than for I_2 and higher for I_2 that I_1. On the other hand, the precision ratio will be lowest for I_3 and highest for I_1.

The notation and arguments of this Appendix are brought into correspondence with those in the First Edition of Centralization and Documentation, July 1963 (cf. IV-3).

IV-7. An Experiment with File Ordering for Information Retrieval, by George Badger, Jr., and William Goffman, in Parameters of Information Science (Proceedings of the American Documentation Institute, 1964 Annual Meeting, Vol. 1, October 5-8, 1964, Philadelphia, Pa.), Washington, D.C., Spartan Books and London, Cleaver-Hume Press, 1964, p. 379-381.

These experiments were conducted at the Center for Documentation and Communications Research (CDCR), Western Reserve University, on a file of approximately 100,000 items accumulated for the American Society for Metals Documentation Service. Questions were provided by paying customers and searches were carried out by a computer. In the first experiment the file was partitioned in three disjoint classes based on the experience of 200 questions posed to the system in the past. Each document was assigned a rank which was equal to the number of times it had been sent to a customer as a response to a query. The three files were interrogated by a set of 26 questions; the test statistic was

the number of responses by the system divided by the number of documents searched. The next step was to repartition the file into six subfiles on the basis of the experience obtained from searching the 26 questions and repeating the process with a set of new requests. A two factor analysis of variance was applied to the data in each case; both variations due to question and due to file ordering were significant. Results indicated that the system was able to retrieve in the first case 33 percent of the documents by searching 19 percent of the file. In the second test this was improved to 44 percent of the documents in 25 percent of the file.

This method of file ordering by ranking was tested to see whether an increase in rank indicated an increase in the probability of relevance as assessed by the evaluator. An analysis of variance on the ratio of relevant to non-relevant evaluations was applied to the data. In both cases the ranking factor was not significant. This result lead to the consideration of the decision processes of the evaluators. Two preliminary studies were conducted. The first of these attempted to assess agreement among a set of evaluators. Each of three evaluators was asked to dissect the computer output to a question into relevant and non-relevant subsets. A Chi-square test was applied to the observed evaluations as compared to those expected if the three evaluators were in complete agreement. The Chi-square of 81.57 was very significant, indicating an absence of agreement.

In the second study the outputs to each of three questions which were dissected into relevant and non-relevant parts were selected, and the non-relevant part was submitted to evaluators to dissect each set into relevant and non-relevant parts. The results showed that a higher percentage of relevant documents was obtained from the non-relevant part of the first dissection than had been obtained in the original dissection. These studies raise fundamental questions concerning factors other than the relationship between the document and the question which influence relevance. Experiments designed to isolate some of these factors are being conducted at the CDCR.

IV-8. Theoretical Considerations in Information Retrieval Systems, by Jack Belzer and William Goffman, Communications of the ACM 7, 439-442 (July 1964).

When a user of a system seeking information poses a question to the system, he must be assured that the system will interpret his question the way he would if he were scanning the file. The documents he would select with the question in mind are designated as relevant; those which he would select to satisfy his need are referred to as pertinent. The difference between his interpretation of the question and the system's interpretation accounts for false drops in the system's output. However, the difference between his need and his ability to express that need will produce false drops in the relevant set. The concept of relevance is an elusive one. It would be useful to develop an evaluation function for the performance of the system, constructed so as to reward the system for retrieving relevant documents and not retrieving the nonrelevant, and to penalize the system for false drops and escaped relevant documents. An evaluation function is going to depend on the purpose of the system: if it were to provide scientists with reading material, the function would be linear and would depend on the relevant part only. (The linearity of the function may hold for a limited range; however, an output of 10,000 documents is not necessarily twice as good as an output of 5,000).

The evaluation function D(A) should read: α [relevant part] - β [irrelevant part] - γ [escaped part] and δ [rest of file]. α, β, γ and δ are nonnegative constants, $\alpha + \gamma$ measure the value of successful retrieval and $\beta + \delta$ measure the nuisance of spurious retrieval. This model has shortcomings, including that not all documents have an equal probability of being relevant to a particular question, and it is not realistic to consider relevant documents without considering their ratio to escaped documents. These shortcomings will be considered in subsequent models. Based on this simple model, however, it is profitable to include a document in the answer if $\alpha f(x) - \beta [1-f(x)] > -\gamma f(x) + \delta [1-f(x)]$, or if $f(x) > (\delta + \beta)/(\alpha + \beta + \gamma + \delta)$, where f (x) is the probability function. Evaluation of the coefficients, and the weights of each, depends on the purpose of the system. Having determined the coefficients, the last inequality presents a clear cut evaluation function for establishing relevance of a document relative to a question.

Identification of information is a function of effectiveness of a system, and file organization is a function of efficiency. Documents will have probabilities of relevance, relative to questions. Operationally, this may be difficult to attain. However, a probability for a tag relative to the total document may be obtained. If for each document in the file a set of tags or terms characterizing the document $\underline{d}$ be obtained, each tag or term would not have the same probability of being relevant to that particular document. Since questions posed to the system use the same terms as those in the file, the term in the question $\underline{q}$ would similarly possess probabilities relevant to $\underline{q}$. The sum of the products of the coefficients of $\underline{q}$ matched with those of $\underline{d}$ would determine a relevance probability of the document relative to the question. A critical probability D(A) would determine whether the document would be included into the set of answers or rejected. This approach has a significant effect on the file structure, because the system does not want to scan the entire file for each question posed to it. It would rather generate a subset of the total file, which has a chance of being relevant, and compute the relevancy probabilities for this subset only.

Alternatively, it is possible to obtain a frequency distribution of the documents in the file appearing as answers to questions and assume that they are the most likely to continue appearing as answers (if the system in general caters to a group of users with similar interests). Such a file would be dynamically self-organizing and a user would have the opportunity of probing a subset of the file where the most likely results would be.

IV-9. A Research Plan for Evaluating the Effectiveness of Various Indexing Systems, by Harold Borko, Report No. FN5649/000/01, Santa Monica, Calif., System Development Corp., July 1961, 23 p.

The plan proposed here is divided into three phases: comparison of effectiveness of three indexing systems, study of common characteristics of the systems, and a factor analysis of significant descriptor terms from the sample documents. For Phase I, a sample of 612 abstracts was obtained from the Vol. 32, No. 1, 1958 issue of Psychological Abstracts; these abstracts would then be indexed by the American Psychological Association (APA) subject heading index, by permutation index of their titles, and by application of descriptor terms automatically derived by computer analysis of titles and abstracts. The APA subject headings were selected from the annual index to Psychological Abstracts in which the sample articles were originally indexed. The automatic indexing technique would make a frequency count of words, derive a score for each word, and then determine which words to include in the index. If fully automatic selection should prove impossible, human judgments would be made in selecting index terms.

The major difficulty in evaluating the effectiveness of the different system deals with defining an acceptable criterion of effectiveness. Several suggestions have been made: Mooers (cf. IV-27) lists user satisfaction, comparison of one system with another, and comparison with an ideal system, as possible methods; Bornstein (cf. IV-10) suggests comparison as to the amount of relevant, partially relevant, peripherally relevant, and nonrelevant information given in response to a request, and the amount of time spent in examining these responses; Cleverdon reports that measures of efficiency should be based on measurements of economic costs. For this study, costs and time are not the main concern, so the criterion selected for use is the ideal system suggested by Mooers, with slight modifications. A procedure approaching this method was also used by Swanson (cf. II-53).

In this project each index system is evaluated independently against an absolute criterion as measured by an equal internal scale of retrieval effectiveness. For this it is necessary to evaluate every document in the sample for relevance to every question --- each of 612 documents rated as to whether it contains information relevant to each of 45 questions. Degree of relevance will be judged on a four-point scale; very relevant, relevant, slightly relevant, and not relevant. The judgments obtained by this method will be made on a rating scale of successive intervals or rank order. The raters' judgments of the relevancy of documents will be checked for reliability.

The formula proposed for measuring the retrieval effectiveness of any index system in response to every question is $R = r - pi$, in which R = retrieval score; r = relevancy score or $\frac{S}{T}$, where S = sum of the relevancy weights of retrieved documents and T = total sum of relevancy weights (for a given question) of all documents; i = irrelevancy score or $\frac{M}{N}$, where M = number of irrelevant documents retrieved and N = total number of documents retrieved; and p = penalty factor by which the system is charged for retrieving irrelevant documents. That is, the retrieval score for a given question is determined by the relevancy weight of documents retrieved, relevancy weight of documents in the library but not retrieved, and number of irrelevant documents retrieved.

The number of questions needed (45 plus three practice questions) is determined by the fact that analysis of variance design will be used to interpret the data; this number of questions will insure that the numbers in each cell will be large enough to provide stability. Representative questions will be provided by asking members of APA to submit questions for which they tried to find answers in Psychological Abstracts indexes, and having those questions screened to select only those which could be answered by the experimental library. Ratings of relevancy will be made by graduate students in psychology, and the ratings converted to an equal interval scale using Likert's Scaling Procedure. Then 30 different students will use each of the indexes to locate all documents relevant to the test questions, keeping a record of terms used and number of documents retrieved for each question.

Retrieval scores will be computed on the basis of number and relevancy of documents retrieved. The analysis of variance design will enable determining whether differences are statistically significant and due to the system, variations among questions, etc. The hypotheses to be tested include that there will be a significant difference among retrieval scores and that the highest score will be obtained by the Permutation Index; and that there will be no significant differences among scores obtained by the three groups of subjects, in response to the three groups of questions, or on the basis of increased experience in searching the index.

For Phase II, the indexing terms used to describe an article in all three systems will be determined, as well as the average number of terms used. The degree of similarity or overlap between the index terms of the systems will be expressed as a percentage. Hypotheses to be tested include that the subject heading index will contain the least number of headings to a given article, and that the terms selected automatically will include a significant proportion of the terms selected by the human indexers.

Phase III proposes a procedure for determining subject categories which is based on the properties of content words in the documents. Such a procedure will help to provide a mathematical basis for subject heading indexing and will likely improve the efficiency of automatic indexing procedures. Methods of procedure include factor analysis to organize content terms into a smaller number of categories, and statistical operations to obtain a symmetrical matrix of interrelationships between descriptor terms.

It is felt that the results and research procedures used in the study should be generally transferable to similar studies in related fields.

IV-10. A Paradigm for a Retrieval Effectiveness Experiment, by Harry Bornstein, American Documentation 12, 254-259 (October 1961).

The criteria established to evaluate indexing systems are the most important aspect of the problem of comparing several indexing systems. The paradigm presented here is based on pragmatic criteria of evaluation --- the user of the information and his judgment of the relevance of the information retrieved. The relative efficiency of different information handling systems can be compared as to: (1) the amount of relevant, partially relevant, peripherally relevant, or non-relevant information obtained in answer to a request; and (2) the amount of time spent in examining the responses. For the suggested experiment a collection of 20, 000 documents is indexed by each of three different indexing systems. A cluster of three analysts (representing users of such systems) from each of 10 classes of analysts is asked to compose search questions designed to solve typical problems in their work; each cluster produces 30 questions, for a total of 300. These questions, phrased to secure the needed information independently of any of the indexing systems, are sorted into three groups of 100 each. Each group is translated into the search criteria for one of the systems; the 300 questions are addressed to half the document collection. The 30 questions with the largest number of responses in each system are retained, and these 90 then translated into search criteria for all three systems. The analysts' reactions to the information supplied are recorded: the time he takes to examine each question, and the hard copy retrieved for it, from each system; his judgment of degree of relevance of each item presented; and whether or not the unit record is a satisfactory substitute for the actual hard copy. For each information system the variables of (1) amount of information and (2) time spent in examining responses, as mentioned above, are determined; in addition, (3) the proportion of acceptable substitutes for hard copy for the responses and (4) the actual coincidence and uniqueness of responses on the scale of relevance are obtained. Statistical analyses of the variances in the first three variables, and simple analysis of the data for variable 4, should result in knowledge of the advantages and limitations of the indexing systems.

In the Technical Appendix to the report the following suggestions and comments are made: that indexing in the paradigm be as thorough as possible; that an ancillary series of small experiments be conducted to yield estimates of several aspects of indexing processing; that a known store of information not be generated (it would affect the comparisons basic to the experiment) and that measures of absolute efficiency not be used (they would require excessive effort to compensate for loss of statistical power); that a much larger

library than the one suggested here must be generated, to determine relationship between retrieval effectiveness and size of store; that the two indices of retrieval effectiveness (relevant and irrelevant documents received, and time spent in examining these documents) are important and both should be included in the experiment; that categories of questions that analysts use in their work need to be developed; and that retrieval effectiveness cannot be inferred from measures of indexer or requestor inconsistency.

IV-11. Requirements, Criteria, and Measures of Performance of Information Storage and Retrieval Systems, by Charles P. Bourne et al, Report to the National Science Foundation, SRI Project No. 3741, Menlo Park, Calif., Stanford Research Institute, December 1961, 132 p.

(The National Academy of Sciences-National Research Council ad hoc Committee, charged with evaluating the American Society for Metals-Western Reserve University Metallurgical Searching Service, recommended that operations research studies be made on criteria for evaluating information retrieval systems. Two limited, exploratory studies directed toward the development of measures of effectiveness were prepared, by Arthur Andersen and Co. (cf. IV-2), and Stanford Research Institute, reported here.)

User requirements need to be measured and described in quantative terms, if possible, before any evaluation procedures are implemented. This study attempted to devise a methodology by which some of the requirements can be so measured; attention was directed to needs for specific information to help with current project work and for exhaustive searches usually performed as a prelude to project work. The evaluation procedures were developed to assess the degree to which storage and retrieval systems satisfied these types of requirements.

After interviewing 92 applied electronics researchers and 11 metallurgists, it was found that a composite agreement was provided on the relative importance of seven different factors; the most important was agreed to be response time, time between request and receipt of the major group of relevant references. There was no strong agreement on the relative importance of the other six factors: relevant material overlooked by the search, irrelevant material produced by the search, form in which references are given, assurance that documents on a given subject do not exist, effort necessary to communicate request, and certainty that specified sources over a certain time period were searched. It was difficult to measure some of the requirements quantitatively; however, some useful results were obtained.

Three separate and complementary tools were developed for the analysis and evaluation of information retrieval systems. The first, a coarse screening procedure, arranged empirical data to show the ranges of parameter values that are likely to be encountered by candidate systems. The second, a performance evaluation procedure, relates system performance to user requirements to arrive at a single figure of merit for each system applied to each user population. This tool can be implemented either by direct measurement and correlation of performance and requirements, with weighting for the relative importance of each requirement; or by reduction of all requirements and performance to a common denominator of time or cost. The third tool, two cost analysis procedures, utilized a specially developed general functional model of a storage and retrieval system and computer simulation programs to determine operating costs over wide ranges in operating conditions such as file size, accession rate, and volume of search requests. Both cost analysis procedures were successfully applied to three representative systems. The computer programs were written in ALGOL so they can be used by any interested group.

The study concluded with suggestions for further research in the following problem areas: (a) development of methodology for determining user requirements; (b) determination of elemental times and costs of the basic operations in retrieval systems; (c) development and use of modelling for analysis of operating costs; (d) development and use of modelling for performance evaluation; (e) pilot tests or evaluations of representative systems; and (f) additional basic studies on such problems as how the user's productivity is related to the type and amount of information services provided, and how search needs are related to the tasks required of the individual.

IV-12. Cost Analysis and Simulation Procedures for the Evaluation of Large Information Systems, by Charles P. Bourne and Donald F. Ford, American Documentation 15, 142-149 (April 1964).

Techniques have been developed for economic analysis or evaluation of large information systems; they should be of value in comparing several different alternative systems, or in determining sensitivity of a given system to changes in its different parts. A computer program, which was written in standard ALGOL language and run on a Burroughs 220 computer, accepts detailed descriptions of operating performance, costs, and interrelationship of all components in a proposed system, then simulates the operation of the system over a 5-year span to compute estimates of operating costs, and equipment and personnel required, during that period. The description of performance includes the equipment used and all the operations performed, as well as operating speeds and basic times to perform manual operations; costs for equipment, supplies, and wages; and statements about file size, input rate, and number of search requests. Using these data the program determines for a one-month period total labor costs, required numbers of each type of equipment, and material costs, obtaining a dollar total for the operation of the entire system for one month. To be realistic, the 5-year costs are computed on a monthly basis using a continuously increasing file size and attendant increasing costs. Where one alternative system is less expensive part of the time only, two methods of comparison useful in making a choice are the "present worth" and "equivalent annual cost" methods; the program can compute these amounts for the candidate systems.

IV-13. Design and Evaluation of a Literature Retrieval Scheme, by Heinrich A. Ernst, in Quarterly Progress Report No. 55, Cambridge, Mass., Massachusetts Institute of Technology, October 15, 1959, p. 130-131.

(This report is an abstract of the author's thesis, Design and Evaluation of a Literature Retrieval Scheme, submitted to the Department of Electrical Engineering, MIT, in partial fulfillment of the requirements for the degree of Master of Science, June 1959.)

In a hypothetical information retrieval system, a machine presents documents, then choice of pertinent ones directs the machine to point out other pertinent information. The system makes use of the reference lists contained in documents which define a relation network among the documents. The machine is guided along the branches of this network to make its selection of pertinent information, determining the next set of possibly pertinent documents by topological procedures in the relation network. Two manual imitations of the system were made to evaluate the design. Using Stumper's Bibliography on Information Theory, Part I, retrieval loss and efficiency were determined: less than 5 percent retrieval loss with an efficiency of 5-15 percent, less than 30 percent retrieval loss with an efficiency of 20-30 percent. No procedures for improving these figures were found when the relation network was used.

IV-14. Basic Parameters of Retrieval Tests, by Robert A. Fairthorne, in Parameters of Information Science (Proceedings of the American Documentation Institute, 1964 Annual Meeting, Vol. 1, October 5-8, 1964, Philadelphia, Pa.), Washington, D.C., Spartan Books and London, Cleaver-Hume Press, 1964, p. 343-345.

Numerical comparisons of system performance depend upon parameters of a two-by-two table in which the row sums are the numbers of acceptable (C) and unacceptable items; the column sums are the numbers of items selected (L) and rejected by the system. The four cells thus contain the numbers of items selected and acceptable (R), acceptable and rejected (C-R), unacceptable and selected (L-R), and unacceptable and rejected (N + R-C-L, where N = total number of distinct items in the test collection). This paper attempts to derive from documentary, rather than mathematical, considerations those parameters that display fundamental retrieval functions.

In conventional retrieval tests, R and L are observed test results, C is asserted, and N is known. The parameters derived from these are usually R/L (precision ratio) and R/C (recall ratio). In addition we need C/N (concentration) and N (size of collection). The precision and recall ratios describe adequately only the properties of systems designed to retrieve all but not only (ABNO) of a few items, not those designed to establish non-existence of items, to supply only but not all (OBNA) items, or to deal with common items in either the ABNO or OBNA mode. We need four parameters (or three, if ratios are adequate) to cover all retrieval situations. The basic retrieval functions are distillation or discrimination, i.e., under suitable invocation to separate the collection into two parts; and identification, i.e., to specify which part has the higher concentration of acceptable items. Discrimination must be numerically independent of how we label the portions. Its value must be unchanged, except for sign, if we interchange the columns of the two-by-two table. Also parameters must be unchanged, except for sign, if we interchange the rows. Whether we specify acceptable items or unacceptable items, our separation of the collection is the same. Thus our parameters are N, C(N-C), L(N-L), RN-LC. The first is the number of distinct items in the collection. The second and third measure the balance of acceptable and unacceptable, selected and rejected, items respectively. The last expression indicates how well the collection has been distilled. Its maximum value of C(N-C) occurs when all the acceptable items are in one part and all the unacceptable in the other.

If the absolute size of the collection is not a primary influence three ratios suffice to describe the test: $C(N-C)/\frac{1}{4}N^2$, $L(N-L)/\frac{1}{4}N^2$, $(RN-CL)/L(N-L)$ or $RN-CL/C(N-C)$. When C and L are small relative to N, the expressions become numerically indistinguishable from C/N, L/N, R/L or R/C. In general, only when all three of these ratios are comparable are the tests or the systems comparable.

The first two expressions above represent the relative balance of acceptable and unacceptable items, and of selected and rejected items, respectively. The first of the two dependent alternatives is called distillation, because it is the difference between the concentration of acceptable items in the selected portion and acceptable items in the rejected portion. The second of the two is called discrimination because it is the difference between the proportion of acceptable items selected and unacceptable items selected. One cannot distill the collection without some discrimination nor discriminate without some potential distillation. Thus the dependence of these parameters has documentary as well as mathematical interpretation. The "trading relation" between recall and precision, whatever it may be, is a pragmatic phenomenon apart from this dependence.

The parameters derived here, if used correctly, describe the test completely. They hold good for all test conditions, including tests of non-existence, and for both ABNO and OBNA systems. They are unchanged, except for sign, by changes of label. They have fundamental retrieval interpretations.

IV-15. Data File Size and Its Relation to the Bayesian Effectiveness of an Information Retrieval System, by Ugo O. Gagliardi, Dunlap and Associates, Inc., Report No. ESD-TR-65-275, Report to U.S. Air Force, Electronic Systems Division, April 1965, 80 p.

A model for evaluating the effectiveness of information retrieval systems which is based on Bayesian statistical theory relates the effectiveness of a retrieval system to the statistical characteristics of the retrieval process and the value structure of the requestor. The model identifies the operational conditions under which one cannot expect to obtain sufficient system performance to justify the development of a system. The approach taken is to characterize the overall error process by two parameters, the recall and precision ratios. These give, respectively, an indication of the average degree of completeness and the average degree of purity of the search. These two parameters are used to express a model of the error process of the retrieval system.

The scalar measure of effectiveness for information systems which has been derived is given by $\epsilon(Q) = [UD_\pi Q]^* \xi - [U\pi]^*$, where $Q = \left\{ P\left(Y^i \mid X^j\right)\right\}$, the Information System Model; $U = \left\{ U\left(A^k \mid X^i\right)\right\}$, the utility structure of the decision maker; $\pi = \left\{ P\left(X^i\right)\right\}$, the prior distribution; D_π = diagonal form of π; ξ = vector of all ones; and $[\ ]^*$ = operation which takes the largest component of the column(s) inside.

A model for the evaluation of a retrieval system using this measure is obtained by introducing recall ratio and precision ratio. The simplest model which is consistent with a Bayesian statistical viewpoint is represented by:

$$\epsilon(Q) = \left| \begin{matrix} r(q\gamma - \beta)(1-q) & q\gamma(1-r) - (1-q)(1-r\beta) \\ -\alpha qr & -\alpha q(1-r) \end{matrix} \right|^* \xi + - \left| \begin{matrix} \gamma q - (1-q) \\ -\alpha q \end{matrix} \right|^*$$

where r = recall ratio, p = precision ratio, q = density of relevant material, α = loss ratio of misses vs. false drops, and γ = utility ratio of hits vs. false drops.

This model was programmed for computational experiments which show that the behavior of the model is determined by the two parameters, $\theta = \alpha q$ and $\lambda = \gamma q$. The number α represents the number of irrelevant documents which a user is willing to examine to avoid missing a relevant document; the number γ represents the number of irrelevant documents which a user is willing to examine to find a relevant document.

Thus $\delta = \left| \begin{matrix} \alpha \\ \gamma \end{matrix} \right|^*$ represents the greatest number of irrelevant documents the user is willing to examine for one positive outcome. Upper and lower bounds for the value of q can be established. Systems which fall above the upper bound correspond to files which are interrogated just as effectively by examination of all the items in the file; systems which fall below the lower bound correspond to systems which are never utilized since elimination of non-relevant material retrieved requires an effort greater than the value of the relevant material retrieved.

It is not possible, using present state-of-the-art indexing, to have effective systems with file sizes much in excess of 100,000 documents. There is critical need for research directed to developing superior indexing and retrieval techniques in order to broaden the size "range".

IV-16. Information Retrieval System Evaluation Technique (Termination Report), by J. L. Garland and Kenneth W. Webb, Task No. 0355, Rockville, Md., International Business Machines Corp., Federal Systems Division, March 1, 1962, 23 p.

The program reported here was undertaken to develop a series of mathematical techniques for evaluating cost and effectiveness of information retrieval systems. The tasks included a review of the literature in the field (25 references are given), development of a detailed characterization of the problem and a method for gathering data, selection of an evaluation strategy applicable to a wide range of systems, and establishment of a set of criteria for evaluating system performance. Library-type systems were examined, particularly those in government agencies. Measures for evaluation of user satisfaction were deferred for later investigation.

The methodology for gathering data places the elements of a system in a matrix that identifies the operations of men, machines, units of information, materials and facilities along the vertical axis and values (identified by code) for these elements in time, dollar costs, and quantity along the horizontal axis. The technique of ratio analysis is used to reduce the data in volume, correlate them, and present them in numerical form as quantitative values which can be evaluated.

The evaluation strategy makes use of the payoff matrix, composed of rows called "strategies" and columns called "states of nature". A strategy is a particular configuration of hardware, software and people to produce a system. A state of nature is a particular set of demands on the system in terms of quality and quantity. The payoff values are calculations which indicate the relative utility of the system under a particular estimate of demands. In calculating payoff measures, the state of nature equal to having no system at all (i.e., no logic to the flow or to organization of the information) must be considered, as well as the possibility of a negative utility in which some system will be worse than no system. Costs of input processing, file maintenance, output processing, and various performance ratios are reflected in the calculations, as are searching time costs and costs of waiting in queue. When the payoff matrices have been calculated it is possible to select the best strategies and therefore the best systems.

The criteria against which library performance can be measured were not completely developed. The parameters estimated in the study are primarily derived from quantitative analysis. Quantitative data are less expensive to collect and were selected on the assumption that much valuable information can be inferred about performance of a system by detailed analysis of its input and output. The technique can provide an indication of expected cost/payoff for each of a series of selected strategies. Overall evaluation, however, will require qualitative judgment by the decision-maker.

IV. 17. The User Approach to Information Systems, by F. Loyal Greer, Bethesda, Md., General Electric Company, Military Communications Department, May 1963, 41 p.

In the design of information storage and retrieval systems, too little attention has been paid to the problems associated with natural word usages in request behavior. Analysis of the words of users in information systems in stating their requirements should be profitable. This paper deals with problems of the user's communication with the inanimate library or mechanized storage and retrieval system. Problems of request formulation are critical both for manual and mechanized systems and for classificatory and search word procedures. This paper focuses on the problems and solution for identifying useful search words for indexing pruposes, but should have implications for information classification also.

The development of the User Approach to storage and retrieval was preceded by empirical investigation of the consistency of human indexers (cf. II-13 and II-14). Even though human indexing consistency were to reach perfection or a highly reliable machine indexing approach is employed, one could not guarantee similarity between the words of the indexing and of the users. Additional investigations focused on word usage responses (cf. IV-18). The User Approach reflects results reported in this latter study and also combines the word response technique with simulation procedures, to evaluate design prior to system implementation. Such evaluation is part of the model for the User Approach, where simulation is used without involving actual documents. In addition, the program suggested here follows a pattern of solution resembling that used in a communications project, in which hypothetical problem situations were delineated and ideas for the achievement of these situations were provided by knowledgeable consultants. It was crucial here that the "consultants" be representative members of the relevant target audience. Potential system users provide through simulation words they will use to search for answers to specific questions. The number of people giving a word is one determinant of its significance. The 11-step model provides a basis for maximizing the probability that the words used for storage will be the words found in the request formulations of system users.

The steps in the model include definition of informational area, enumeration of information subdivisions, listing of problems in each division, description of clientele, selection of clientele sample, listing questions for the problems, sifting questions, search word listing for questions, and application of criteria for search words. Step 10 tests the products of previous steps by repeating them with some differences, such as using a separate but comparable group of subjects. Step 10 provides a general test of the adequacy of the search words so far identified, measures the extent to which the same questions are repeated and the same words evolved, and also measures the degree to which search words already determined cover the words used for new questions. Step 11 is integration of results for final concordance.

The user's concordance may be compared with other concordances. For example, it is proposed that in subject-indexing (as contrasted to word-indexing) the subject terms will be much more directly related to user requirements than the author's terms. The User Approach includes a method for implementation and consequently the possibility for testing this thesis. Also the User Approach incorporates provision for varied indexing as appropriate to particular information systems. Use of the model stated here, with enumeration of pertinent questions and search words, offers a basis for criteria for indexers in selecting significant words in a text. The user concordance may very likely lead to greater indexer consistency.

IV-18. Word Usage and Implications for Storage and Retrieval, by F. Loyal Greer, Washington, D.C., General Electric Company, Information Systems Operation, July 1962, 74 p.

It is believed that attention needs to be paid to the problems of retrieval associated with word usage in the request behavior of the user of information systems. The emphasis here is on search words which a user employs in seeking information. The "User Approach" is contrasted with the "Expert Approach", using judges to determine the important words for storage, and the "Statistical Approach", representing analyses of an author's written terms for use as storage and retrieval vehicles. The approaches are in danger of failing to employ terms which will match the user's expression of his needs. But the user's request formulations must match the storage subsystem in a way that maximizes the probability of retrieving relevant material. This paper reports on two preliminary studies assessing word usage behavior. The problem under study is the user's communication with an inanimate library or automated storage and retrieval system. The approach

developed can provide leads for exploring a series of potentially profitable requests in retrieving information for any particular question. Various systems embodying elaborate encoding procedures or even total text of documents in storage have been proposed; their effectiveness will be largely contingent on the human's ability to communicate with them.

In the Topic Response Study words given in free recall to topic categories were examined in relationship to interest. The words were analyzed as a function of degree of usage in the literature and rate or number of emissions. Eight graduate students were asked to write as many relevant words as possible in the time limit (seven 30-second intervals) for each of five topics presented in random order --- business, communism, religion, science, and sports. In the second part, the subjects rank ordered the topics according to their interest in them. Each word recalled was assigned a value representing its frequency of usage in the literature according to the Thorndike-Lorge word count study, with the high values representing words infrequently used. The number of words given was determined for each topic; additionally the mean usage value for the words was determined. Results include that there were more words given for areas of high and low interest than moderate interest and that rare words were given at the beginning and again at the end of the time periods. No attempt was made to evaluate the appropriateness of the words given. The study suggests that the more knowledge one has in an area the more words he will reproduce for that area.

The Search Word Study sought to ascertain the feasibility of research on word usage in the retrieval context and to derive some hypotheses about differences in the use of search words. Two technical papers by Rodgers (cf. II-13 and II-14) gave added impetus to the research. Twelve individuals with some background in storage and retrieval served as subjects. The study was conducted concurrently with the Inter-Indexer Study referred to above. The subjects were asked to list in columns all the words they thought might be used as search words in the general area of information storage and retrieval, within a 15-minute time limit. The words given by the subjects were used to provide an index for the usage values for the words. The number of words, of common words, and of rare words were counted; the rare word and common word percentages were calculated, as well as these percentages within the first and last third of a subject's words. Combined Mean Word Percentage Consistency attempted to assess the agreement among subjects in the use of words and was based on comparison of one individual with each of the others. In addition the mean percentage of a subject's words given by others and of other words agreeing with the subject's were calculated.

The results of the Search Word and Inter-Indexer studies were compared for overlap in words designated as significant and for similarity of subject performance. Results of the former include that 633 words were written down, of which 321 were different. Approximately 69 percent of the total were given by two or more individuals. The combined mean word percentage of inter-subject agreement for all 12 subjects was 14.4; the other two inter-subject consistency indexes yielded a mean percentage consistency agreement of 26.1 among subjects. There was a 59 percent overlap between search words in free response and the words judged significant in the text of the "first" article in the Inter-Indexer Study; for the 20th article there was still a 20 percent overlap in words. Overlap between search words and the 10 most frequently mentioned words for the Inter-Indexer Study was 7.7 words on the average.

The studies reported here attest to the complexity of variables which can influence a requestor's choice of words. These variables must be considered if there is to be a maximum articulation between search words employed by a requestor and the manner in which information has been stored.

The paper concludes with a discussion of the User Approach to Information Systems, reported elsewhere by the author (cf. IV-18).

IV-19. A Proposed Pilot Project, by C. Dake Gull, in Annual Report, 1956-1957, Division of Engineering and Industrial Research, National Academy of Sciences-National Research Council, Washington, D.C., National Academy of Sciences-National Research Council, February 18, 1958, p. 54-61.

The American Society for Metals sponsored a classification for metallurgical literature which can be used with punched cards. It also supported a pilot-plant project at Western Reserve University (WRU) for machine methods of searching and retrieving coded metallurgical literature. The pilot project proposed here would expand the revised schedules prepared for the second edition of the classification scheme into three products: a list of terms suitable for coordinate indexing; a systematic classification based on those terms; and another listing acceptable as a conventional authority list of subject headings, also based on the same terms. The methods when developed would be applied to the same 25,000 abstracts which were then being put into machine language at WRU. This would afford an opportunity to obtain some statistics on how long it took to prepare the four different methods and how much it cost. It would then be desirable to put the four methods in four different metallurgical libraries and to interchange among the libraries all questions recorded at all of them. At the end of a testing period of a year or two all of the four libraries would have answered all of the questions. The questions and answers would then be turned over to an independent group of metallurgists and information specialists to study the effectiveness of retrieval under the four methods. By comparing this information with the cost of introducing the information into the system and maintaining and operating the service, it should be possible to have a reasonable indication of the relative effectiveness of these four methods for the storage and retrieval of information.

IV-20. Measurement of File Operating Effectiveness --- Time, Cost, and Information, by Robert M. Hayes, Part 6 of the Final Report on the Organization of Large Files, NSF Contract No. C-280, Sherman Oaks, Calif., Hughes Dynamics, Inc., Advanced Information Systems Division, April 30, 1964, 50 p.

The design of an information system represents the attempt to provide balance among the factors of required response time, degree of required reliability, and cost and value. All other considerations are a result of the compromise between time and reliability on the one hand, and cost and value on the other. This report demonstrates the importance of the compromise. The relationship between decisions and their value is a function of the time at which they are made and of the information entering into them. As a function of the amount of information, the value may be characterized by a normal growth curve; as a function of time, by a normal decay curve. The problem in the former case is that of providing a consistent measure for information. In the latter case the overwhelming role of time becomes evident. If the operation of an information system can be characterized as a function of both time and information, the relationship between the size, F, of data base on which a decision is based, the two parameters, I and T, of information and actual response time, and the cost, C, of the system must be defined.

Consider a file F of bits consisting of items, x, each of N bits. If a request is matched against each item in the file over a specified n bits of the N, the information from the file in response to the set of requests is measured as a function of F, N, and n, considered in three parts: noiseless, significant, and noisy communication. In noiseless communication there is an apparent contradiction to the feeling that information should increase as the size of the file increases. Balance is sought between the new information provided by the file and the extent of agreement with known information input to the file by the request. Problems in combatting noise, on the one hand, and maintaining the agreement with the requested information, on the other, must be reconciled. As n or F/N is increased, the

likelihood of finding a good match is increased but the new information obtained is decreased. The usual entropy measure for information is extended to include, as an equally significant feature, the relevancy of the information received --- determined, for example, by the degree of similarity to a request. The standard approach to noise is to assume a binomially distributed error based on the probability that an error might occur at any one bit of the n over which the match is made. The effect of this type of error source on file operation is best combatted by responding with a set of items rather than a single one.

Error in file operation can arise when any indexing structure is imposed on the file. An index can be constructed by establishing a "sequence of significance" on the bits and using successive bits as index criteria; by establishing class numbers, encompassing groups of bits, by factor analysis or a priori classification of terms; or by establishing groups of items, on the basis of similarity, and using a representative item as a group identifier. In the first method the process of searching can be interpreted as one of successively screening items from the file on the basis of matching the request with successive portions identifying the index term. It should be relatively simple to screen out the potentially relevant items for further analysis or to eliminate most items from detailed consideration. Concentration upon selecting the set of most probable items, by discarding all improbable items, is the key concept in the decoding strategy discussed, extending the methods of sequential decoding to file operation.

To reduce the expected average number of computations in such an index searching process, the methods of activity organizing the file can be used. The aim is to produce a hierarchical arrangement of levels of grouping which will represent a compromise among the various distribution measures so as to optimize the selected measure of efficiency. These levels of grouping are analogous to the structure of a normal classification scheme although their method of derivation is dependent upon the character of usage rather than a priori decision. The number of levels open at any time is dependent upon the distribution of probable activity and the aim is that the probable activity will be equal for all groupings which can be seen at any time. The sequence of scanning may be organized so as to look at the groupings in their order of probable relevancy, that is, given a measure of the probable activity, arrange the groups in order according to that measure so that the ones of interest will be seen sooner than under an organization which ignores this distribution.

IV-21. Mathematical Theories of Relevance with Respect to Systems of Automatic and Manual Indexing, by Donald J. Hillman, in Automation and Scientific Communication, Short Papers, Part 2 (papers contributed to the 26th Annual Meeting of the American Documentation Institute, Chicago, Ill., October 1963), Ed. by H. P. Luhn, Washington, D.C., American Documentation Institute, 1963, p. 323-324.

A program of research is underway to provide a theory of query-to-document relevancies in connection with problems of indexing. Several reasonable alternatives are needed so that experimental trials can help in deciding which alternative theory is most adequate. The problem of relevance is essentially that of characterizing the relatedness of concepts so as to make the characterizations independent of any notion of relevance determined by a given system. In one promising theory similarity-judgments are described in terms of a principle of conceptual matching. This theory results in a topology of concepts. Certain metrics can be defined on the basis of this theory to provide formulas expressing query-to-document relevancies.

Theories of Farradane, Maron and Kuhns (cf. III-18), and Baker fall within the scope of this study. All these theories are being examined from the standpoint of greater generality. Since their adequacy depends upon the results of experimental validation, testing procedures have been set up in a number of ways. For example, the mutual relevancies predicted by a theory are checked against such relevancies provided by expert human judgment. When these relevancies differ, a basis is offered for comparison and modification to be incorporated into new theories.

In another example, a theory of relevancy can be incorporated into a set of indexing instructions. Sets of such instructions have been set up and index terms have been assigned to individual documents. Results of searches based on these different indexing methods are being submitted to various specialists and users for comment as to the accuracy of the methods in retrieving relevant items.

These indexing methods are being programmed for the GE 225 computer for a small collection of documents. Retrieved documents will be judged by each theory of relevance incorporated in a set of indexing instructions for that collection. Results obtained by these methods will be compared with results obtained from human assignment of index terms for the same documents, to provide a realistic basis for discussion of automatic indexing.

The crucial problem of indexing is to assign symbols to documents so as to be able to find documents related to a request. Since the documents may not be indexed by the exact terms of the request, relevancy criteria must be used to retrieve documents in order of their relatedness to the request. For this part of the study a more extensive list of prescription-terms will be generated by the computer to retrieve more documents. To each term in the collection there will be assigned an "Association Factor", expressing its conceptual relation to prescription-terms. Relevance measures are then assigned to documents and accession numbers compiled with respect to these measures. The possible mechanization of this total process is being investigated.

IV-22. Criteria for Total Information System Evaluation, by Paul W. Howerton, in Information Handling: First Principles, Ed. by Paul W. Howerton, Washington, D.C., Spartan Books, and London, Cleaver-Hume Press, 1963, p. 195-207.

Any process now carried on by humans which is routine in nature or can be made routine can be automated, whenever it can be shown that automatic handling would save time, achieve greater efficiency, conserve space, or show financial economy. Studies which set forth conclusive comparisons of the effectiveness and cost of operating a going system with those of proposed systems may be prepared against criteria developed empirically.

General criteria are concerned with identification of those subsystems within the total information system which would be affected by adoption of the proposed system; and identification of records and statistical data needed to make an evaluation including descriptions of document holdings, intellectual processing techniques, coding schemes, servicing of document files and indexes, equipment used and personnel requirements. Preliminary steps can be summarized as: determining the points of similarity between the present and proposed system at each stage of the processing cycle; identifying the first point in the processing cycle to be affected by the change; summarizing all changes in the present system necessitated by adoption of the new system; setting forth advantages of the present system and of the proposed system; and making comparative tables of cost factors, showing requirements for manpower, space, time, and equipment.

As an integral part of the process for evaluation of proposed systems, a list of questions has been accumulated for use during the preparation of the preliminary study. The questions are grouped according to stages of processing and include: physical handling of documents; intellectual problems, such as to what extent the proposed system provides for recovery of relevant material, reduces redundancy and prevents false drops, and what effect the proposed system will have on indexing procedures and training of indexers; creation, maintenance and servicing of machine indexes, as to whether the proposed system advocates one arrangement of all inputs or a series of subsystems for special purposes, and whether manual retrieval of known documents will be possible; establishment, maintenance and servicing of document holdings, such as the structure of the document store and correction of existing files with new equipment; space requirements; conversion problems; human factors of manpower requirements and retraining of present staff; and flexibility.

Any proposal for a new system must not be so general and lacking in details as to prevent such tabulated comparison with a going system.

Systems for data processing, meaning the logical manipulation of information for analytical purposes, can be analyzed by most of these criteria. The definition of the desired end product is the first order of business with economic evaluations as the concluding step in the routine.

IV-23. Methodologies for System Design, by Hughes Dynamics, Report No. RADC-TDR-63-486, Final Report to U.S. Air Force, Rome Air Development Center, Los Angeles, Calif., Hughes Dynamics, February 24, 1964, 98 p.

This report presents the results of work on development, programming, and testing of methodologies for aiding information system design and evaluation. The approach has been the formalization of well-defined methodologies and associated general purpose computer programs; the aim has been to allow communication of the characteristics of file tasks and equipment in familiar terms with subsequent automatic derivation of quantitative evaluations of effectiveness. The work includes the complete programs for an "evaluation and assignment model" which provide for mechanized determination of the "optimum" assignment of components and functions to points in a hierarchical reporting structure; and test results on the relative effectiveness (in terms of quality of results and time and cost for the design process) of the measure of system efficiency, the system design model, and the evaluation and assignment programs.

The evaluational model relates the cost and effectiveness of the component parts and functions of a system to provide a single measure of its efficiency. The model is based on the following set of definitions: a job consists of a set of defined operations and a volume for each operation; a component is a defined means of performing some parts of some of these operations; the efficiency, a, with which a given component performs in the execution of a given operation is measured as a function of cost, time, and the size of the operation (i.e., $a = CT/N$, where C = units of dollars per time, T = time, and N = bits); and a total system, consisting of a set of components required to perform the set of operations, is represented by a matrix of their respective efficiencies (the "efficiency matrix" for the system). The problem is to find a reasonable measure of "total system efficiency" in terms of the relevant parameters a_{ij}, v_j, x_i, and c_i (efficiencies, volume, required costs and actual costs respectively). One approach is to use the normalized quadratic cost as the measure. A more interesting variant is to consider the generalization of the classical eigen value theory to handle a nonpositive-definite, nonsquare matrix A directly. A third and fourth measure of efficiency are provided by the reciprocals of one and the other of the

first two, corresponding to the reciprocal eigen value problems. Chebyshev-type measures can also be considered. None of these measures provides any consideration of actual cost, and may therefore be considered as theoretical efficiencies. In the relationship between theoretical efficiency and operating efficiency, three values are of interest: theoretical efficiency, $(A\bar{v} \cdot A\bar{v})/(\bar{v} \cdot \bar{v})$; operating efficiency, $(c \cdot Av)/(\bar{v} \cdot \bar{v})$; and rms cost per unit of rms volume, $(\bar{c} \cdot \bar{c})/(\bar{v} \cdot \bar{v})$. In effect, operating efficiency attempts to compromise between actual and theoretical costs.

There are a number of problem areas in the model as defined. The first relates to determining the efficiency values in the matrix and more fundamentally to measuring the concept of "complexity", to provide some consistent measure among functions. This problem is discussed somewhat more deeply below. A second class of problems relates to the variability in the basic volume data. Any average value is only one of the statistics describing a variable distribution and provides no reflection of the effects of peak load activity. It is felt that the model can be extended to accommodate these considerations by describing the volume vector in terms of the fundamental parameters of the distributions of its components. The third area of difficulty lies in the definition and selection of the functions and components to be considered in the systems design; they are assumed, in the model, to have been prespecified. The purpose of the development of a calculus of operations is to provide a tool for making these selections.

In the evaluation model, as it was developed and programmed, the measure of component efficiency was CT/N. Testing of such a measure has the purpose of demonstrating the degree of intuitive validity which the results of the measure provide when used for comparison of alternative devices. The approach here in testing the measure has been to apply it to various classes of searching methods and file storage components. The various methods of file searching include linear scan, indexing, and as a special case of indexing, logarithmic search. The efficiency of the first type appears to be characterized by the rate of data transfer; the efficiency of the second by the relationship between available access time and the capacity of the storage unit. A unified measure which provides a method for comparison among devices which are either linear scan or random access, or both, requires definition of a measure of file storage unit efficiency, in terms of C, T, and N: the number of "active characters" is defined as $N = (e_1 r)(e_2 R) = (e_1 e_2) F = eF$, where F = capacity of file, r = size of each record, R = number of records, e_1 = percentage of each active record which is of interest, and e_2 = percentage of active records. The transfer rate per active character = t, the average access time to an active record = T_a, and the total time to process the specified activity = T. Then for sequential access devices $\alpha = \frac{CT}{F} \frac{1}{e_1 e_2}$, and for random access devices $\alpha = \frac{C(e_2 RT_a)}{(e_1 e_2 F)} = \frac{(CT_a)}{(F/R)} \frac{1}{e_1}$

The number of active characters involved in a search can be calculated if the address position is known, if the address is not known but the file is sequential, and if the file is random. For files with total size F greater than the capacity of a file unit F_o, multiple units (for random access) or sequential replacement (for sequential access) are used. The C, T, and R are recalculated, and linear access efficiency becomes $\alpha(e) = \frac{1}{e} \alpha_o$ and random access efficiency $\alpha(e_2) = \frac{(F)}{(F_o)} \frac{1}{e_2} \alpha_o$. If α (e, F) is plotted for various devices, efficiency curves can be obtained. These show, for example, that disks are competitive over an apparently narrow range. The following are apparently valid: $e F > 10^5$ is a reasonable range of batch size for economic utilization of linear access, and random access is economic for $e F < 10^5$; present tape speeds begin to fail, for time reasons, to search a file $F > 10^9$ in a day; disk memories begin to fail, for capacity reasons, with a file $F > 10^9$.

The characteristics of various storage devices, both sequential and random access, have been tabulated, including the resulting efficiency values; the two types of devices have been consolidated for particular parameters of a problem, $F = 3 \times 10^9$, $e = 10^{-3}$.

One major question in the use of the measure CT/N is how to measure complexity consistently among functions. It is proposed to define the complexity of a function over all devices capable of performing the function. Several alternative methods for approximating the complexity from characteristics of a searching problem are possible: program it and use the number of decision instructions as the normalizing factor; number of "NOR" circuits in a special purpose device designed for the function; or Turing machine decomposition. Finally, the complexity of a device is defined as the most complex operation which it is capable of performing.

IV-24. Language Structure and Interpersonal Commonality, by Robert V. Katter, Report No. SP-1185/000/01, Santa Monica, Calif., System Development Corp., June 17, 1963, 30 p.

(The last 11 pages of this report are designated SP-1185/000/01A and dated September 17, 1963.)

Meta-theories of grammar consist of ideas about the nature of language behavior and communication, coupled with corresponding methods for analyzing language behavior to arrive at a structural theory. This paper takes a descriptive normative viewpoint toward meta-theoretical problems and develops empirical measures of the degree of interpersonal commonality of responses made by members of a language community to various language stimuli. One of these measures is of the commonality of predictive judgments made to pairs of elements in a long string (a corpus) without regard to successional effects.

In the field of extracting segments from a text to serve as representations for it, the procedure for extraction must (eventually) be fast and inexpensive but a criterion procedure for evaluating extraction results need not be; once segments of a corpus of text have been assigned criterion values, they can be used to validate many different extraction procedures. It is possible that no criterion measure currently in use is satisfactory because too much emphasis has been placed on obtaining inexpensive criterion measures. Among the procedures that have been proposed or employed, some as a method of extraction in one context and a criterion method for evaluating extraction results in another, are: (a) holistic valuative judgments, in which text segments (words, phrases, sentences) are selected and sometimes rank ordered in terms of their being judged representative of or relevant to specific texts; (b) quantitative valuative judgments, in which text segments are rank ordered or rated on a scale of relevance, or on a scale of probability of relevance to specific texts; (c) analytic descriptive judgments, in which text segments are judged as belonging (or not belonging) to predefined desired content rubrics of text, such as purpose, procedures, findings, recommendations, and in which a set of segments judged desirable from a single text are also judged and filtered for redundancy and coherence; (d) frequency based formulas, in which the absolute or relative frequencies of occurrence, co-occurrences, or contingent co-occurrences of segments are tallied for a text, and the decision to extract certain segments is a function of these frequencies; and (e) position based formulas, in which segments such as the title, section headings, and first and last sentences in paragraphs are designated for extraction. Historically, the most common evidence of validation has been for an investigator to present a procedure as an extraction method and then to note that the results appear to him to correlate well with his own criterion judgments: Baxendale, Borko, Doyle, Maron and Kuhns (cf. III-18), and Montgomery and Swanson (cf. II-36). A few investigators have used extended programs of interaction among members of a research team to produce high intersubjective agreement and then

used these group judgments to select text representations: Cleverdon (cf. I-2). A natural step was to try to use judgments of typical users as criteria, but studies of the degree of intersubjective agreement of judgments of typical users found it to be low: Rath, Resnick and Savage (cf., for example, II-45).

One source of this low intersubjective agreement among users may be that it is often not clear what is intended by the words relevant and representative. Considerations such as the validity of the material, its usefulness, style, understandability, etc., can all enter the judgments in unknown amounts. None of these considerations are the right ones for selecting text segments for indexing, classification, or abstracting purposes. It is important not to confuse valuation with representation. Descriptive analytic procedures overcome some of these difficulties but have two apparent shortcomings: none specify the relationship between the extracted segments and the remainder, and none measure the degree of intersubjective agreement of typical users on such a relationship. The concept and measure proposed here attempt to take account of these shortcomings.

The concept of mean commonality of predictive certainties poses a single retrieval-relevant question: when an extracted segment is presented to typical users, what does this cue do to their pattern of predictions regarding what other content segments would be contained in that document? A predictive pair (cue segment with predicted segment) has high commonality of predictive certainties when there is a high degree of inter subjective agreement on the predicted presence (or absence) of the predicted segment, and the average degree of subjective certainty is high regarding that prediction.

A format for eliciting responses from a sample of subjects (S) would include: Given sentence A from some article, check whether a sentence with a meaning close to sentence B would more likely be present or absent in the same article; check degree of certainty (from six choices) of that guess. The commonality of predictive certainties of a sample of S's to a single predictive pair is given by $C = \frac{\Sigma X}{5N_r}$, where X = response value from a single S with any value from + 5 to - 5, and N_r = number of S's giving one response each. C takes values ranging between + 1 and - 1. A particular cue segment may be paired with many different predicted segments, and C values obtained for these pairings. The mean commonality of predictive certainties associated with that cue segment is given by $\bar{C} = \frac{\Sigma /C/}{N_p}$, where /C/ = absolute C value and N_p = number of predictive pairs.

On the average, factually correct predictions should have significantly higher C values than factually incorrect ones; this hypothesis can be tested experimentally.

IV-25. Statistical Decision Criteria in the Evaluation of Information System Performance, by Charles H. Kriebel, ONR Research Memorandum No. 131, Pittsburgh, Pa., Carnegie Institute of Technology, Graduate School of Industrial Administration, August 1964, 33 p.

An information system is a service facility to the management of an organization: a "better" system will furnish "better" information, facilitating the execution of "better" decisions to improve performance overall. Aspects of statistical decision theory deal with the mathematical analysis of decisions when the outcome is uncertain. It is possible within this framework to evaluate explicitly the relative worth of different information in terms of a decision objective. The purpose of this paper is to indicate how comparable criteria can be developed which may be used to evaluate the performance of information and decision systems in modern organizations.

In a linear team situation the possibility for interaction between decisions by members is omitted and the requirements for an information system, as such, are nonexistent. If the payoff functional is quadratic in the action and state variables, the information system is distinguished as a component of "information structure" in the team decision model. Employing the prior-to-posterior framework of Bayes analysis, a measure of the economic performance capability of two simple information systems is illustrated. The first system (A), complete informational decentralization, corresponds to the case where each team member specializes his operating information to his own observations on a particular state variable and no communication takes place between team members. The second system (B), limited dissemination of information, corresponds to the case where members have the opportunity to revise their a priori knowledge about the state variables in subsequent periods on the basis of individual observations and messages received from other members.

For the team payoff expressed in terms of cost, the relative worth to the firm of employing a system A is the cost reductions that can be realized by including consideration for the variance in the state variables a posteriori. The system A exhibits constant returns to scale relative to the posterior variance of the first order component. The distinction between this system and the null information system is that in the null case no observations are made on the information processes, and hence there is no formal basis on which to revise the a priori distribution. Under the system A the decision makers possess the option of obtaining additional information by observing the state variables over time. Thus the a priori probability distribution can be revised at each stage of the process in light of the new observations.

In system B, each member communicates some function of his observations to a central staff which compiles all such information received and periodically disseminates this compilation to each member. Under this system, the posterior distribution is obtained with respect to the pooled observations for each period and the system exhibits constant returns to scale in both the variance and covariance terms of the posterior distribution.

In summary, information to the decision maker will be more valuable or "better" if it provides a correspondingly larger reduction in the relative amount of uncertainty present, where reductions in uncertainty are expressed relative to the certainty counterpart, perfect information. Ability to evaluate the performance of different information systems relative to a particular organization is limited only by the ability to describe formally the expectation computations conditioned on the information provided by these systems.

IV-26. System Effectiveness Study, by J. L. Kuhns, Section 7 of Research on an Advanced NASA Information System, Report No. NASw-538, Canoga Park, Calif., The Bunker Ramo Corp., October 11, 1963, 16 p.

We shall regard an information system as a three-term relation between (1) an input consisting of a user's information requirement, (2) a collection of information "packages" (e.g., documents, abstracts, titles of documents, photographs, tabular data, etc.), and (3) an output consisting of a certain selection of the collection. Within the scope of the present study, time will be omitted as a variable. We equate the information requirement with the communication of the requirement to the system and we assume that this communication is in natural language. This allows us to treat the input and output of the system as entities of the same logical type. To complete the model, we next assume that there is a referee who can decide for any item in the collection not only whether it is relevant to the information requirement or not, but also decide between any two relevant items whether one is more relevant than another. These comparative evaluations are reflected in a numerical scale of degree of relevance. Similarly, it is appropriate to regard the

selection procedure as a procedure for assigning numbers to items. (These numbers can be regarded as the system-calculated strength-of-match between collection item and information requirement.)

Development of a measure of retrieval effectiveness must take into account both the proportion of relevant material retrieved and the proportion of retrieved material relevant. We look for ways of comparing a result with the perfect retrieval. The measurement of this effectiveness requires the specification of the values of a certain function of these variables: $f(n, x)$. We define $f(n, x) = \frac{kx - mn}{m(k - m)}$, where n = number of items retrieved, m = number of relevant items, x = number of items relevant and retrieved, and k = value of n at which retrieval is deemed worthless, based on the user's personal estimate of the utility of the information required.

Let us now look at the problem from a completely different viewpoint. We will examine a hypothetical situation of system usage and define our measure of retrieval effectiveness to depend on the actual cost to the user of a single search. Thus $f(n, x) = \frac{V}{V_{max}} = \frac{\text{net gain}}{\text{maximum possible net gain}}$. It is easy to show that the quantity $1/V_{max}$ can be interpreted as a penalty coefficient for irrelevant retrieval. We rewrite $f(n, x)$ in the simple form $f(n, x) = R - pI$. The quantity R is the proportion of relevant material retrieved, I is the effective amount of irrelevant material retrieved (in the unweighted case I is simply the number of irrelevant documents), and p is the irrelevance penalty and may take on arbitrarily assigned values. The use of this measure of retrieval effectiveness has been reported elsewhere (cf. II-53).

The measure appears to be a complete solution for the retrieval effectiveness of a system acting on a single query. The problem now is to extend this to an evaluation for an entire set of queries. It is quite plausible to assume that in the entire set of queries the sets of numbers are independent. The definition of independence that appears appropriate here is that the coefficient of linear correlation between the number sets is zero. This will be the case, for instance, if the queries are sufficiently uniform in value so that p is constant; or that the file content and retrieval procedure is sufficiently stable to yield constant I. This allows the computation of the mean-value score: $\bar{f} = \bar{R} - \bar{p}\bar{I}$

To compare two systems, a decision must be made based on an estimate of the magnitude of $\bar{p}$. Recall that $\bar{p}$ is the cost to the user of processing a single retrieved item relative to the maximum net gain possible from the search. Exact values are not necessary, but only gross estimates. For example, in a bibliographic search which produces abstracts the value of $\bar{p}$ can be taken as rather small. Another consideration is that estimates can be obtained by an intuitive evaluation of the equivalence of search outputs in particular situations. Such considerations will allow estimates of the range of value of $\bar{p}$. Although subjective, it is plausible to assume that there will be some uniformity among different evaluators.

In the case of a ranked retrieval the first step is the interpretation of what constitutes the output. Recall that we have been compelling the user to process all output items; we now permit the user to process the output items successively, and choose at each step whether or not he will proceed. Then we can define the output to consist of the processed items, and the effectiveness can be measured by the same graphical scheme as before. This seems reasonable as it certainly incorporates the ranking data. However, new complexity is introduced into the implementation of the method; namely, test participants are required to process the output.

A second approach is to correlate the two rankings given by perfect retrieval and by the selection procedure. We must "weight" the rankings, however. The decision to continue processing will be based on the information already obtained, and the gain to be obtained by continuing will be reduced accordingly. Thus p will be a non-decreasing step function, which increases after each relevant item is processed.

The steps in implementation are to select an appropriate sample of user information requirements, a test "library", and judgments of perfect retrieval.

It is recommended that this tool be given consideration as part of a systematic procedure for system improvement and quality control.

IV-27. Some Approaches to the Theory of Information Systems, by Borje Langefors, BIT 3, 229-254 (1963).

A systematic technique for establishing the real needs for information within an organization has to define the information needed, its volume, the time intervals at which it is required and at which it is available, the data from which it can be produced, the process needed for its production and the form for presentation of the results. This analysis, basic to any rational design of an information system, must be hardware independent. Only after such an analysis has been made it is timely to consider the common characteristics of hardware systems in general and then the suitability of specific hardware systems. At this phase of the analysis it is appropriate to consider the impact of the organization of information on the total work load and on the data processing system and try to find an optimal --- or at least efficient --- solution.

One problem of organizing data is concerned with the data transport needed to bring data from the storage place to the processor. Another is to organize data in a way that is meaningful for people. It seems appropriate to consider such data as are stored in the main memory of the processor to be used for processing as being available without transport work. The problem of data organization then is how to store data in mass storage so as to minimize the work needed to transport these data to the main store when needed or to make them available within the time required.

When basic functions and information sets required for these functions have been found, the sets must be defined in terms of more elementary concepts. Thus when a set of information sets b, c, and d is needed in order to produce an information set a, then b, c, d constitute the precedence set for a. Or $\rho(a) = b, c, d$; a is the succedence set of $\rho(a)$. It is always supposed that the precedence relations include updated versions of the information sets and that precedence relations are grouped by subset relations. Precedence relations can be specified by a precedence matrix, where precedents are listed to the left of the matrix and succedents are listed above their associated columns. Some information sets are terminal, some are initial, and the rest have an intermediate position and can be ordered by levels. If the precedence matrix is rewritten to reflect this ordering, it can be used to draw a corresponding graph of the relations of information sets.

The set of all feasible operations which uses $\rho(a)$ to produce some approximation to a is called computation, written Comp (a). To each precedence relation there is associated a computation, for the implementation of which there may be different systems of actual procedures. If the computations are all taken as separate computer runs, there will be a multiplicity of data input transport for all files. Some computations may call for multiple scanning of some files. If the (relative) volume or scan time for each file is indicated along with the precedence matrix, it will be possible to see how much transport is saved by taking together two or more computations into the same process. The system so

described is called the basic topological system. The computations system implies a (relative) transport volume for input against the volume for single input, giving a topological transport factor. The multiple transport can be reduced by taking some computations into the same computer process, if the computer memory is sufficient to store the programs. Grouping computations corresponds to grouping columns of the precedence matrix, and it is easy to see how much transport is saved by grouping any computation with any other one.

Grouping of computations may call for shorter data blocks in the files in order to leave memory space for the group programs. This will then lead to an increase of transport volume which will also have to be considered. In general it will be desirable (and possible) to group more than two computations. One will have to test for different combinations of computations where the memory space and the space requirements for the different programs have to be considered. The subset relations give valuable aid in suggesting efficient grouping.

The precedence matrix also gives information about the input and output (transport) requirements. The amount of transport equipment with a computer has to be small for cost reasons. Several of the elementary files will have to be consolidated to larger files, corresponding to grouping of rows of the precedence matrix. Consolidation of files may cause excessive data transport; grouping of computations on the other hand may eliminate such excessive transport. When we consider consolidation of files, the picture of excessive data transport becomes somewhat complicated, being composed of duplicated input of some files to several processes and of deadweight transport of some data in these files.

The precedence relations and precedence matrix are suitable to describe some of the relations in this study, but cannot specify the number of file passes in a process for both input and output files and do not offer a convenient description of how computations are grouped to composite processes. To remedy this, the concept of incidence between process and input and output is introduced, and the precedence matrix is replaced by the incidence matrix. One row is taken for each process (or computation) and one column for each data set. The number of scans that a file is treated by a process is used as the incidence number between the file and the process.

A change in the graphical representation may be introduced which is analogous to the change in passing from precedence matrix to incidence matrix.

In designing the information system we have to consider limitations imposed by hardware cost. Filed data, in a storage much slower than the main memory of the computer, are transported to and from the main memory. Limited memory space may make it uneconomic to group too many computations together, because all necessary programs cannot be kept in the main memory simultaneously and have to be transported to and from the memory. This may be a larger transport volume than the data transport it saves. Again, the number of file storage (or data transport) units cannot be chosen large enough to transport all data sets separately, which may be a reason for not grouping some computations although this might have saved data transport. Finally, using memory space to group computations and thereby reduce the number of file scans may necessitate shorter data blocks, which will increase data transport by increasing the number of interblock gaps which are a source of deadweight transport.

Optimum design of the data structure for a system requires establishing different feasible solutions and finding one among them which corresponds to minimum transport. In trying to find a feasible solution it appears that a lot of programming for earlier procedures will have been done before it becomes clear that a restart will be necessary. Therefore the manner of analyzing for feasible solutions using the incidence matrix or graph may save man years of programming. In order to estimate the quality of a solution

reached this way the resulting data transport volume is compared with that of the system without transport limitation and with the basic topological system as well as with the theoretical minimum.

An example, using fairly realistic data volumes and size relations between input data, programs and memory space, illustrates the nature of the problem when small memories are used.

IV-28. Information System Design Computations using Generalized Matrix Algebra, by Borje Langefors, BIT 5, 96-121 (1965).

Information system design implies two main problems: deciding on capacity level and designing the most economic system which realizes this capacity level. The common way of designing systems is to choose a capacity level, then do an ad hoc design job, involving much programming and reprogramming. This paper is concerned with some methods for designing the system by using analysis and design methods which minimize the cost for reprogramming and restructuring during the design phase. The methods, contained in an earlier paper (cf. IV-27), can be described as generalized matrix operations. The matrix operations involve the incidence matrix, or modified versions of it, which resulted through earlier operations, and also the vector defining file transport volumes as well as matrices which specify the process groupings and the file consolidations wanted.

The matrix routines can easily be performed on a computer if a suitable generalized matrix program system is available. This is a convenient method for comparing different design alternatives. To choose those design alternatives to be tested, system properties which reduce the number of processes or files that would profit from being joined are utilized. Thus the "manual" choice of design alternatives becomes tractable. In that case the possibility of automating the calculations made necessary by the matrix operations becomes a great advantage.

A complete formalization (and hence possible automatization) is established of the computation of transport saving --- or resulting transport amount --- for any specified process grouping (if also specification is given of how possible multiple file scans of a process are to be interpreted). In order to develop this procedure into a fully automatic routine an algorithm for selecting the set of different groupings to be tested must be established. In many practical systems the number of feasible groupings is fairly small. Therefore, a formalization of the selection of feasible groupings may make possible a complete analysis of such combinations. In such cases a fully automatic solution of this design problem would have been achieved. For more complex systems, full automatization may not be reached. Then such a procedure would still be useful, for it will reduce the amount of intuitive --- or "manual" --- choice of the different designs to be tested. It also seems useful to test a random selection of the set of feasible groupings to define the best of them which, while not exactly optimal, may well be a better solution than would be obtained from a few intuitive choices.

The result of the recommended analysis can be displayed in a matrix with one column and one row for each process and where each element in a column measures the transport saved by grouping the process for the column with that associated with the row position of the element. The procedure reduces the choice for system design to an extent which is dependent on the topological properties of the basic information system (or of the incidence matrix). Another procedure for further reduction of the choice problem, a logically necessary operation, is related to what conditions must be satisfied in order that two processes may at all be grouped together. The precedence matrix described earlier (cf. IV-27) should be sufficient for testing such conditions.

In addition to those properties of an information system which have been used so far to reduce the number of design configurations to be tested, one significant property remains --- the different sorting sequences used with different files. It appears realistic at the present state of the art to regard file sequence to have been specified before the analysis starts. The sorting sequence of the different files will also define a partitioning of the set of processes into one subset for each set of file sequences. In most practical systems this sorting condition will bring about a very significant subdivision of the system into much smaller subsystems, which are separated with respect to the process grouping analysis and with respect to file consolidation problems.

To keep input-output equipment requirements within economic bounds it is necessary to consolidate small elementary files into larger and fewer ones. The problem of efficient file consolidations is a problem of general importance for information system design, whether it is built on serial access memories, random access memories, or both. The incidence matrices for a system and its modification after file consolidation can be compared to find rules of general validity for a prescribed consolidation of two standing files. In general, file consolidation increases the amount of file transport but most or all of the increase in transport time can be eliminated by suitable process grouping. Thus the potential for file consolidation that a certain grouping carries with it has to be considered in combination with the reduction in transport it brings. A procedure for aiding design in this stage would compute equipment reduction. Assistance in selecting candidate groups of files to be tested for possible consolidation can be obtained from the computer, in analogy with the grouping phase. Such consolidations can be found by scanning the incidence matrix --- similarity of columns in the matrix indicates feasibility for consolidation and can be printed out by simple algorithms analyzing the matrix.

IV-29. The Intensive Sample Test for the Objective Evaluation of the Performance of Information Retrieval Systems, by Calvin N. Mooers, Report No. ZTB-132, Report to Rome Air Development Center, Cambridge, Mass., Zator Company, August 1959, 20 p.

The method of retrieval system testing proposed here has the following features: it can give an absolute determination of retrieval system performance rather than a relative determination, since it does not depend upon comparison with the performance of some other retrieval system; performance is measured against information known to be in the file prior to the retrieval experiment; and relevance of all the citations taking part in the test procedure is settled before the test begins. In an ideal situation, we should like to test the performance of a retrieval system against a hypothetically perfect system. A perfect system can be defined as that one which would produce for each customer and his inquiry the exact set of documents (no more and no less) which the customer himself would select if he were able thoroughly to read the entire library and then were to pick out the documents of relevance to him and his inquiry. It should be emphasized that the perfect retrieval system is defined only for the population of documents and for the population of customers-with-inquiries that are settled on beforehand. It is clear that we shall never be able to achieve such a system for any reasonably large collection and group of customers. But we should consider methods of approximating a perfect system, and some kind of approximation to the performance of a perfect system for an actual retrieval system to be tested.

A sampling method is suggested as a device for these approximations. Samples are chosen from the population of documents, a sample size not too large, as the time, effort and trouble in administering the test will increase with sample size; samples are also chosen from the population of customers, with care, since the value of the test depends upon the intelligence of these people. Each sample customer is given a few of the samples

of documents, and asked to read them carefully and formulate a set of inquiries such that at least one document provides relevant information for each of his inquiries. The question should not be framed in such a way that the document alone will provide the entire answer to the problem; also, the question should not be tailored closely to the peculiarities of the document. This procedure leads to a set of document-to-inquiry associations that is a small but perfect retrieval system. To make possible the analysis of test results, a scale of relevance is provided for each of the question-document pairs: not relevant, relevant (this document could be omitted in the search without too great loss), and crucial (the omission of this document would be a serious loss).

The set of inquiries is now given to the operators of the actual retrieval system, who do not know what samples have been drawn. They may process the questions in any way but must of necessity process their entire file for each question. Termination of the test shall be at the option of the system operators. The documents retrieved are checked to find which of the sample documents have been retrieved and in response to which of the questions. The following three test ratios are computed: crucial documents ratio (number of crucially relevant documents cited to the number in the sample), relevant documents ratio (number of relevant or crucially relevant documents cited to the number in the sample), and not-relevant documents ratio (the number of not-relevant documents cited to the number in the sample).

Excellence of performance is indicated by a high crucial documents ratio, very close to the value of one, along with a low not-relevant documents ratio, which should be less than 0.01. The relevant documents ratio should be close to one in almost all cases. Only a sample of the performance of the actual system is taken, but the sample matches completely the coverage of the approximating-perfect system and makes the analysis of the test simpler.

IV-30. Evaluation of Science Communication Systems, by Lawrence S. Papier, CRDL Special Publication 4-65, Edgewood Arsenal, Md., U.S. Army Edgewood Arsenal, Chemical Research and Development Laboratories, February 1965, 49 p.

All appropriate factors within specific systems, as well as external factors and their relations, must be identified in order to evaluate information systems. Evaluation is part of the process of analysis and design; the fact of multiple criteria and complexity must be recognized and accepted. Primary emphasis in the general method proposed here is on storage and retrieval aspects; specialized technical and intellectual problems, as well as detailed operating problems, are excluded. Criteria for evaluation are derived from consideration of management objectives, user requirements, system properties, system efficiency, and technical and operating aspects.

Primary organizational objectives include the growth and increase of efficiency and effectiveness of the organization. More specifically, the organization needs to attract and retain productive scientific personnel, to improve the quality of information supplied to those personnel, to cut costs, to develop new products, and to meet competition. These objectives can be translated into specific and useful criteria for evaluation of scientific communication systems. Management must specify objectives and requirements of its system in order to be clear as to what constitutes improvement in quality of information furnished. In cutting costs the issues of system requirements, user requirements, long and short range effects, and the interrelationship of objectives must be considered. The organization of a system designed to aid in the development of new products will be different from one designed to raise general skill levels or to meet competitive challenge. Again, management must clearly specify the aims of the system in order to simplify evaluation.

The direct goal of a science communication system is to fulfill the requirements of its users or, more accurately, the wants of its users. There have been several attempts at expressing user requirements from various standpoints, but efforts to determine general significance of the factors is difficult. Significance and weight vary in individual cases. However the items can be grouped into several major categories: form of material and oral contacts (books, reports, meetings, etc.), control tools (reference works), types of communication (for example, data and information), and properties of system response (accuracy, currency, completeness, etc.). The listings can be summed and combined into a coordinate grid simplified into a study of the properties of system response for the various types of communication. The grid may be detailed to any degree necessary. The nature of properties and communication must, of course, be understood before their ability to fulfill user requirements can be determined.

The evaluation of a system must include not only its effectiveness in meeting organizational objectives and user requirements but also its efficiency in so doing. Evaluation of efficiency relates to cost, time and volume studies. System properties of compatibility, adaptability, and capability of self-improvement must be studied, as well as operating characteristics and feasibility. In the proposed analysis, areas of evaluation include ability to meet management objectives (discussed above), dealing with external relations, ability to fill user requirements (discussed above), system properties, and system efficiency. The area of external relations works with questions of services, cooperation, and geographical proximity; it also includes relations with, for example, purchasing, the patent department and computation center, as well as problems of interlibrary loans and compatibility factors. The system properties to be considered are adaptability, vitality, improvement and self-correction, and multi-purpose capability (browsability and accurate retrieval, for example). For system efficiency the total system is divided into parts and each one analyzed for the factors mentioned above. Cost, time and volume figures are collected; material, personnel and equipment are examined; and linkage and flow of material are detailed. The analysis may proceed until unit operations are reached.

Science communication systems are complex; evaluation should be neither strictly intuitive nor over-simplified. The analysis suggested here begins with an investigation of the objectives of the total system, broadens into the seven areas, divides into an analysis of subsystems, and proceeds in increasing depth. The final evaluation includes a determination of vital considerations and performance for each area, between areas, and for the total system.

IV-31. Operational Criteria for Designing Information Retrieval Systems, in Machine Literature Searching, by James W. Perry, Allen Kent and Madeline M. Berry, New York, Interscience Publishers, Inc., 1956, p. 47-48.

It may be said that an information retrieval system embracing a totality of "n" documents will direct our attention to a smaller number of "m" documents as being of possible pertinent interest. Personal inspection will, in general, reveal that a still smaller number of "w" documents are of actual pertinent interest. We thus define the following quantities: n = number of documents embraced by a given system, m = number of documents which use of the system indicates to be of possible pertinent interest, and w = number of documents that are found to be of actual pertinent interest by personal inspection of the m selected documents. Definition of these quantities leads immediately to consideration of certain relationships between them: $\frac{m}{n}$ = fraction of total documents to which attention is directed, the "resolution factor" which may be expected to have a different value when a given system is used to direct attention to documents of possible pertinent interest for dif-

ferent information requirements; $\frac{n-m}{n}$ = fraction of total documents from which attention is diverted, the "elimination factor" which obviously becomes larger as m/n becomes smaller and vice versa; $\frac{w}{m}$ = fraction of documents to which attention is directed that are found on inspection to be pertinent, the "pertinency factor" which becomes larger as the proportion of pertinent documents increases among those to which attention has been directed by the system; $\frac{m-w}{m}$ = fraction of documents to which attention is directed that are found, on inspection, not to be pertinent, the "noise factor" which may be used advantageously in estimating time lost in reviewing documents not even of marginal interest to which the system has directed attention; and x = number of documents of pertinent interest which are among the totality n embraced by a given system, important in evaluating the reliability with which a system directs attention to documents of pertinent interest. In determining x, considerable time may often be saved by applying statistical analysis to properly selected representative samples of the totality of n documents.

Relationships between x, w, and m lead immediately to definition of additional operational criteria: $\frac{w}{x}$ = fraction of pertinent documents to which the system directed attention, the "recall factor" which measures the proportion of pertinent documents to which the information retrieval system directed attention when a given search was conducted; and $\frac{x-w}{x}$ = fraction of pertinent documents to which the system failed to direct attention, the "omission factor" which measures the proportion of pertinent documents from which the system diverted attention when a given search was conducted.

Effectiveness and efficiency of an information retrieval system as an aid in meeting an information requirement cannot be measured solely in terms of either the recall factor, w/x, or the pertinency factor, w/m. Both factors must be taken into account. Proving over-all effectiveness must be based on evaluation of the performance of the system for a statistically significant sample of varying information requirements. Each of the information requirements, by virtue of its being different, would require that a corresponding search be made. The scope, both of each information requirement and of the corresponding search, would be defined in terms of different criteria, taken singly or in combination. For each search, the values for m, w, and x would be determined and the corresponding values for w/m and w/x would be calculated. Examination of the data so obtained would reveal that a certain percentage of searches would have values for w/m within various ranges of values of w/m. A corresponding plot of data for w/x would be the same. The percentage of searches corresponding to each of the selected increments in the value of w/x would be determined. In general, it may be said that high average values of w/x (the recall factor) will characterize an information retrieval system that functions with high reliability. With such a system there is high probability that attention will be directed to all documents of pertinent interest. Correspondingly, consistent high values for w/m (the pertinency factor) indicate that attention is not being misdirected to documents of no pertinent interest. The selectivity of such a documentation system may be said to be excellent.

In accord with well-known statistical methods the data may be used to plot frequency distribution curves. Such curves may be characterized by various parameters, e.g., arithmetic mean, standard deviation, skewness, kurtosis. These parameters may then be used to characterize the performance characteristics of the information retrieval system under investigation.

IV-32. A Technique for the Organization of Large Files, by Ralph M. Shoffner, American Documentation 13, 95-103 (January 1962).

The technique presented here relates the indexing function and the physical system characteristics to the time and cost of retrieval. An important feature is that the technique, being quantitative, focuses attention on some of the system characteristics which must be measured and related before the system operation can be considered satisfactory on the basis of a technical, as well as conceptual, knowledge of information retrieval.

A retrieval file consists of a set of basic file elements, R, and additional elements used to refer to the R elements for search purposes. The individual elements can be brought together in groups and then be retrieved through their group membership. Thus the file may be represented by a tree structure, the terminal nodes representing the elements of R while the non-terminal (or reference) nodes represent the groups. A search provides one of the elements of R as a response. The process of searching the file is repetitive and the three basic operations are: acquire, process and select. When a terminal node is selected, the search terminates with this R element as the search result.

An efficient system is one which accomplishes a task with minimum expenditure of effort measured in terms of the cost incurred. The task will be taken as a single search on the file. The cost of a search, CT/N, is defined as the product of the time per search, T/N, and the system cost per unit time, C. The cost and time are each defined as functions of the system characteristics: facility costs and operating rates, personnel costs and operating rates, and the sequence of operations. The average time for a search as a function of the defined search tree is expressed as

$$\frac{T}{N} = \sum_{i=1}^{L} \sum_{k=1}^{g_i} Pr_{i,k} \left(t_{a_{ik}} + t_{r_{ik}} \right)$$, where Pr_{ik} = probability of selecting k^{th} group, level i; $t_{a_{ik}}$ = time to access k^{th} group, level i; $t_{r_{ik}}$ = time for selection process k^{th} group, level i; g_i = number of groups on level i; and L = number of file levels.

Those parts of the system which may be changed are varied to obtain the minimum value of CT/N, which defines the most efficient system obtainable within the given constraints of the problem. Given an estimate of the file activity, the tree structure may be modified to reduce the average time required for a search. The order in which the nodes within a group are processed may be revised, allowing a reduction in either or both of the access and process times (activity organization). Alternatively, terminal nodes may be moved higher in the search tree, to allow the possibility of terminating search without processing all levels of the file (hierarchical organization). Activity organization does not affect the tree structure. The cost function, C, has a constant value and in CT/N occurs when T/N is minimum. For a given structure, min (T/N) occurs when the elements with the highest probability are highest in the tree. For hierarchical organization, the highest probability elements are considered first. Where one is attempting to determine the file structure, min CT/N does not necessarily occur at min T/N. The criterion for moving the k^{th} element from level n to n-1 is $CT/N(k, n-1) < CT/N(k, n)$. All elements are first assigned to the lowest level of the file and a minimum CT/N is obtained. Moves of the elements to the next higher level are evaluated in order of their probability. When the criterion is violated, no other items on that level need be evaluated due to their lower probability. Evaluation is then carried out for the next higher level. After the moves to a new level are made, the evaluation starts again at the lowest level and moves up through the levels. The procedure terminates when the element with the highest probability violates the criterion, or when no more file levels are available.

To demonstrate the effect of the tree structure upon an explicit equation for CT/N, a model file is considered, file relations are derived from assumptions about the file, and an equation obtained using those relations:

$$\frac{CT}{N} = \left[\sum_{i=1}^{L} (a + mn_i)\left(1 - \sum_{j=1}^{i-1} Pr_j\right)\right] \cdot \left(dR + eS + \sum_{j=1}^{J} F_j\right)$$ where a = constant accessing time; m = accessing time per node; n = nodes per group, level i; r = processing time per node; Pr_j = probability of search termination on level j, d = storage cost per element per unit time; R = number of terminal elements; e = storage cost per non-terminal element per unit time; S = number of non-terminal elements; and $\sum_{j=1}^{J} F_j$ = other facility costs per unit time.

Interpreted as an expected value, the equation is a basic model of information retrieval. Then values a, mn_i, and rn_i are the probability weighted average of the operating times experienced on level i of the file. The equation allows both activity and hierarchical organizations. The function is minimized to provide the least cost search. Systems with different equipment are then compared on the basis of their respective minima. In such systems, with specified equipment, the file organization (n_j, a, m, k_i, Pr_j, and r) determines the minimum cost.

The technique is demonstrated on a sample system, a rotary card file, and some comments are made concerning computer files. In application of this model to a specific system, two problems are of potential importance. Two assumptions of the model were: that a search produced one and only one file element, R, and that the retrieval system was integral and independent of any larger system. In systems where either assumption is inapplicable, the functions will have to be modified.

IV-33. Information Retrieval Systems, by John A. Swets, Science 141, 245-250 (July 19, 1963).

(The material presented in this paper appears in fuller detail in Measures of Effectiveness of Information Retrieval Systems --- A Review and a Proposal, by John A. Swets, Report No. 982, Report to Council on Library Resources, Cambridge, Mass., Bolt Beranek and Newman, March 1963, 23 p.)

Various measures for evaluating the performance of information retrieval systems have been suggested. Bourne et al (cf. IV-11) recommend determining a measure of agreement between an aspect of system performance and a related user requirement for each of 10-12 requirements, weighting the measures according to the relative importance of the requirements, and summing the figures to obtain a single figure of merit. Bornstein (cf. IV-10) proposes consideration of four variables: number of pertinent, partially pertinent, peripherally pertinent, and non-pertinent responses to each question; time spent in examining materials in each category of pertinence; proportion of acceptable substitutes for hard copy in each category; and coincidence and uniqueness of responses on the scale of pertinence. Analyses of variance for the first three variables will show relative measures of efficiency and effectiveness in the retrieval systems being compared. Wyllys (cf. IV-34) derives a single figure of merit by obtaining the product of four variables: the "restriction ratio" or reduction in potentially pertinent items effected by the search tool; the cost of using the tool, weighting such factors as time, money, inconvenience, and indexing costs;

the number of pertinent documents eliminated from further consideration; and a loss function incorporating the degree of pertinence of the pertinent documents eliminated.

Verhoeff et al (cf. III-30) would maximize the measure for a given system by defining a critical probability for each item relative to each query, and by retrieving those items having a probability of pertinence greater than the critical probability (pertinence is defined by each system user). Swanson (cf. II-53) proposed a measure M = R-pI, where R is the sum of relevance weights of retrieved items divided by the sum of relevance weights of all items; I is the effective amount of irrelevant material, defined as N-RL, where N = the total number of items retrieved and L = the total number of pertinent items in the store; and p is a penalty which takes on arbitrary values. Borko (cf. IV-9) modified Swanson's measure by redefining the irrelevancy score I, as the number of nonpertinent items retrieved, without the arbitrary penalty on non-pertinent retrievals. Mooers (cf. IV-29) proposes three ratios: the number of crucially pertinent items retrieved to the number of crucially pertinent in the store; the number of pertinent or crucially pertinent retrieved to the number of pertinent or crucially pertinent in the store; and the number of non-pertinent items retrieved to the number of non-pertinent in the store. That is, Mooers proposes consideration of three conditional probabilities: an important hit, a hit, and a false drop. Perry and Kent (cf. IV-31) focus on two quantities, the conditional hit probability and the inverse hit probability. Cleverdon (cf. I-2) considers the same two variables, which he calls recall ratio and relevance ratio respectively. He emphasizes the relationship between the two (as the relevance ratio increases the recall ratio can be expected to drop) and speculates that a highly restrictive query would lead to a high relevance ratio and a low recall ratio, and a less restrictive query would increase the recall ratio at the expense of the relevance ratio. Swanson's second proposal plots the percentage of pertinent items retrieved against the number of non-pertinent items retrieved.

The measure proposed here is supplied by statistical-decision theory, and provides an index of effectiveness that is invariant over changes in the breadth of the search query or in the total number of items retrieved. (The technique has the drawback, at present, that the model on which it is based has not been validated in information retrieval problems.) As an analogy to decision theory, a retrieval system takes a measure of a given item in the store, relative to a specific query, and assigns the item to one of two categories --- the retrieval system rejects the item as not being pertinent, or it retrieves the item. It is assumed that the retrieval system assigns a fallible index of pertinence to items, that the values of the index assigned to non-pertinent items vary about a mean and the values assigned to pertinent items vary about a higher mean, and that the two distributions, which overlap, are normal and of equal variance. Complete description of the retrieval system's performance can be obtained from an operating characteristic curve (as used in statistics) calculated on the basis of these assumptions. A family of theoretical operating characteristic curves can be drawn, and the parameter of this family of curves will serve as a measure of effectiveness. The slope of the curve at any point serves as an index of the particular acceptance criterion, and of the breadth of the search query, at that point; the difference in slopes of two points is a measure of the effective change in the breadth of the search query. The validity of this technique can be tested by determining the operating characteristic curve experimentally. A large amount of data is needed but there is no substitute for adequate data to evaluate retrieval systems empirically.

IV-34. Document Searches and Condensed Representations, by Ronald E. Wyllys, in Joint Man-Computer Indexing and Abstracting, Report No. Mitre SS-13, Bedford, Mass., The Mitre Corp., 1962, p. 37-60.

This paper will attempt to touch upon certain problems involved in both the theory and the implementation of semi-automated document-retrieval systems. We shall view condensed representations as running the gamut from a one-word subject-heading at one extreme to a book condensation at the other, with such document representations as traditional library tools and coordinate indexes falling in between these extremes. We begin by asking a question: What are the purposes, or the functions, to be served by various kinds of condensed representations of documents? There are at least two functions: the function of serving as a search tool, and the function of revealing the essential content of the represented document. In the present state-of-the-art of computers and of information storage and retrieval, it is more practical to try to automate the first function than the second. Document-retrieval will, in general, be a multi-stage process. The theoretical generation solution to the document-retrieval problem is expressed as the discovery and use of that sequence of search tools which, under the constraints of the given document-retrieval system, most rapidly restricts the set of potentially pertinent documents while still retaining all actually pertinent documents within the set. This theoretical solution is immeasurably easier to state in words than to achieve in practice. Practical solutions will have to take other factors into account, e.g., the search time, the search cost, the initial indexing cost, the desired level of accuracy, and the feasible level of accuracy.

It would be desirable to be able to talk about the "effectiveness" of a search tool in terms of how much it cuts down the size of the set of potentially pertinent documents, how much it costs in time or money to use, and its accuracy in not eliminating actually pertinent documents. A possible measure $E(t_k)$ of the effectiveness of search tool t_k is $E(t_k) = (n[D(t_k)] / n[D(t_{k-1})])\, C(t_k) L(p[D(t_{k-1}) - D(t_k)])$, where t_k = the search tool used at the k-th stage in a search process; $D(\cdot)$ = the set of documents remaining after the tool "$\cdot$" is applied; $n(\cdot)$ = a function whose value is the number of documents in the set "$\cdot$"; $p(\cdot)$ = a function whose value is the number of actually pertinent documents in the set "$\cdot$"; $C(\cdot)$ = a function whose value is the cost of using the tool "$\cdot$"; and $L(\cdot)$ = a loss function. This measure is so defined that a low effectiveness number corresponds to an effective search tool and a high effectiveness number to an ineffective tool.

The cost function could be thought of as a function of time, money, inconvenience, or other similar factors. Costs will be specific to the peculiar requirements of a given system. Here is a general framework within which to deal with individual cases. With regard to the loss function, consideration should be given to the relative pertinence of the retained documents as opposed to that of the eliminated documents. In other words, the degree of pertinence of the eliminated documents should be a factor in the definition of the loss function. It is interesting to note that both Swanson (cf. II-53) and Borko (cf. IV-9) include an unspecified irrelevance penalty factor in their retrieval scores. Unfortunately, the pertinence of a document to a request cannot be strictly and accurately defined, which leads us to the observation that the actual degree of relevance of a document to a given question can never be measured by anything other than the questioner's judgment. Verhoeff, Goffman, and Belzer (cf. III-30) on the one hand, and Maron and Kuhns (cf. III-18) on the other, propose statistical estimators that seem intuitively related to the degree of relevance and are useful concepts.

If several tools are to be compared for use at the k-th search stage, the factor $n[D(t_{k-1})]$ will be the same for all of them, and it will be necessary only to compare the modified effectiveness measures. Since we are mainly interested in comparing sequences that are applied to the same initial set of documents and yield identical sets as their results, we shall find it convenient to define a sequence T_i to be coextensive with a sequence

T_j, when T_i and T_j are both coinitial and coterminal. We can define an effectiveness measure for coextensive sequences of search tool and an optimal sequence can then be defined.

The concept of coextensiveness enables us to compare search-tool sequences that take, so to speak, different paths through a set of documents to reach the same final set. But there are three cases to be considered in comparing sequences that are coinitial but not coterminal. Case I is that in which we can add a suitable additional search tool to yield a new sequence. Then T_i and T_j are coterminal, and we can compare their effectiveness. Case II is that in which sequences T_i and T_j yield different resultant sets of documents such that neither of the resultant sets contains the other but both contain the same group of actually pertinent documents. Thus we can again add search tools to form new sequences such that T_i' and T_j' are coterminal, and hence coextensive. Case III is that in which again sequences T_i and T_j yield different resultant sets of documents; however, one contains some actually pertinent documents that are not contained in the other. How to handle this case remains an open question.

In a typical manner of using a conventional library to seek books that might be pertinent to one's interest of the moment, the set of documents under consideration is successively restricted from the class of all literature down to a few pages. If one is searching for pertinent articles in journals, the process still has several successively more restrictive stages. Now let us speculate about a document-retrieval system that includes an effective semi-automated searching system. What form should the search process take in such a system? Complete scanning may be possible for small collections of documents but is hardly likely to be economically feasible for large ones; and the only substitute for complete scanning is the use of some form of condensed representations of the documents in the system. We should probably be well advised to seek guidance for semi-automated document-retrieval systems from existing systems. We observe that the search tools used at each successive stage above are successively longer, and successively more costly, for the length of a tool is an allowable measure of its cost. We observe also that, in general, at the same time as the successive search tools are growing longer, the number of documents eliminated by each successive search tool is successively smaller. Thus it appears that the costs of the successive tools are increasing while the corresponding restriction ratios are decreasing, with the result that successive tools would appear to be less effective. This conclusion, however, overlooks the effect of the loss function. In designing automated document-retrieval systems to follow the existent practice, the first search stage might consist in the restriction to a particular reel of magnetic tape; the second search stage might be the computer's scanning of sets of descriptors for the presence of certain descriptors in certain specified Boolean combinations; and the third stage might consist in the display of automatic abstracts.

One must not overlook the fact that in designing automated systems we shall not necessarily be restricted to analogs of the traditional library search tools. We are free to develop new tools that will take advantage of the speed of computers in performing routine tasks, and we can utilize the communication between the user and the system to make both the old and the new tools more effective. A possible search method might be to have the computer scan automatic indexes and compare the index terms therein with the request, then obtain the possibly pertinent documents and display their association maps for the user to examine, and finally after his examination and further rejection of documents, provide the user with automatic abstracts for the remaining documents. It might be possible to select by various methods a small set of documents possibly pertinent to a user's interest and then, as the final stage in the search process, to prepare what might be called "ad hoc" or "tailor-made" automatic abstracts that would reflect (a) the pertinence of each document to the request, and (b) the distinctions among the documents in the set. The concept of preparing "ad hoc" abstracts could be extended to other types of condensed

representations as well. We re-emphasize the absence of and need for an adequate theory of the search process.

We have been talking about the search process in a document-retrieval system and about tools for the search process. For theoretical purposes all the search tools that we have used as examples can be viewed as condensed representations of the contents of some document or set of documents. Recall the notion that all types of condensed representations serve two functions to a great or lesser degree: the search-tool function, and the content-revealing function. It is possible to list a variety of types of condensed representations in the order of their decreasing service to the search-tool function and their simultaneously increasing service to the content-revealing function.

We can distinguish two basic kinds of human-produced types of condensed representations: on the one hand, representations during whose formation a human other than the author exercises some degree of critical judgment about the content of the document; on the other hand, representations in whose formation such judgment is not exercised. For a long time to come, document-content analysis by automatic means will not be able to simulate or in any way replace those condensed representation types that fall into the critical-judgment category. All available and prospective techniques for automatic document-content analysis produce only representations in the non-critical-judgment category. Lists of the critical-judgment and non-critical-judgment categories of condensed representations are essentially in the previously suggested order of decreasing service to the search tool function and increasing service to the content revealing function. Also, the condensed-representation types in the two lists are listed in order of increasing cost, if we interpret our cost function $C(\cdot)$ in terms of the length of the search tool.

Individual examples of the different types of condensed representations of documents will vary in the degree to which they serve the search tool and the contnet-revealing functions. Criteria for representations that are to be used primarily as search tools include accuracy, discrimination, interpretability, and brevity. Unfortunately, it is a fact of nature that these criteria are impossible to satisfy simultaneously. A variety of search tools is needed, each designed for a specific stage of the search process, in order to emphasize the criterion most important at that stage. The number of search tools employed in traditional document-retrieval systems is a reflection of this need; but the advent of semi-automated document-retrieval systems will make possible a hitherto unknown flexibility and variety in search tools.

The section designation in abstract numbers may serve to categorize the entries under a subject. Thus, costs: in I are costs of systems being compared, in II are costs of one system being described, in III are discussions of cost criteria, and in IV are proposed methods to measure and record costs.

BIBLIOGRAPHY ON EVALUATION OF INFORMATION SYSTEMS

INTRODUCTION

The citations listed in this bibliography include references to the literature on evaluation of information systems through 1966. Selected items have been annotated and appear in the main body of this publication; others were reviewed but not annotated, because they were deemed to be outside the scope of the annotated bibliography; still others were noted but not reviewed, mainly due to pressures of time. Those items that have been annotated are identified with the designation of the corresponding abstract.

Several valuable sources were consulted in preliminary collection of material for the annotated bibliography. In particular, the bibliography by Pronko in his report, Comparative Studies of Retrieval Systems: Problems and Prospects; the bibliography by Newman in the National Academy of Sciences-National Research Council report, The Metallurgical Searching Service of the American Society for Metals-Western Reserve University: An Evaluation; and that by Taulbee, Bibliography on the Evaluation of Information Storage and Retrieval Systems, all were of great help. In addition, C. P. Bourne's Bibliography on the Performance Evaluation of Information Systems (unpublished, prepared January 30, 1964) was a helpful source of references. As work on the bibliography progressed, personal reading and communication with co-workers added to the list of references. This bibliography is the result of such activities.

[Numbers in brackets refer to abstracts]

Abraham, C. T., Techniques for Thesaurus Organization and Evaluation, in Parameters of Information Science (Proceedings of the American Documentation Institute Annual Meeting, Philadelphia, Pa., October 5-8, 1964), Washington, D.C., Spartan Books, 1964, p. 485-497.

Ackoff, Russell L., Measurement of Value of Recorded Information, Report to the National Science Foundation, Cleveland, Ohio, Case Institute of Technology, July 12, 1961, 61 p., AD 260 734.

Adams, Scott, MEDLARS: Performance, Problems, Possibilities, Bulletin of the Medical Library Association 53, 139-151 (April 1965).

Adams, W. Mansfield, Relationship of Keywords in Titles to References Cited, American Documentation 18, 26-32 (January 1967).

Aitchison, Jean and Cyril W. Cleverdon, A Report on a Test of the Index of Metallurgical Literature of Western Reserve University, Report to the National Science Foundation on the ASLIB-Cranfield Research Project, Cranfield, England, The College of Aeronautics, October 1963, 270 p., AD 419 956. [I-1]

Altmann, Berthold, A Multiple Testing of the ABC Method and the Development of a Second-Generation Model. Part I. Preliminary Discussions of Methodology, Report No. TR-1295, Washington, D.C., U.S. Army Materiel Command, Harry Diamond Laboratories, April 1965, 86 p., AD 617 118. [II-17]

Altmann, Berthold, A Multiple Testing of the ABC Method and the Development of a Second-Generation Model. Part II, Test Results and an Analysis of "Recall Ratio", Report No. TR-1296, Washington, D.C., U.S. Army Materiel Command, Harry Diamond Laboratories, October 1965, 35 p., AD 625 924.

Altmann, Berthold, A Multiple Testing of the Natural Language Storage and Retrieval ABC Method: Preliminary Analysis of Test Results, American Documentation 18, 33-45 (January 1967).

American Psychological Association, Reports of the APA's Project on Scientific Information Exchange in Psychology, Vol. I, published under NSF Grant G-18494, December 1963, 283 p., PB 164 496; Vol. II, published under NSF Grant G-[illegible]1 (continuation of G-18494), December 1965, 292 p., PB 169 005.

Anderson, Alan A., Guidelines for the Utilization of Statisticians in the Design and Execution of Information Retrieval System Evaluation Studies, Santa Monica, Calif., System Development Corp., July 12, 1966, 11 p.

Anderson, Alan A., Multi-Level File Structure as a Frame of Reference for Measuring User Interest, Report No. SDC-SP-2137, Santa Monica, Calif., System Development Corp., July 12, 1966, 12 p., AD 636 832.

ARINC Research Corporation, System Effectiveness Concepts and Analytical Techniques, ARINC Pub. No. 267-01-7-419, Washington, D.C., ARINC Research Corp., January 1964.

Artandi, Susan, Measure of Indexing, Library Resources and Technical Services 8, 229-235 (Summer 1964). [II-19]

Artandi, Susan, Thesaurus Controls Automatic Book Indexing by Computer, in Automation and Scientific Communication. Short Papers, Part 1 (papers contributed to the 26th Annual Meeting of American Documentation Institute, Chicago, Ill., October 1963), Ed. by H. P. Luhn, Washington, D.C., American Documentation Institute, 1963, p. 1-2. (A condensed version of Book Indexing by Computer, Ph. D. Thesis, New Brunswick, N.J., Rutgers, The State University Press, 1963, 207 p.) [II-18]

Artandi, Susan, Investigation of Systems for the Intelligent Organization of Information, Report to the National Science Foundation from the Graduate School of Library Service, Rutgers, The State University, New Brunswick, N.J., June 1964, 40 p., PB 166 202. [I-11]

Arthur Andersen & Co., Research Study of Criteria and Procedures for Evaluating Scientific Information Retrieval Systems, Final Report to the National Science Foundation, Contract NSF-C218, New York, March 1962, 103 p. plus appendices, AD 273 115. [IV-2]

Arthur D. Little, Inc., A Model for Study and Evaluation of Coordinate Retrieval Systems, in Centralization and Documentation, Report No. C-64469, Final Report to the National Science Foundation, Cambridge, Mass., Arthur D. Little, Inc., Second Edition, June 1964, p. 19-31. (First Edition, July 1963, 70 p., PB 181 548). [IV-3]

Arthur D. Little, Inc., Models of Performance of Coordinate Search Systems, Appendix II, in Centralization and Documentation. Appendices to a Final Report to the National Science Foundation, Cambridge, Mass., Arthur D. Little, Inc., July 1963, p. 28-69, P 181 548 A. [IV-4]

Arthur D. Little, Inc., Evaluation of Performance of Large Information Retrieval Systems, Appendix III, in Centralization and Documentation. Appendices to a Final Report to the National Science Foundation, Cambridge, Mass., Arthur D. Little, Inc., July 1963, p. 70-109, PB 181 548 A. [IV-5]

Arthur D. Little, Inc., Quantitative Relations Useful for the Evaluation of Information Retrieval Systems, Appendix to Centralization and Documentation, Report No. C-64469, Final Report to the National Science Foundation, Cambridge, Mass., Arthur D. Little, Inc., Second Edition, June 1964, 18 p. [IV-6]

Atherton, Pauline A., A Proposed Standard Description for Evaluation Tests of Retrieval Systems, New York, American Institute of Physics, Documentation Research Project, 1965, 6 p.

Atherton, Pauline and Harold Borko, A Test of a Factor-Analytically Derived Automated Classification Method Applied to Descriptions of Work and Search Requests of Nuclear Physicists, Report No. AIP/DRP 65-1, New York, American Institute of Physics, Documentation Research Project; Report No. SP-1905, Santa Monica, Calif., System Development Corporation, January 1965, 15 p.

Atherton, Pauline A. and J. C. Yovich, Three Experiments with Citation Indexing and Bibliographic Coupling of Physics Literature, New York, American Institute of Physics, Documentation Research Project, April 1962, 39 p. [I-12]

Auerbach Corporation, DOD User Needs Study. Phase I, Final Technical Report No. 1151-TR-3, in 2 Vol., Philadelphia, Pa., Auerbach Corp., May 14, 1965, 484 p., Vol. 1, AD 615 501; Vol. 2, AD 615 502.

Bacon, F.R. et al, Application of a Telereference System to Divisional Library Card Catalogs: A Feasibility Analysis, Final Report to the Council on Library Resources, Inc., Ann Arbor, Mich., University of Michigan, Engineering Research Institute, May 1958, 91 p. [II-20]

Badger, George Jr. and William Goffman, An Experiment with File Ordering for Information Retrieval, in Parameters of Information Science (Proceedings of the Americal Documentation Institute Annual Meeting, Philadelphia, Pa., October 5-8, 1964), Washington, D.C., Spartan Books, p. 379-381. [IV-7]

Baer, H., Organisation, Kosten, and Leistungen der Dokumentation, Industriellen Organisation, Heft 1/2, (1959).

Balz, Charles F., The Need for a Thesaurus in Automated Information Retrieval, Report No. 62-825-481, Owego, N.Y., IBM Corp., Space Guidance Center, September 1962, 10 p.

Barhydt, Gordon C., A Comparison of Relevance Assessment by Three Types of Evaluator, in Parameters of Information Science (Proceedings of the American Documentation Institute Annual Meeting, Philadelphia, Pa., October 5-8, 1964), Washington, D.C., Spartan Books, 1964, p. 383-385. [II-21]

Bartels, W., Vergleiche der Wirtschaftlichkeit von Sichloch -, Kerbloch - und Steilkarteien (Comparative Economics of Punched -, Edge-Notched -, and Optical Coincidence-Card Files), in Kurzberichte und Sonderdrucke zur 8. Öffenthichen Arbeitrsitzung des Arbeitsausschusces für Kostengrundlagen der Dokumentation der Deutschen Gesellschaft für Dokumentation e. V. am Oktober 9, 1961, in Bad Dürkheim an der Weinstrasse, (in Short Notes and Reprints of the Eighth Public Workshop of the Committee for Cost Bases of Documentation), Frankfurt am Main, 1961, p. 12-16.

Bartels, W., Sichtlochkarten - Kerhlochkarten, Vorteile und Grenzen heider Methoden hei der Karteibefragung, dargestellt am Beispiel einer Pflanzenschutz - Literaturkartei (Punched Cards - Notched Cards - Advantage and Limits of Both Methods in Searching Card Files, with an Example from Plant Care Literature), Nachrichten für Dokumentation 12, 137-146 (September 1961).

Batten, William E., Information Retrieval - The Economic Aspect, Journal of Documentation 22, 87-92 (June 1966).

Becker, Joseph, Getting to Know the User of an IR System, in Information Systems Workshop: The Designer's Responsibility and His Methodology, Washington, D.C., Spartan Books, 1962, p. 61-67.

Becker, Joseph, System Analysis, Prelude to Library Data Processing, ALA Bulletin 59, 293-296 (April 1965).

Belzer, Jack and William Goffman, Theoretical Considerations in Information Retrieval Systems, Communications of the ACM 7, 439-442 (July 1964). [IV-8]

Bernard, Jessie and Charles Shilling, Accuracy of Titles in Describing Content of Biological Science Articles, BSCP Communique 10-63, Washington, D.C., American Institute of Biological Sciences, Biological Sciences Communication Project, May 1963, 49 p. plus appendices.

Bernier, Charles L., Correlative Indexes. X. Subject-Index Qualities, Journal of Chemical Documentation 4, 104-107 (April 1964).

Bernier, Charles L., Indexing Process Evaluation, American Documentation 16, 323-328 (October 1965).

Binford, Richard Lee, A Comparison of Keyword-in-Context (KWIC) Indexing to Manual Indexing, M.S. Thesis, University of Pittsburgh, 1965, 77 p., AD 620 420.

Black, Donald V., Indexing Techniques, Description and Background, Appendix to Documentation Storage and Retrieval Techniques, Report No. PRC D-634A, Los Angeles, Calif., Planning Research Corp., June 13, 1963, 29 p., AD 414 713.

Bloomfield, Masse, Evaluation of Coordinate Indexing at the Naval Ordnance Test Station, American Documentation 8, 22-25 (January 1957).

Bloomfield, Masse, Simulated Machine Indexing, Part 4: A Technique to Evaluate the Efficiency of Indexing, Special Libraries 57, 400-403 (July/August 1966).

Blunt, Charles R., An Information Retrieval System Model, Report No. 352. 14-R-1, State College, Pa., HRB-Singer, Inc., October 1965, 144 p., AD 623 590.

Blunt, Charles R. et al, A General Model for Simulating Storage and Retrieval Systems, Report No. 352. 14-R-2, State College, Pa., HRB-Singer, Inc., April 1966, 183 p.

Bohnert, Lea M., Two Methods of Organizing Technical Information for Search, American Documentation 6, 134-151 (October 1955).

Booth, Andrew D., Characterizing Documents - a Trial of an Automatic Method, Computers and Automation 4, 32-33 (November 1965).

Borko, Harold, A Research Plan for Evaluating the Effectiveness of Various Indexing Systems, Field Note FN 5649/000/01, Santa Monica, Calif., System Development Corporation, July 10, 1961, 23 p., AD 278 624. [IV-9]

Borko, Harold, Evaluating the Effectiveness of Information Retrieval Systems, Report No. SP-909/000/00, Santa Monica, Calif., System Development Corporation, August 2, 1962, 7 p. [III-1]

Borko, Harold, Determining User Requirements for an Information Storage and Retrieval System: A Systems Approach, in Information Systems Workshop: The Designer's Responsibility and His Methodology, Washington, D.C., Spartan Books, 1962, p. 37-46.

Borko, Harold, The Construction of an Empirically Based Mathematically Derived Classification System, in AFIPS Conference Proceedings, Volume 21, 1962 Spring Joint Computer Conference, Palo Alto, Calif., National Press, 1962, p. 279-289. (Also available as Report No. SP-585, Santa Monica, Calif., System Devel. Corp., October 26, 1961, 23 p.)

Borko, Harold, Studies on the Reliability and Validity of Factor-Analytically Derived Classification Categories, in Statistical Association Methods for Mechanized Documentation (Symposium Proceedings), NBS Misc. Publ. 269, Washington, D.C., National Bureau of Standards, December 15, 1965, p. 245-257.

Borko, Harold, Measuring the Reliability of Subject Classification by Men and Machines, American Documentation 15, 268-273 (October 1964).

Borko, Harold and Myrna Bernick, Automatic Document Classification, Journal of the ACM 10, 151-162 (April 1963). (Also available as TM-771, Santa Monica, Calif., System Development Corp., November 15, 1962, 19 p.)

Borko, Harold and Myrna Bernick, Automatic Document Classification. Part II. Additional Experiments, Journal of the ACM 11, 138-151 (April 1964). (Also available as TM-771/001/00, Santa Monica, Calif., System Development Corp., October 18, 1963, 33 p.)

Bornstein, Harry, A Paradigm for a Retrieval Effectiveness Experiment, American Documentation 12, 254-259 (October 1961). [IV-10]

Bourne, Charles P., A Review of the Methodology of Information System Design, in Information Systems Workshop: The Designer's Responsibility and His Methodology, Washington, D.C., Spartan Books, 1962, p. 11-35. [III-2]

Bourne, Charles P., Evaluation of Indexing Systems, in American Documentation Institute Annual Review of Information Science and Technology, Vol. 1, Carlos A. Cuadra, Editor, New York, Interscience Publishers, 1966, p. 171-190.

Bourne, Charles P. and Donald F. Ford, Cost Analysis and Simulation Procedures for the Evaluation of Large Information Systems, American Documentation 15, 142-149 (April 1964). [IV-12]

Bourne, Charles P. et al., Requirements, Criteria, and Measures of Performance of Information Storage and Retrieval Systems, Final Report to the National Science Foundation on SRI Project 3741, Menlo Park, Calif., Stanford Research Institute, December 1961, 132 p., AD 270 942. [IV-11]

Boyd, D. F. and H. S. Krasnow, Economic Evaluation of Management Information Systems, IBM Systems Journal 2, 2-23 (March 1963).

Brandenberg, W. et al., Selective Dissemination of Information. SDI 2 System, Report No. 17-031, Yorktown Heights, N.Y., IBM Corporation, Advanced Systems Development Division, April 18, 1961, 95 p. [II-22]

Brownson, Helen L., Evaluation of Document Searching Systems and Procedures, Journal of Documentation 21, 261-266 (December 1965).

Bryant, Edward C., Some Notes on Associative Retrieval, January 28, 1964, 6 p. Further Notes on Associative Retrieval, February 28, 1964, 5 p. Denver, Colo., Westat Research Analysts, Inc., PB 166 511.

Bryant, Edward C., A Stochastic Model for the Patent Examining Process, in Proceedings of the Social Statistics Section, American Statistical Association Meeting, Cleveland, Ohio, 1963, p. 91-102.

Bryant, Edward C., On Stochastic Models of the Patent Office Examining System, Report No. WRA PO 8, Denver, Colo., Westat Research Analysts, Inc., April 1963, 31 p., PB 166 486.

Bryant, Edward C., A Progress Report on Search-System Evaluation in the United States Patent Office, Report No. WRA PO 14, Denver, Colo., Westat Research Analysts, Inc., May 1964, 27 p. [II-23]

Bryant, Edward C., Progress Toward Evaluation of Information Retrieval Systems, in Information Retrieval Among Examining Patent Offices (Proceedings of Fourth Annual Meeting of the Committee for International Cooperation in Information Retrieval Among Examining Patent Offices, Washington, D.C., October 1964), Ed. by H. Pfeffer, Washington, D.C., Spartan Books, 1966, p. 362-377. [III-4]

Bryant, Edward C., Evaluation of Information Retrieval Systems in Patent Office Environments. Part I. Statistical Concepts, Bethesda, Md., Westat Research Analysts, Inc., February 1965, 37 p., PB 168 000. [III-3]

Bryant, Edward C., Measurement, Modeling, Experimentation, and Evaluation in Information Retrieval Systems, in Information Retrieval Among Examining Patent Offices (Proceedings of Third Annual Meeting of the Committee for International Cooperation in Information Retrieval Among Examining Patent Offices, Vienna, September 1963), Baltimore, Md., Spartan Books, Inc., 1964, p. 198 - 207. [III-3]

Bryant, Edward C., Control of Indexing Errors, Report No. WRA PO 16, Bethesda, Md., Westat Research Analysts, Inc., April 1965, 15 p., PB 168 336.

Bryant, Edward C., A Status Report on Research in Information Retrieval, in Management Information Systems and the Information Specialist (Proceedings of a Symposium held at Purdue University, July 12-13, 1965), Ed. by John M. Houkes, Lafayette, Ind., Purdue University, 1966, p. 87-95.

Bryant, Edward C., Notes on Sampling for Missed Documents in IR Systems, Bethesda, Md., Westat Research Analysts, Inc., September 1964, 3 p.

Bryant, Edward C. and Donald W. King, Experiments with Search Systems of the U.S. Patent Office (Paper presented to American Association for the Advancement of Science, Berkeley, Calif., December 29, 1965), Bethesda, Md., Westat Research Analysts, Inc., 11 p.

Bryant, Edward C. and Donald W. King, Potential Improvement of Retrieval by Associative Adjustment of the File (Paper prepared for presentation at Sixth Annual Meeting of the Committee for International Cooperation in Information Retrieval Among Examining Patent Offices, The Hague, October 1966), Bethesda, Md., Westat Research, Inc., 11 p.

Bryant, Edward C., Donald W. King and P. James Terragno, Analysis of an Indexing and Retrieval Experiment for the Organometallic File of the U.S. Patent Office, Report No. WRA PO 10, Denver, Colo., Westat Research Analysts, Inc., August 1963, 52 p. plus appendices, PB 166 488. [II-24]

Bryant, Edward C., Donald W. King and P. James Terragno, Some Technical Notes on Coding Errors, Report No. WRA PO 7, Denver, Colo., Westat Research Analysts, Inc., July 1963, 30 p., PB 166 487. [II-25]

Bryant, Edward C., Donald W. King and P. James Terragno, Revised Design for Coding Experiment, 307/88.5 File, Report No. WRA PO 9, Denver, Colo., Westat Research Analysts, Inc., April 1965, 15 p., PB 168 488.

Bryant, Edward C., D. T. Searls and R. H. Shumway, Some Theoretical Aspects of the Improvement of Document Screening by Associative Transformations, Contract AF 49(638)1484, Bethesda, Md., Westat Research Analysts, Inc., November 30, 1965, 48 p., AD 628 191.

Buscher, William C., The Analysis of Medical Documents with a Comparative Evaluation of Three Indexing Procedures (Paper presented at the Rochester Conference on Data Acquisition and Processing in Biology and Medicine, July 1963), Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, 13 p. [I-13]

Carlson, Gary, Letter to the Editor, American Documentation 14, 328-329 (October 1963). [II-36]

Carnovsky, L., Evaluation of Library Services, UNESCO Bulletin for Libraries 13, 221-225 (October 1959).

Causey, Robert L., Some Mathematical Problems Arising in Information Retrieval from Inverted Files, Final Report, Huntsville, Ala., Alabama University, Huntsville Research Institute, June 3, 1965, 17 p., AD 619 955.

Chodrow, Mark M., David E. Sparks and David P. Waite, Information Service System Modeling -- Analytical Tools for Management Evaluation, Report No. R-1061-4, Wakefield, Mass., Information Dynamics Corp., December 1963, 1 v.

Cleverdon, Cyril W., An Investigation into the Comparative Efficiency of Information Retrieval Systems, UNESCO Bulletin for Libraries 12, 267-270 (November-December 1958). [I-2]

Cleverdon, Cyril W., The Evaluation of Systems Used in Information Retrieval, in Proceedings of the International Conference on Scientific Information (Washington, D.C., November 1958), Vol. I, Washington, D.C., National Academy of Sciences-National Research Council, 1959, p. 687-698. [I-2]

Cleverdon, Cyril W., Report on the First Stage of an Investigation into the Comparative Efficiency of Indexing Systems, Report to the National Science Foundation on the ASLIB-Cranfield Research Project, Cranfield, England, The College of Aeronautics, September 1960, 166 p. [I-2]

Cleverdon, Cyril W., Interim Report on the Test Programme of an Investigation into the Comparative Efficiency of Indexing Systems, Report to the National Science Foundation on the ASLIB-Cranfield Research Project, Cranfield, England, The College of Aeronautics, November 1960, 80 p., plus figures. [I-2]

Cleverdon, Cyril W., The ASLIB-Cranfield Research Project on the Comparative Efficiency of Indexing Systems, ASLIB Proceedings 12, 421-431 (December 1960). [I-2]

Cleverdon, Cyril W., Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems, Report to the National Science Foundation on the ASLIB-Cranfield Research Project, Cranfield, England, The College of Aeronautics, October 1962, 305 p. plus figures, PB 162 342. [I-2]

Cleverdon, Cyril W., The Cleverdon-WRU Experiment: Conclusions, in Information Retrieval in Action (papers presented at 1962 Conference at the Center for Documentation and Communication Research), Cleveland, Ohio, The Press of Western Reserve University, 1963, p. 101-107. [I-3]

Cleverdon, Cyril W., The In-House Testing and Evaluation of the Operating Efficiency of the Intellectual Stages of Information Retrieval Systems, Cranfield, England, The College of Aeronautics, September 1964, 13 p. [III-5]

Cleverdon, Cyril W., The Cranfield Hypothesis, The Library Quarterly 35, 121-125 (April 1965).

Cleverdon, Cyril W. and Jack Mills, The Testing of Index Language Devices, ASLIB Proceedings 15, 106-130 (April 1963). [II-1]

Cleverdon, Cyril W. and Jack Mills, The Analysis of Index Language Devices, in Automation and Scientific Communication. Proceedings, Part 3 (papers presented at the 26th Annual Meeting of the American Documentation Institute, Chicago, Ill., October 1963), Ed. by P. C. Janaske, Washington, D.C., American Documentation Institute, 1964, p. 451-454. [II-1]

Cleverdon, Cyril W. and R. G. Thorne, A Brief Experiment with the Uniterm System of Coordinate Indexing for the Cataloging of Structural Data, Library Memorandum No. 7, Farnsborough, England, Royal Aircraft Establishment, January 1954, 20 p., AD 35 004.

Cleverdon, Cyril W., F. Wilfred Lancaster and Jack Mills, Uncovering Some Facts of Life in Information Retrieval, Special Libraries 55, 89-91 (February 1964).

Cleverdon, Cyril W., Jack Mills and Michael Keen, Factors Determining the Performance of Indexing Systems. Vol. 1, Design (Part 1. Text, Part 2. Appendices), Report to the National Science Foundation on the ASLIB-Cranfield Research Project, Cranfield, England, The College of Aeronautics, 1966, 377 p.

Cleverdon, Cyril W. and Michael Keen, Factors Determining the Performance of Indexing Systems. Vol. 2, Test Results, Report to the National Science Foundation on the ASLIB-Cranfield Research Project, Cranfield, England, The College of Aeronautics, 1966, 299 p.

Cohen, Stanley M., Carol M. Lauer and Bettina C. Schwartz, An Evaluation of Links and Roles as Retrieval Tools, Journal of Chemical Documentation 5, 118-121 (May 1965).

Conger, C. Richard, The Simulation and Evaluation of Information Retrieval Systems, State College, Pa., HRB-Singer, Inc., April 1965, 99 p.

Cordero, J. A., Letter to the Editor, American Documentation 14, 272 (July 1963).

Costello, John C. Jr., Uniterm Indexing Principles, Problems and Solutions, American Documentation 12, 20-26 (January 1961).

Costello, John C. Jr., Computer Requirements for Inverted Coordinate Indexes, American Documentation 13, 414-419 (October 1962).

Costello, John C. Jr., A Basic Theory of Roles as Syntactical Control Devices in Coordinate Indexes, Journal of Chemical Documentation 4, 116-124 (April 1964).

Cuadra, Carlos A., On the Utility of the Relevance Concept, Report No. SP-1595, Santa Monica, Calif., System Development Corporation, March 18, 1964, 9 p. [III-6]

Cuadra, Carlos A., Emory H. Holmes, Robert V. Katter and Everett M. Wallace, Experimental Studies of Relevance Judgments: Second Progress Report, SDC Report No. TM-3068/000/00, Santa Monica, Calif., System Development Corp., July 1, 1966, 53 p.

Cuadra, Carlos A., Emory H. Holmes, Robert V. Katter and Everett M. Wallace, Experimental Studies of Relevance Judgments: Third Progress Report, SDC Report No. TM-3347, Santa Monica, Calif., System Development Corp., January 20, 1967, 68 p.

Curtice, Robert M., Letters to the Editor, American Documentation 16, 248 (July 1965).

Curtice, Robert M. and Victor Rosenberg, Optimizing Retrieval Results with Man-Machine Interaction, in Abstracts (of papers submitted to International Federation for Documentation 1965 Congress, Washington, D.C., October 10-15, 1965), Washington, D.C., Secretariat, 1965 FID Congress, p. 64.

Curtice, Robert M. and Victor Rosenberg, Optimizing Retrieval Results with Man-Machine Interaction, Bethlehem, Pa., Lehigh University, Center for the Information Sciences, February 1965, 24 p.

Dale, Alfred G., Retrieval System Experimentation and Evaluation at LRC, Austin, Texas, University of Texas, Linguistics Research Center, July 1965, 9 p., PB 168 284.

Dale, Alfred G. and Nell Dale, Clumping Techniques and Associative Retrieval, in Statistical Association Methods for Mechanized Documentation (Symposium Proceedings), NBS Misc. Publ. 269, Washington, D.C., National Bureau of Standards, December 15, 1965, p. 230-235.

Dale, Alfred G. and Nell Dale, Some Clumping Experiments for Associative Document Retrieval, American Documentation 16, 5-9 (January 1965).

Damerau, Fred J., An Experiment in Automatic Indexing, American Documentation 16, 283-289 (October 1965).

Davis, Ruth M., Classification and Evaluation of Information System Design Techniques, in Second Congress on the Information System Sciences (Hot Springs, Va., November 1964), Washington, D.C., Spartan Books, Inc., 1965, p. 77-83.

Dekker, Jacob, Evaluation of the Mechanical Search System for Analog-Digital Converters of the Netherlands Patent Office, in Information Retrieval Among Examining Patent Offices (Proceedings of Fourth Annual Meeting of the Committee for International Cooperation in Information Retrieval Among Examining Patent Offices, Washington, D.C., October 1964), Ed. by Harold Pfeffer, Washington, D.C., Spartan Books, 1965, p. 111-178.

DeLucia, Al, Index-Abstract Evaluation and Design, American Documentation 15, 121-125 (1964).

Dennis, Bernard K., J. J. Brady and J. A. Dovel, Jr., Five Operational Years of Inverted Index Manipulation and Abstract Retrieval by an Electronic Computer, Journal of Chemical Documentation 2, 234-242 (October 1962).

Dougherty, Richard M., The Scope and Operating Efficiency of Information Centers as Illustrated by the Chemical-Biological Coordination Center of the National Research Council, College and Research Libraries 25, 7-12, 20 (January 1964). (Also available as Ph. D. Dissertation, Rutgers, The State University, 1963, 242 p.)

Doyle, Lauren B., Is Relevance an Adequate Criterion for Retrieval System Evaluation?, in Automation and Scientific Communication. Short Papers, Part 2 (papers contributed to the 26th Annual Meeting of the American Documentation Institute, Chicago, Ill., October 1963), Ed. by H. P. Luhn, Washington, D.C., American Documentation Institute, 1963, p. 199-200. (Also available as SDC Report No. SP-1262, Santa Monica, Calif., System Development Corp., July 1, 1963, 6 p.) [III-7]

Doyle, Lauren B., Is Automatic Classification a Reasonable Application of Statistical Analysis of Text?, Journal of the ACM 12, 473-489 (October 1965). (Also available as SDC Report No. SP-1753, Santa Monica, Calif., System Development Corp., August 31, 1964, 34 p.) [II-26]

Doyle, Lauren B., Breaking the Cost Barrier in Automatic Classification, SDC Report No. SP-2516, Santa Monica, Calif., System Development Corp., July 1, 1966, 62 p., AD 636 837.

Dyke, H. Gordon, A Figure-of-Merit Ordering System for a Search Output, American Documentation 10, 85-86 (January 1959).

Dzubak, B. J. and C. R. Warburton, The Organization of Structured Files, Communications of the ACM 8, 446-452 (July 1965).

Edmundson, Harold P., Problems in Automatic Abstracting, Communications of the ACM 7, 259-263 (April 1964). [II-27]

Einhorn, Sheldon, Effectiveness Analysis for Large-Scale Complex Systems, in Information Processing 1965, Vol. 2, (Proceedings of IFIP Congress 1965), Ed. by Wayne A. Kalenich, Washington, D.C., Spartan Books, 1966.

Ernst, Heinrich A., Design and Evaluation of a Literature Retrieval Scheme, in Quarterly Progress Report No. 55, Cambridge, Mass., Massachusetts Institute of Technology, Research Laboratory of Electronics, October 15, 1959, p. 130-131, AD 230 230. [IV-13]

Ernst, Martin L., Evaluation of Performance of Large Information Retrieval Systems, in Second Congress on the Information System Sciences (Hot Springs, Va., November 1964), Washington, D.C., Spartan Books, Inc., 1965, p. 239-249.

Fairthorne, Robert A., Criteria for the Organs of Information Systems, in Information Systems Workshop: The Designer's Responsibility and His Methodology, Washington, D.C., Spartan Books, 1962, p. 113-119.

Fairthorne, Robert A., Basic Parameters of Retrieval Tests, in Parameters of Information Science (Proceedings of American Documentation Institute Annual Meeting, Philadelphia, Pa., October 5-8, 1964), Washington, D.C., Spartan Books, 1964, p. 343-345[IV-14]

Fairthorne, Robert A., Some Basic Comments on Retrieval Testing, Journal of Documentation 21, 267-270 (December 1965).

Fairthorne, Robert A., Implications of Test Procedures, in Information Retrieval in Action (papers presented at the 1962 Conference at Center for Documentation and Communications Research), Cleveland, Ohio, The Press of Western Reserve University, 1963, p. 109-113. [III-8]

Farris, Rhodes N., Computers Cut the Cost of Literature Searches, Chemical Engineering Progress 62, 89-91 (May 1966).

Fels, E. M., Evaluation of an Information Retrieval System by Modified Mooers Plan, American Documentation 14, 28-34 (January 1963). [II-28]

Fossum, Earl G. and Gilbert Kaskey, Some Notes on the Use and Data Processing Aspects of Association Factors in IS and R Systems, Report No. AFOSR 65-1702, Blue Bell, Pa., UNIVAC Division of Sperry Rand Corp., March 30, 1965, 30 p., AD 624 241.

Fossum, Earl G., et al, Optimization and Standardization of Information Retrieval Language and Systems, Report No. AFOSR-3216, Blue Bell, Pa., UNIVAC Division of Sperry Rand Corp., July 1962, 117 p., AD 287 117.

Freeman, Robert R. and G. Malcom Dyson, Development and Production of Chemical Titles, Journal of Chemical Documentation 3, 16-20 (January 1963).

Gagliardi, Ugo O., Data File Size and Its Relation to the Bayesian Effectiveness of an Information Retrieval System, Report No. ESD-TR-65-275, Bedford, Mass., Air Force Systems Command, Electronic Systems Division, April 1965, 80 p., AD 618 311. [IV-15]

Garfield, Eugene and I. H. Sher, New Factors in the Evaluation of Scientific Literature through Citation Indexing, American Documentation 14, 195-201 (July 1963).

Garland, J. L. and Kenneth W. Webb, Information Retrieval Systems Evaluation Technique, Termination Report, Task No. 0355, Rockville, Md., IBM Corp., Federal Systems Division, March 1962, 23 p. [IV-16]

Giuliano, Vincent E., Document Retrieval System Evaluation Principles, in Abstracts (of papers submitted to International Federation for Documentation 1965 Congress, Washington, D.C., October 10-15, 1965), Washington, D.C., Secretariat, 1965 FID Congress, p. 32.

Giuliano, Vincent E. and Paul E. Jones, Jr., Linear Associative Information Retrieval, in Vistas in Information Handling, Vol. 1, Ed. by P. W. Howerton and D. C. Weeks, Washington, D.C., Spartan Books, 1963, p. 30-54.

Giuliano, Vincent E. and Paul E. Jones, Jr., Study and Test of a Methodology for Laboratory Evaluation of Message Retrieval Systems, Report No. ESD-TR-66-405, Cambridge, Mass., Arthur D. Little, Inc., August 1966, 183 p., AD 642 829.

Glazer, Ezra, Increasing the Efficiency of the Use of Information -- A Background Review, in Abstracts (of papers submitted to International Federation for Documentation 1965 Congress, Washington, D.C., October 10-15, 1965), Washington, D.C., Secretariat, 1965 FID Congress, p. 37-38.

Gluss, Brian, A Record Storage Model and Its Information Retrieval Strategy, in Proceedings of the Second International Conference on Operational Research 1960, Ed. by J. Banbury and J. Maitland, New York, John Wiley and Sons, Inc., 1961, p. 82-88.

Goffman, William, On Relevance as a Measure, Information Storage and Retrieval 2, 201-203 (December 1964).

Goffman, William, A Searching Procedure for Information Retrieval, Information Storage and Retrieval 2, 73-78 (July 1964).

Goffman, William, On the Logic of Information Retrieval, Information Storage and Retrieval 2, 217-220 (August 1965).

Goffman, William and Vaun A. Newill, A Methodology for Test and Evaluation of Information Retrieval Systems, Information Storage and Retrieval 3, 19-25 (August 1966). (Also available as Report No. CSL:TR-2, Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, July 1964, 19 p., AD 614 005). [III-9]

Goldwyn, Alvin J., Purpose and Objectives of the Comparative Systems Laboratory, Report No. CLS:TR-1, Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, August 1964, 21 p. [III-11]

Goldwyn, Alvin J., The Place of Indexing in the Design of Information Systems Tests, in Automation and Scientific Communication. Short Papers, Part 2 (papers contributed to the 26th Annual Meeting of the American Documentation Institute, Chicago, Ill., October 1963), Ed. by H. P. Luhn, Washington, D.C., American Documentation Institute, 1963, p. 321-322. [III-10]

Goodman, Arnold F., John D. Hodges, Jr., and Forrest G. Allen, Final Report. DOD User-Needs Study, Phase II. Flow of Scientific and Technical Information Within the Defense Industry. Volume I. Overview, Report No. C6-2442/030, Anaheim, Calif., North American Aviation, Inc., Autonetics Division, November 30, 1966, 65 p., AD 647 111.

Greer, F. Loyal, The User Approach to Information Systems, Bethesda, Md., General Electric Co., Information Systems Operation, May 1963, 41 p. [IV-17]

Greer, F. Loyal, Word Usage and Implications for Storage and Retrieval, Washington, D.C., General Electric Co., Information Systems Operation, July 1962, 74 p. [IV-18]

Gull, C. Dake, Seven Years of Work on the Organization of Materials in the Special Library, American Documentation 7, 320-329 (October 1956). [I-14]

Gull, C. Dake, A Proposed Pilot Project, in Annual Report, 1956-57, Division of Engineering and Industrial Research, National Academy of Sciences-National Research Council, Washington, D.C., National Academy of Sciences-National Research Council, February 18, 1958, p. 54-61. [IV-19]

Gurk, Herbert M. and Jack Minker, The Design and Simulation of an Information Processing System, Journal of the ACM 8, 260-270 (April 1961).

Hagelback, E., Manual vs. Mechanical Search in Transistor and Non-Linear Conductor Systems, in Information Retrieval Among Examining Patent Offices (Proceedings of Fourth Annual Meeting of the Committee for International Cooperation in Information Retrieval Among Examining Patent Offices, Washington, D.C., October 1964), Ed. by Harold Pfeffer, Washington, D.C., Spartan Books, 1965, p. 518-540.

Hammond, William, Convertibility of Indexing Vocabularies, in Proceedings of Conference on The Literature of Nuclear Sciences: Its Management and Use, Oak Ridge, Tenn., U.S. Atomic Energy Commission, Division of Technical Information Extension, December 1962, p. 223-234. [I-15]

Hammond, William, Statistical Association Methods for Simultaneous Searching of Multiple Document Collections, in Statistical Association Methods for Mechanized Documentation (Symposium Proceedings), NBS Misc. Publ. 269, Washington, D.C., National Bureau of Standards, December 15, 1965, p. 237-243.

Hammond, William and Staffan Rosenborg, Experimental Study of Convertibility between Large Technical Indexing Vocabularies -- with Table of Indexing Equivalents, Technical Report No. IR-1, Contract NSF C-259, Silver Spring, Md., Datatrol Corp., August 1962, 297 p. [I-15]

Hammond, William, Staffan Rosenborg and Josephine Jaster, A Search Strategy for Retrieving Legal Information, Technical Report No. IR-2, Silver Spring, Md., Datatrol Corp., December 1, 1962, 19 p., PB 164 212.

Hammond, William et al, Common Vocabulary Approaches for Government Scientific and Technical Information Systems, Technical Report No. IR-10, Contract NSF C-342, Silver Spring, Md., Datatrol Corp., December 1963, 108 p., AD 430 000.

Hayes, Robert M., The Analysis of Information Systems, in Information Retrieval and Machine Translation, Ed. by Allen Kent, New York, Interscience Publishers, Inc., 1960, p. 429-443.

Hayes, Robert M., Mathematical Models for Information Systems Design and A Calculus of Operations, Report No. RADC-TR-61-96, Final Report to Rome Air Development Center, Torrance, Calif., Magnavox Research Lab., October 27, 1961, 178 p., AD 266 577.

Hayes, Robert M., Measurement of File Operating Effectiveness - Time, Cost, and Information, Part 6 of The Organization of Large Files, Final Report to the National Science Foundation, Sherman Oaks, Calif., Hughes Dynamics, Inc., Advance Information Systems Division, April 30, 1964, 50 p. [IV-20]

Hayes, Robert M., Theories of System Design, in Information Storage and Retrieval: Tools, Elements, Theories, by J. Becker and R. M. Hayes, New York, John Wiley and Sons, Inc., 1963, p. 398-425.

Hayes, Robert M., The Measurement of Information from a File, in Statistical Association Methods for Mechanized Documentation (Symposium Proceedings), NBS Misc. Publ. 269, Washington, D.C., National Bureau of Standards, December 15, 1965, p. 161-162.

Heckman, Ralph P., A Method for Investigating the Behavior of Attributes which Belong to Information Storage and Retrieval Systems, M.S. Thesis, Georgia Institute of Technology, August 1965, 98 p., AD 624 658.

Heilprin, Lawrence B. and S. S. Crutchfield, Project Lawsearch: A Statistical Comparison of Coordinate and Conventional Legal Indexing, in Parameters of Information Science (Proceedings of the American Documentation Institute Annual Meeting, Philadelphia, Pa., October 5-8, 1964), Washington, D.C., Spartan Books, 1964, p. 215-234. [I-16]

Herner, Saul, The Relationship of Information-Use Studies and the Design of Information Storage and Retrieval Systems, Report No. RADC TN 59-136, Washington, D.C., Herner and Co., December 1958, 25 p.

Herner, Saul and Mary Herner, Determining Requirements for Atomic Energy Information from Reference Questions, in Proceedings of the International Conference on Scientific Information (Washington, D.C., November 1958), Vol. I, Washington, D.C., National Academy of Sciences-National Research Council, 1959, p. 181-187.

Herner, Saul, F. Wilfred Lancaster and Walter F. Johanningsmeier, A Case Study in the Application of Cranfield System Evaluation Techniques, Journal of Chemical Documentation 5, 92-95 (1965). (Also available as Report No. AFOSR-64-2259, Report to U.S. Air Force Office of Scientific Research, Washington, D.C., Herner and Co., 14 p., AD 608 743.) [II-30]

Herner, Saul et al, Recommended Design for the U.S. Medical Library and Information System. Volume I - System Design, Implementation, Costs, Report to the National Science Foundation, Contract NSF-C442, Washington, D.C., Herner and Co., 1966.

Hersey, D.C. and Monroe E. Freeman, User Responses in the Evaluation of a Flexible Indexing and Retrieval System, in Automation and Scientific Communication. Short Papers, Part 2 (papers contributed to the 26th Annual Meeting of the American Documentation Institute, Chicago, Ill., October 1963), Ed. by H. P. Luhn, Washington, D.C., American Documentation Institute, 1963, p. 117-118.

Hewer, David J., An Evaluation of Methods for Improving Search Strategies in a Coordinate Searching System, M.S. Thesis, University of Pittsburgh, 1966, AD 481 441.

Hillier, James, Management's Evaluation of Information Services, in Information Retrieval Management, Ed. by L. Hattery and E. McCormick, Detroit, Mich., American Data Processing, Inc., 1963, p. 54-60.

Hillier, James, Measuring the Value of Information Services, Journal of Chemical Documentation 2, 31-33 (1962). [III-12]

Hillman, Donald J., Mathematical Theories of Relevance With Respect to Systems of Automatic and Manual Indexing, in Automation and Scientific Communication. Short Papers, Part 2 (papers contributed to the 26th Annual Meeting of the American Documentation Institute, Chicago, Ill., October 1963), Ed. by H. P. Luhn, Washington, D.C., American Documentation Institute, 1963, p. 323-324. [IV-21]

Hillman, Donald J., An Empirical Testing Program for Models of Information Storage and Retrieval Systems, Report No. AFOSR-64-2385, Final Report to the Air Force Office of Scientific Research, Bethlehem, Pa., Lehigh University, November 6, 1964, 10 p., AD 608 704.

Hillman, Donald J., Mathematical Theories of Relevance With Respect to the Problems of Indexing. Report No. 1: The Formal Basis of Relevance Judgments, Report to the National Science Foundation, NSF Grant GN-177, Bethlehem, Pa., Lehigh University, July 9, 1964, 21 p.

Hillman, Donald J., The Notion of Relevance (1), American Documentation 15, 26-34 (January 1964). [III-13]

Hirayama, Kenzo, Comparison of Keyword Indexing and Indexing by Systematic Classification, in Abstracts (of papers submitted to International Federation for Documentation 1965 Congress, Washington, D.C., October 10-15, 1965), Washington, D.C., Secretariat, 1965 FID Congress, p. 69.

Hoffman, J.M., Experimental Design for Measuring Intra- and Inter-Group Consistency of Human Judgment of Relevance, M.S. Thesis, Georgia Institute of Technology, August 1965, 111 p., AD 620 342.

Hoisman, A.J. and A.M. Daitch, Techniques for Relating Personnel Performance to System Effectiveness Criteria: A Critical Review of the Literature, Report to Bureau of Naval Personnel, Santa Monica, Calif., Dunlap and Associates, Inc., September 1964, 45 p. [III-14]

Holm, Bart E., Searching Strategies and Equipment, American Documentation 13, 31-39 (January 1962).

Hooper, R.S., Indexer Consistency Tests -- Origin, Measurements, Results and Utilization, Report No. TR-95-96, Bethesda, Md., IBM Corporation, Federal Systems Division, 1965, 19 p.

Houston, Nona and Eugene Wall, The Distribution of Term Usage in Manipulative Indexes, American Documentation 15, 105-114 (April 1964). [I-7]

Howerton, Paul W., Criteria for Total Information System Evaluation, in Information Handling: First Principles, Washington, D.C., Spartan Books and London, Cleaver-Hume Press, 1963, p. 195-207. [IV-22]

Hughes Dynamics, Methodologies for System Design, Report No. RADC-TDR-63-486, Final Report to U.S. Air Force, Rome Air Development Center, Los Angeles, Calif., Hughes Dynamics, February 24, 1964, 98 p., AD 434 749. [IV-23]

Hyslop, Marjorie R., A Compatibility Study of Two Information Systems, American Documentation 14, 292-298 (October 1963).

Hyslop, Marjorie R., Role Indicators and Their Use in Information Searching --- Relationship of ASM and EJC Systems, in Parameters of Information Science (Proceedings of American Documentation Institute Annual Meeting, Philadelphia, Pa., October 5-8, 1964), Washington, D.C., Spartan Books, 1964, p. 99-107. [II-3]

International Business Machines Corp., Implementation, Test and Evaluation of a Selective Dissemination System for NASA Scientific and Technical Information, Report No. NASA-CR-62020, Yorktown Heights, N.Y., IBM Corp., January 1966, 90 p.

Jackson, S.L., Sears and L.C. Subject Headings: A Sample Comparison, Library Journal 86, 755-756 (February 15, 1961).

Jacoby, J., Methodology for Indexer Reliability Tests, Report No. RADC-TN-62-1, Report to U.S. Air Force, Rome Air Development Center, Bethesda, Md., Documentation, Inc., March 1962.

Jacoby, J. and Vladimir Slamecka, Indexer Consistency Under Minimal Conditions, Report No. RADC-TDR-62-426, Report to U.S. Air Force, Rome Air Development Center, Bethesda, Md., Documentation, Inc., November 1962, 42 p. plus appendices, AD 288 087. [II-4]

Jahoda, Gerald, Correlative Indexing Systems for the Control of Research Records, DLS Dissertation, Columbia University School of Library Science, 1960.

Jahoda, Gerald, Development of a Combination Manual and Machine-Based Index to Research and Engineering Reports, Special Libraries 53, 74-78 (February 1962).

Jahoda, Gerald, A Technique for Determining Index Requirements, American Documentation 15, 82-85 (April 1964).

Janning, E. A., Establishment of a Coordinate Indexing Retrieval System for the Air Force Materials Laboratory, Report No. RTD-TDR-63-4263, Dayton, Ohio, University of Dayton Research Institute, November 1963, 71 p., AD 428 423.

Janning, E. A., The Modification of an Information Retrieval System by Improving Vocabulary Control, Indexing Consistency and Search Capabilities, Report No. AFML-TR-65-20, Dayton, Ohio, University of Dayton Research Institute, March 1965, 8 p. plus appendices, AD 613 301. [II-5]

Janning, E. A., Operations of a Document Retrieval System Using a Controlled Vocabulary, Report No. TR-66-36, Dayton, Ohio, University of Dayton Research Institute, March 1966, 20 p., AD 633 614.

Jaster, J. J., Barbara R. Murray and Mortimer Taube, Evaluating Coordinate Indexing Systems, in The State of the Art of Coordinate Indexing, Report to the National Science Foundation, Contract NSF-C-147, Washington, D. C., Documentation, Inc., February 1962, p. 81-107, AD 275 393. [II-29]

Johanningsmeier, Walter F. and F. Wilfred Lancaster, Project SHARP (SHips Analysis and Retrieval Project) Information Storage and Retrieval System: Evaluation of Indexing Procedures and Retrieval Effectiveness, Report No. NAVSHIPS 250-210-3, Washington, D. C., Department of the Navy, Bureau of Ships, June 1964, 49 p. [II-30]

Jonker, Frederick, The Descriptive Continuum: A "Generalized" Theory of Indexing, in Proceedings of the International Conference on Scientific Information (Washington, D. C., November 1958), Vol. II, Washington, D. C., National Academy of Sciences-National Research Council, 1958, p. 1291-1311. (Also available as Report No. AFOSR-TN-57-287, Report to the Air Force, Office of Scientific Research, Bethesda, Md., Documentation, Inc., June 1957, 26 p., AD 132 358.

Katter, Robert V., Language Structure and Interpersonal Commonality, Report No. SP-1185/000/01, Santa Monica, Calif., System Development Corporation, June 17, 1963, 30 p. [IV-24]

Kelley, J. Hilary, Entropy as a Measuring Tool for Information Systems, in Abstracts (of papers submitted to the International Federation for Documentation 1965 Congress, Washington, D. C., October 10-15, 1965), Washington, D. C., Secretariat, 1965 FID Congress, p. 70.

Kent, Allen, Comments in Discussion of Symposium on Advanced Methods in Information Storage and Retrieval, in Information Processing 1962, Ed. by Cicely M. Popplewell, Amsterdam, North-Holland Publishing Co., 1963, p. 296. [I-4]

Kent, Allen, The Cleverdon-WRU Experiment: Purpose, in Information Retrieval in Action (papers presented at 1962 Conference at the Center for Documentation and Communication Research), Cleveland, Ohio, The Press of Western Reserve University, 1963, p. 73-84.

Kent, Allen, Minimum Criteria for a Coordinated Information System, Letter to the Editor, American Documentation 11, 84-87 (January 1960). [III-15]

Kent, Allen, System Design Criteria, in Textbook on Mechanized Information Retrieval, by A. Kent, New York, John Wiley and Sons, Inc., 1962, p. 196-219.

Kent, Allen, Evaluating Procedures for Processing and Searching Metallurgical Literature, Communications of the ACM 3, 599, 623 (November 1960).

Kent, Allen and Harriet Geer, Evaluation of Literature Searching Systems, American Documentation 8, 149-151 (April 1957).

Kessler, M. M., Bibliographic Coupling Between Scientific Papers, American Documentation 14, 10-25 (January 1963).

Kessler, M. M., Comparison of the Results of Bibliographic Coupling and Analytic Subject Indexing, American Documentation 16, 223-233 (July 1965). (Also available as Report No. R-7, Cambridge, Mass., Massachusetts Institute of Technology, January 28, 1963, 30 p.) [I-18]

Kessler, M. M., Bibliographic Coupling Extended in Time: Ten Case Histories, Information Storage and Retrieval 1, 169-187 (November 1963).

Kessler, M. M., An Experimental Communication Center for Scientific and Technical Information, Report No. 4G-0002, Cambridge, Mass., Massachusetts Institute of Technology, Lincoln Laboratory, March 1962, 20 p.

King, Donald W., Some Notes on Search Strategies, in Information Retrieval Among Examining Patent Offices (Proceedings of the Fourth Annual Meeting of the Committee for International Cooperation in Information Retrieval Among Examining Patent Offices, Washington, D.C., October 1964), Ed. by H. Pfeffer, Washington, D.C., Spartan Books, 1965, p. 378-386.

King, Donald W., Evaluation of Coordinate Index Systems During File Development, Journal of Chemical Documentation 5, 96-99 (May 1965). [II-31]

King, Donald W., Evaluation of Analog-to-Digital Converter Patent Information Retrieval Systems, Report No. WRA PO 12, Denver, Colo., Westat Research Analysts, Inc., June 1964, 42 p. plus appendices, PB 166 491. [I-19]

King, Donald W., Designs of Experiments in Information Retrieval, in Proceedings of the Social Statistics Section, American Statistical Association Meeting, Cleveland, Ohio, 1963, p. 103-118.

King, Donald W., Experimental Designs in Work and Time Studies, Journal of Chemical Documentation 5, 205-209 (November 1965).

King, Donald W. and J.M. Daley, Quality Control of Coordinate Indexing, in Parameters of Information Science (Proceedings of the American Documentation Institute Annual Meeting, Philadelphia, Pa., October 5-8, 1964), Washington, D.C., Spartan Books, 1964, p. 389-392.

King, Donald W. and Patricia M. McDonnell, Evaluation of Coordinate Index Systems During File Development. Part II, An Application, Journal of Chemical Documentation 6, 235-240 (November 1966).

King, Donald W. and P. James Terragno, Some Techniques for Measuring System Performance, in Parameters of Information Science (Proceedings of the American Documentation Institute Annual Meeting, Philadelphia, Pa., October 5-8, 1964), Washington, D.C., Spartan Books, 1964, p. 393-398.

Klempner, Irving M., Methodology for the Comparative Analysis of Information Storage and Retrieval Systems: A Critical Review, American Documentation 15, 210-216 (July 1964). [III-16]

Knable, J. P., II, Experiment Comparing Keywords Found in Indexes and Abstracts Prepared by Humans with Those in Titles, American Documentation 16, 123-124 (April 1965).

Kochen, Manfred, Toward Document Retrieval Theory: Relevance-Recall Ratio for Text Containing One Specified Query Term, in Automation and Scientific Communication. Proceedings, Part 3 (papers presented at the 26th Annual Meeting of the American Documentation Institute, Chicago, Ill., October 1963), Ed. by P. C. Janaske, Washington, D.C., American Documentation Institute, 1963, p. 439-442. [III-17]

Kochen, Manfred, Preliminary Operational Analysis of a Computer-Based On-Demand Document Retrieval System Using Coordinate Indexing, in Some Problems in Information Science, New York and London, Scarecrow Press, Inc., 1965, p. 47-60. (Also in Report No. AFCRL-64-87, Final Report to Air Force Cambridge Research Lab., Yorktown Heights, N.Y., IBM Corp., Thomas J. Watson Research Center, April 2, 1964, p. 81-112), AD 600 113.

Kochen, Manfred, Research or Search, IBM Research Note NC-37, Yorktown Heights, N. Y., IBM Corp., Thomas J. Watson Research Center, 1961, 6 p.

Kochen, Manfred and E. Wang, Concerning the Possibility of a Cooperative Information Exchange, IBM Journal of Research and Development 6, 270-271 (April 1962).

Koller, Herbert R., Ethel Marden and Harold Pfeffer, The Haystaq System: Past, Present and Future. Appendix A: Evaluation of Literature Searching Systems, in Proceedings of International Conference on Scientific Information (Washington, D.C., November 1958), Vol. II, Washington, D.C., National Academy of Sciences-National Research Council, 1959, p. 1169-1171.

Korotkin, Arthur L. and Lawrence H. Oliver, The Effect of Subject Matter Familiarity and the Use of an Indexing Aid Upon Inter-Indexer Consistency, Bethesda, Md., General Electric Co., Information Systems Operation, February 14, 1964, 17 p. [II-7]

Korotkin, Arthur L. and Lawrence H. Oliver, A Method for Computing Indexer Consistency, Bethesda, Md., General Electric Co., Information Systems Operation, February 1964, 8 p. [II-6]

Korotkin, Arthur L., Lawrence H. Oliver and D.R. Burgis, Indexing Aids, Procedures and Devices, Report No. RADC-TR-64-582, Bethesda, Md., General Electric Co., Information Systems Operation, April 1965, AD 616 342.

Kraft, Donald H., A Comparison of Keyword-In-Context (KWIC) Indexing of Titles With a Subject Heading Classification System, American Documentation 15, 48-52 (January 1964). [I-20]

Kriebel, Charles H., Statistical Decision Criteria in the Evaluation of Information System Performance, ONR Research Memorandum No. 131, Pittsburgh, Pa., Carnegie Institute of Technology, Graduate School of Industrial Administration, August 1964, 33 p. [IV-25]

Kriebel, Charles H., Optimum Design of Production Decision and Information System, Ph.D. Dissertation, Massachusetts Institute of Technology, Department of Industrial Management, Cambridge, Mass., June 1964.

Krug, J.F., Comparison of Uniterm, Descriptor and Role-Indicator Methods of Encoding Literature for Information Retrieval, M.A. Thesis, University of Chicago Graduate Library School, 1964.

Kuhns, J.L., Appendix C: Measures of Retrieval Effectiveness, in Word Correlation and Automatic Indexing, Progress Report No. 2 to Council on Library Resources, Canoga Park, Calif., Ramo-Wooldridge Corporation, December 21, 1959, p. 1C-10C.

Kuhns, J.L., Section 7: System Effectiveness Study, in Research on an Advanced NASA Information System, Report to National Aeronautics and Space Administration, Canoga Park, Calif., Bunker Ramo Corporation, October 11, 1963, p. 66-82. [IV-26]

Kuhns, J.L. and Christine A. Montgomery, A Comparative Study of Fragment Versus Document Retrieval, in Parameters of Information Science (Proceedings of the American Documentation Institute Annual Meeting, Philadelphia, Pa., October 5-8, 1964), Washington, D.C., Spartan Books, 1964, p. 369-377. [I-21]

Kurmey, William John, An Evaluation of Automatically Prepared Abstracts and Indexes, Thesis University of Chicago Graduate Library School, June 1964, 64 p. [II-32]

Kyle, Barbara, Consistency Analysis of Two Indexers in Using K.C. for Political Science Material, London, England, National Book League, May 1962, 4 p. [II-8]

Kyle, Barbara, Information Retrieval and Subject Indexing: Cranfield and After, Journal of Documentation 20, 55-69 (June 1964).

Lancaster, F. Wilfred, Engineering Information Storage: Indexing vs. Classification, Machine Design 37, 105-107 (January 7, 1965).

Lancaster, F. Wilfred, Some Observations on the Performance of EJC Role Indicators in a Mechanized Retrieval System, Special Libraries 55, 696-701 (December 1964). [II-9]

Lancaster, F. Wilfred and J. Mills, Testing Indexes and Index Language Devices: The ASLIB Cranfield Project, American Documentation 15, 4-13 (January 1964).

Langefors, Borje, Information System Design Computations Using Generalized Matrix Algebra, BIT (Nordisk Tidskrift for Informations-Behandling) 5, 96-121 (1965). [IV-28]

Langefors, Borje, Some Approaches to the Theory of Information Systems, BIT (Nordisk Tidskrift for Informations-Behandling) 3, 229-254 (1963). [IV-27]

Lefkovitz, David, Automatic Stratification of Descriptors, Report No. 64-03, Philadelphia, Pa., University of Pennsylvania, Moore School of Electrical Engineering, September 15, 1963, 138 p., AD 423 647.

Lehman, M., Serial Matrix Storage System, IRE Transactions on Electronic Computers EC-10, 247-252 (June 1961).

Leibowitz, Jacob and E. D. Lewis, Design of a Mechanized Patent Search System with the Aid of Preliminary Evaluation Studies, in Information Retrieval Among Examining Patent Offices (Proceedings of Fourth Annual Meeting of the Committee for International Cooperation in Information Retrieval Among Examining Patent Offices, Washington, D. C., October 1964), Ed. by H. Pfeffer, Washington, D. C., Spartan Books, 1965, p. 387-412.

Levy, N. P. and R. W. Sigmon, Economic Analysis of a Technical Information Dissemination System, in Abstracts (of papers submitted to International Federation for Documentation 1965 Congress, Washington, D. C., October 10-15, 1965), Washington, D. C., Secretariat, 1965 FID Congress, p. 73.

Linder, L. H., Indexing Costs for 10,000 Documents, in Automation and Scientific Communication. Short Papers, Part 2 (papers contributed to the 26th Annual Meeting of the American Documentation Institute, Chicago, Ill., October 1963), Ed. by H. P. Luhn, Washington, D. C., American Documentation Institute, 1963, p. 147-148. [II-33]

Linder, L. H., Comparative Costs of Document Indexing and Book Cataloging, Special Libraries 56, 724-726 (December 1965).

Lipetz, Ben-Ami, Labor Costs, Conversion Costs, and Compatibility in Document Control Systems, American Documentation 14, 117-122 (April 1963).

Lipetz, Ben-Ami, Design of an Experiment for Evaluation of the Citation Index as a Reference Aid, in Automation and Scientific Communication. Short Papers, Part 2 (papers contributed to the 26th Annual Meeting of the American Documentation Institute, Chicago, Ill., October 1963), Ed. by H. P. Luhn, Washington, D. C., American Documentation Institute, 1963, p. 265-266.

Lipetz, Ben-Ami, The Effect of a Citation Index on Literature Use by Physicists, in Abstracts (of papers submitted to International Federation for Documentation 1965 Congress, Washington, D. C., October 10-15, 1965), Washington, D. C., Secretariat, 1965 FID Congress, p. 26.

Lipetz, Ben-Ami, Evaluation of the Impact of a Citation Index in Physics, Report No. AIP/DRP-C1-3, New York, American Institute of Physics, Documentation Research Project, September 1, 1964, 64 p. plus appendices.

Macha, H. R., A Technique for Evaluating Information Analysis Methods and Personnel, American Documentation 4, 35-49 (April 1953).

MacMillan, Judith T. and Issac D. Welt, A Study of Indexing Procedures in a Limited Area of the Medical Sciences, American Documentation 12, 27-31 (January 1961). [II-10]

Magnavox Research Laboratories, Mathematical Models for Information Systems Design and a Calculus of Operations Report No. RADC TR-61-96, Final Report to Rome Air Development Center, Torrance, Calif., Magnavox Research Lab., October 27, 1961, 178 p., AD 266 577.

Maizell, Robert E., Standards for Measuring the Effectiveness of Technical Library Performance, IRE Transactions on Engineering Management PGEM-7, 69-72 (June 1960).

Maizell, Robert E., Value of Titles for Indexing Purposes, Revue de la Documentation 27, 126-127 (August 1960). [I-22]

Maloney, Clifford J., Abstract Theory of Retrieval Coding, in Proceedings of the International Conference on Scientific Information (Washington, D.C., November 1958), Vol. II, Washington, D.C., National Academy of Sciences-National Research Council, 1959, p. 1365-1382.

Manley, Ron, Inadequacy of Varying Depth of Indexing and Other Document Collection Approaches to Information Retrieval for Researchers, American Documentation 12, 204-205 (July 1961).

Margolis, J., Citation Indexing and Evaluation of Scientific Papers, Science 155, 1213-1219 (March 10, 1967).

Markel, Gene A., A Concept for Modeling and Evaluating Information - Producing Systems, Report No. 352.13-R-1, State College, Pa., HRB-Singer Inc., January 28, 1966, 36 p., AD 628 495.

Maron, M.E., Automatic Indexing: An Experimental Inquiry, in Machine Indexing, Progress and Problems, Washington, D.C., American University, 1961, p. 236-265.

Maron, M.E. and J.L. Kuhns, On Relevance, Probabilistic Indexing and Information Retrieval, Journal of the ACM 7, 216-244 (July 1960). [III-18]

Mathieu, J. and S. Barlen, Guiding Principles for Time and Cost in Documentation Work, Report No. TIL/T 4966, Translation JPRS No. 591 369, Great Britain, Ministry of Aviation, Technical Information and Library Services, January 1959, 34 p., AD 229 229.

Mayer, S.E., Model for Description of Information Retrieval Systems, Preliminary Report No. 2, McLean, Va., Human Sciences Research, Inc., December 1965.

McGrath, Joseph E. et al, A Systematic Framework for Comparison of System Research Methods, Report No. HSR-TN-59/7 SM, Arlington, Va., Human Sciences Research, Inc., November 1959, 65 p., AD 229 923.

Meetham, A.R., Probabilistic Pairs and Groups of Words in a Text, Language and Speech 7, 98-106 (April-June 1964).

Meetham, A.R., Preliminary Studies for Machine Generated Index Vocabularies, Language and Speech 6, 22-36 (January-March 1963). [II-34]

Meier, Richard L., Efficiency Criteria for the Operation of Large Libraries, The Library Quarterly 31, 215-234 (July 1961).

Melton, Jessica S., A Use for the Techniques of Structural Linguistics in Documentation Research, Report No. CSL: TR-4, Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, September 1964, 20 p. [III-19]

Melton, Jessica S. and William Buscher, The Cleverdon-WRU Experiment: Search Strategies, in Information Retrieval in Action (papers presented at 1962 Conference at the Center for Documentation and Communication Research), Cleveland, Ohio, The Press of Western Reserve University, 1963, p. 85-91.

Miller, Eugene et al, Comparison of Conventional Grouping and Inverted Grouping of Codes for the Storage and Retrieval of Chemical Data, Report No. AFOSR-TN 58-366, Washington, D.C., Documentation, Inc., May 1958, 17 p.

Minder, Thomas L. and Gerald J. Lazorick, Automation of the Penn State University Acquisitions Department, in Automation and Scientific Communication. Proceedings, Part 3 (papers presented at the 26th Annual Meeting of the American Documentation Institute, Chicago, Ill., October 1963), Ed. by P. C. Janaske, Washington, D. C., American Documentation Institute, 1963, p. 455-460. [II-35]

Mitnick, L. L., Indexing Criteria, Technical Memo No. 64-75, Washington, D. C., General Electric Co., August 12, 1964, 14 p. [III-20]

Montague, Barbara A., Analysis of Designing, Installation and Operation of a Coordinate Indexing System Using Links and Roles for the Plastics Department of DuPont, Journal of Chemical Documentation 4, 251-255 (October 1964).

Montague, Barbara A., Testing, Comparison, and Evaluation of Recall, Relevance and Cost of Coordinate Indexing with Links and Roles, in Parameters of Information Science (Proceedings of the American Documentation Institute Annual Meeting, Philadelphia, Pa., October 5-8, 1964), Washington, D. C., Spartan Books, 1964, p. 357-367.

Montague, Barbara A., Testing, Comparison and Evaluation of Recall, Relevance and Cost of Coordinate Indexing with Links and Roles, American Documentation 16, 201-208 (July 1965). [II-11]

Montgomery, Christine and Don R. Swanson, Machine-like Indexing by People, American Documentation 13, 359-366 (October 1962). [II-36]

Mooers, Calvin N., The Intensive Sample Test for the Objective Evaluation of the Performance of Information Retrieval Systems, Report No. RADC-TN-59-160, Report to U.S. Air Force, Rome Air Development Center, Cambridge, Mass., Zator Co., August 1959, 20 p. [IV-29]

Mooers, Calvin N., Seven System Models, Report No. ZTB-133, Pt. 2, Cambridge, Mass., Zator Co., August 1959, 39 p.

Morris, Jack C., Evolution or Involution: Notes Critical of the Uniterm System of Indexing, Journal of Cataloging and Classification 10, 111-118 (July 1954).

Mueller, Max W., An Evaluation of Information Retrieval Systems, Memorandum Report No. 7170, Burbank, Calif., Lockheed Aircraft Corp., Operations Research Division, September 30, 1959, 114 p. [I-23]

Mueller, Max W., Time, Cost and Value Factors in Information Retrieval (paper presented at Information Retrieval Systems Conference, Poughkeepsie, N.Y., September 21-23, 1959), White Plains, N.Y., IBM Corp., Data Processing Division, 16 p.

Myatt, DeWitt O. and T. E. Upham, A Quantitative Technique for Designing the Technical Information Center, Journal of Chemical Documentation 1, 18-24 (November 1961).

National Academy of Sciences-National Research Council, The Metallurgical Searching Service of the American Society for Metals-Western Reserve University: An Evaluation, a Report by an Ad Hoc Committee of the Office of Documentation, Pub. No. 1148, Washington, D. C., National Academy of Sciences-National Research Council, 1964, 96 p. [II-37]

National Library of Medicine, Information Systems Division Design for a Test Program to Evaluate Demand Search Performance of Medlars, Bethesda, Md., National Library of Medicine, June 1966.

National Science Foundation, Summary of Study Conference on Evaluation of Document Searching Systems and Procedures, Washington, D.C., National Science Foundation, February 10, 1965, 24 p. plus appendices.

Neelameghan, A., The Keyword-In-Context Index: An Evaluation, in Documentation Periodicals: Coverage. Arrangement. Scatter. Seepage. Compilation, Bangalore, India, Documentation Research Training Center, 1963, p. 127-140. [II-38]

Newill, Vaun A. and William Goffman, Searching Titles by Man, Machine, and Chance, in Parameters of Information Science (Proceedings of the American Documentation Institute Annual Meeting, Philadelphia, Pa., October 5-8, 1964), Washington, D.C., Spartan Books, 1964, p. 421-423. [II-39]

Newman, Simon M., Economic Justification - Factors Establishing System Costs, in Information Retrieval Management, Ed. by Lowell Hattery and Edward McCormick, Detroit, Mich., American Data Processing, Inc., 1963, p. 117-119.

O'Connor, John, Correlation of Indexing Headings and Title Words in Three Medical Indexing Systems, American Documentation 15, 96-104 (April 1964). [I-24]

O'Connor, John, Review of ASLIB-Cranfield Research Project: Report on the First Stage and Interim Report on the Test Programme, Journal of Documentation 17, 252-261 (December 1961). [I-5]

O'Connor, John, Mechanized Indexing Studies of MSD Toxicity, Report No. AFOSR-64-0682, Philadelphia, Pa., Institute for Scientific Information, January 1964, 37 p. plus appendices, AD 436 523.

O'Connor, John, Some Remarks on Mechanized Indexing and Some Small Scale Empirical Results, in Machine Indexing, Progress and Problems, Washington, D.C., American University, 1961, p. 266-277.

O'Connor, John, Some Suggested Mechanized Indexing Investigations Which Require No Machines, American Documentation 12, 198-203 (July 1961). [III-22]

Ohlman, Herbert, The Activity Spectrum: A Tool for Analyzing Information Systems, in Proceedings of the Symposium on Education for the Information Sciences, September 1965, Washington, D.C., Spartan Books, p. 155-166.

Oliver, Lawrence H. et al, An Investigation of the Basic Processes Involved in the Manual Indexing of Scientific Documents, Bethesda, Md., General Electric Co., February 1966, PB 169 415.

Overmyer, LaVahn, Test Program for Evaluating Procedures for the Exploitation of Literature of Interest to Metallurgists, American Documentation 13, 210-222 (April 1962). [II-40]

Overmyer, LaVahn, The Dollars and Cents of Basic Operations in Information Retrieval, in Information Retrieval In Action (papers presented at 1962 Conference at the Center for Documentation and Communication Research), Cleveland, Ohio, The Press of Western Reserve University, 1963, p. 199-211. [II-42]

Overmyer, LaVahn, An Analysis of Output Costs and Procedures for an Operational Searching Service, American Documentation 14, 123-142 (April 1963). [II-41]

Painter, Ann F., An Analysis of Duplication and Consistency of Subject Indexing Involved in Report Handling at the Office of Technical Services, U.S. Department of Commerce, Springfield, Va., Clearinghouse for Federal Scientific and Technical Information, March 1963, 135 p., PB 181 501. [II-12]

Painter, Ann F., Convertibility Potential Among Government Information Agency Indexing Systems, Library Resources and Technical Services 7, 274-281 (Summer 1963).

Palatt, P.E., Testing the Effectiveness of a Thesaurus-Controlled Subject Index, in Abstracts (of papers submitted to the International Federation for Documentation 1965 Congress, Washington, D.C., October 10-15, 1965), Washington, D.C., Secretariat, 1965 FID Congress, p. 82.

Papier, Lawrence S., Evaluation of Science Communication Systems, CRDL Special Publication 4-65, Edgewood Arsenal, Md., U.S. Army Edgewood Arsenal, Chemical Research and Development Laboratories, February 1965, 49 p., AD 615 108. [IV-30]

Parker, Edwin B., The User's Place in an Information System, American Documentation 17, 26-27 (January 1966).

Payne, Dan and John F. Hale, Automatic Abstracting Evaluation Support, Report No. RADC-TDR-64-30, Final Report to Rome Air Development Center, Griffis Air Force Base, N.Y., Rome Air Development Center, Information Processing Branch, February 1964, 38 p. plus appendices. [I-25]

Payne, Dan, John F. Hale and Sara J. Munger, An Informative Abstracting Technique: Development and Verification, in Abstracts (of papers submitted to the International Federation for Documentation 1965 Congress, Washington, D.C., October 10-15, 1965), Washington, D.C., Secretariat, 1965 FID Congress, p. 82.

Perry, James W., Defining the Query Spectrum - The Basis for Developing and Evaluating Information Retrieval Methods, IEEE Transactions on Engineering Writing and Speech EWS 6, 10-27 (September 1963). [III-23]

Perry, James W. and Allen Kent, Selectivity Criteria for Systems Evaluation, in Documentation and Information Retrieval. An Introduction to Basic Principles and Cost Analysis, Cleveland, Ohio, Western Reserve University Press, 1957.

Perry, James W. and Allen Kent, Introduction to Machine Literature Searching, in Tools for Machine Literature Searching, New York, Interscience Publishers, Inc., 1958, p. 3-18.

Perry, James W., Allen Kent and Madeline M. Berry, Operational Criteria for Designing Information Retrieval Systems, in Machine Literature Searching, New York, Interscience Publishers, Inc., 1956, p. 41-48. [IV-31]

Pietsch, Eric H.E., Evaluation of Mechanized Documentation at the Gmelin Institut, in Punched Cards. Their Applications to Science and Industry, 2nd Ed., Ed. by R.S. Casey, J.W. Perry, M.M. Berry and A. Kent, New York, Reinhold Publishing Corp., 1958, p. 571-618.

Postley, John A., Behavioral Factors in Information Systems, in Information Systems Workshop: The Designer's Responsibility and His Methodology, Washington, D.C., Spartan Books, 1962, p. 83-95.

Postley, John A., Report on a Study of Behavioral Factors in Information Systems, Report to the National Science Foundation on NSF Contract C-265, Los Angeles, Calif., Hughes Dynamics, Inc., 1962, 88 p. plus appendices, AD 419 622.

Pratt, L., Analysis of Library Systems; A Bibliography, Special Libraries 55, 688-695 (December 1964).

Pronko, Eugene, Comparative Studies of Retrieval Systems: Problems and Prospects (prepared for UNESCO Working Group No. 2 on Automatic Documentation-Storage and Retrieval, Moscow, November 11-16, 1963), Washington, D.C., National Science Foundation, November 1963, 26 p. plus bibliography.

Proposed Scope of Area 5, Organization of Information for Storage and Retrospective Search, in Proceedings of the International Conference on Scientific Information (Washington, D.C., November 1958), Vol. II, Washington, D.C., National Academy of Sciences-National Research Council, 1959, p. 817-821.

Rafter, S., Analysis of Citations in Source Papers in Three Retrospective Bibliographies on (1) Hydrofoil Craft, (2) Continental Drift, and (3) Tsunamis (Seismic Sea Waves), in Abstracts (of papers sumbitted to the International Federation for Documentation 1965 Congress, Washington, D.C., October 10-15, 1965), Washington, D.C., Secretariat, 1965 FID Congress, p. 83.

Rath, G.J., A. Resnick and T.R. Savage, Comparisons of Four Types of Lexical Indicators of Content, American Documentation 12, 126-131 (April 1961).

Reed, David M. and Donald J. Hillman, Document Retrieval Theory, Relevance, and The Methodology of Evaluation. Report No. 4, Canonical Decomposition, Bethlehem, Pa., Lehigh University, Center for the Information Sciences, August 12, 1966, 33 p., PB 173 129.

Rees, Alan M., Semantic Factors, Role Indicators et Alia - Eight Years of Information Retrieval at Western Reserve University, ASLIB Proceedings 15, 350-363 (December 1963).

Rees, Alan M., Relevancy and Pertinency in Indexing, American Documentation 13, 93-94 (January 1962).

Rees, Alan M., The Cleverdon-WRU Experiment: Search Results, in Information Retrieval in Action (papers presented at the 1962 Conference at the Center for Documentation and Communication Research), Cleveland, Ohio, The Press of Western Reserve University, 1963, p. 93-99. [I-6]

Rees, Alan M., Review of a Report of the ASLIB-Cranfield Test of the Index of Metallurgical Literature of Western Reserve University, Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, October 1963, 32 p. [I-7]

Rees, Alan M., The ASLIB-Cranfield Test of the Western Reserve University Indexing System for Metallurgical Literature: A Review of the Final Report, American Documentation 16, 73-76 (April 1965).

Rees, Alan M., The Evaluation of Retrieval Systems, Report No. CSL:TR-5, Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, July 1965, 21 p. [III-24]

Rees, Alan M., The Relevance of Relevance to Testing and Evaluation of Document Retrieval Systems, ASLIB Proceedings 18, 316-324 (November 1966).

Rees, Alan M. and Tefko Saracevic, Measurability of Relevance, in Progress in Information Science and Technology (Proceedings of the American Documentation Institute Annual Meeting, Santa Monica, Calif., October 3-7, 1966), Santa Monica, Calif., Adrianne Press, 1966, p. 225-234. (Also available as Report No. CSL:TR-7, Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, August 1966, 19 p.).

Rees, Alan M. and Douglas G. Schultz, A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching, Progress Report No. 3, Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, December 1966, 11 p.

Resnick, A., Relative Effectiveness of Document Titles and Abstracts for Determining Relevance of Documents, Science 134, 1004-1005 (October 6, 1961). [II-43]

Resnick, A. and C. B. Hensley, The Use of Diary and Interview Techniques in Evaluating a System for Disseminating Technical Information, American Documentation 14, 109-116 (April 1963). (Also available as Technical Report No. 17-055, Yorktown Heights, N.Y., IBM Corp., Advanced Systems Development Division, December 29, 1961, 83 p.) [II-44]

Resnick, A. and T.R. Savage, A Re-evaluation of Machine-Generated Abstracts, Human Factors 2, 141-146 (1960).

Resnick, A. and T.R. Savage, The Consistency of Human Judgments of Relevance, American Documentation 15, 93-95 (April 1964). [II-45]

Rial, J.F., Final Report on the ROUT Document Retrieval System, Report No. ESD-TDR-64-96, prepared for U.S. Air Force, Air Force Systems Command, Bedford, Mass., The Mitre Corp., May 1964, 66 p.,AD 601 145. [II-46]

Richmond, Phyllis A., Review of the Cranfield Project, American Documentation 14, 307-311 (October 1963). [I-8]

Riddles, A.J., Computer Based Concept Searching of U.S. Patent Claims, Report No. ITIRC-004, Yorktown Heights, N.Y., IBM Corp., Technical Information Retrieval Center, August 1, 1965.

Rocchio, Joseph, Performance Indices for Document Retrieval Systems, in Information Storage and Retrieval, Scientific Report No. ISR-8, Report to the National Science Foundation, Cambridge, Mass., Harvard University, Computation Laboratory, December 1964, p. III-1 - III-18. [II-47]

Rocchio, Joseph and Margaret Engel, Test Design and Detailed Retrieval Results, in Information Storage and Retrieval, Scientific Report No. ISR-8, Report to the National Science Foundation, Cambridge, Mass., Harvard University, Computation Laboratory, December 1964, p. X-1 - X-25. [II-48]

Rocchio, Joseph and Gerard Salton, Information Search Optimization and Interactive Retrieval Techniques, in AFIPS Conference Proceedings, Volume 27, Part I, 1965 Fall Joint Computer Conference, Washington, D.C., Spartan Books, 1965, p. 293-305.

Rodgers, Dorothy J., A Study of Intra-Indexer Consistency, Washington, D.C., General Electric Co., Information System Operation, January 1961, 25 p. [II-14]

Rodgers, Dorothy J., A Study of Inter-Indexer Consistency, Washington, D.C., General Electric Co., Information System Operation, September 29, 1961, 59 p. [II-13]

Rothstein, S., The Measurement and Evaluation of Reference Service, Library Trends 12, 456-472 (January 1964).

Ruhl, Mary Jane, Chemical Documents and Their Titles: Human Concept Indexing vs. KWIC-Machine Indexing, American Documentation 15, 136-141 (April 1964). [I-26]

Salton, Gerard, The Evaluation of Automatic Retrieval Procedures --- Selected Test Results Using the SMART System, in Information Storage and Retrieval, Scientific Report No. ISR-8, Report to the National Science Foundation, Cambridge, Mass., Harvard University, Computation Laboratory, December 1964, p. IV-1 - IV-36. [II-49]

Salton, Gerard, The Evaluation of Automatic Retrieval Procedures --- Selected Test Results Using the SMART System, American Documentation 16, 209-222 (July 1965).

Salton, Gerard, An Evaluation Program for Associative Indexing, in Statistical Association Methods for Mechanized Documentation (Symposium Proceedings), NBS Misc. Publ. 269, Washington, D.C., National Bureau of Standards, December 15, 1965, p. 201-210.

Salton, Gerard, Evaluation Procedures for Computer-Based Retrieval Systems, in Abstracts (of papers submitted to International Federation for Documentation 1965 Congress, Washington, D.C., October 10-15, 1965), Washington, D.C., Secretariat, 1965 FID Congress, p. 30.

Salton, Gerard and M.E. Lesk, The SMART Automatic Document Retrieval System - An Illustration, Communications of the ACM 8, 391-398 (June 1965).

Saracevic, Tefko and Alan M. Rees, Identification and Control of Variables in Information Retrieval Experimentation, Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, January 1966, 24 p.

Savage, T., A Boyle's Law for Indexer Consistency, American Documentation 16, 33 (January 1965).

Savage, T., C.B. Hensley and A. Resnick, On the Problem of Testing Hypothesis in Information Retrieval, Yorktown Heights, N.Y., IBM Corp., Advanced Systems Development Division, April 12, 1960

Schreiber, A.L., Measuring Relevance of an Item of Information to the Command of a Complex Man-Machine System, Report No. HSR-TN-61/1-SM, McLean, Va., Human Sciences Research, Inc., January 1961, AD 257 607.

Schüller, J.A., Experience with Indexing and Retrieving by UDC and Uniterms, ASLIB Proceedings 12, 373-389 (November 1960). [I-27]

Schultz, Claire K., Some Characteristics of an Efficient Information Retrieval System, Journal of Chemical Documentation 2, 103-105 (April 1962).

Schultz, Claire K. and Richard H. Orr, Evaluating Indexing by Reference to Term-Choice Patterns of a Criterion Group, in Abstracts (of papers submitted to International Federation for Documentation 1965 Congress, Washington, D.C., October 10-15, 1965), Washington, D.C., Secretariat, 1965 FID Congress, p. 86.

Schultz, Claire K. and Clayton A. Shepherd, A Computer Analysis of the Merck Sharpe and Dohme Indexing System, American Documentation 12, 83-92 (June 1961). [II-50]

Schultz, Claire K., Wallace L. Schultz and Richard H. Orr, Comparative Indexing: Terms Supplied by Biomedical Authors and by Document Titles, American Documentation 16, 299-312 (October 1965).

Schultz, Claire K., Phyllis D. Schwartz and Leon Steinberg, A Comparison of Dictionary Use Within Two Information Retrieval Systems, American Documentation 12, 247-253 (October 1961). [I-28]

Shaffer, Barbara B., An Analysis of Factors Causing Irrelevant Answers to Machine Literature Searches and Proposed Solutions to the Problem, in Parameters of Information Science (Proceedings of the American Documentation Institute Annual Meeting, Philadelphia, Pa., October 5-8, 1964), Washington, D.C., Spartan Books, 1964, p. 433-436. [II-51]

Sharp, John R., Review of Aitchison-Cleverdon, Alan M. Rees, and National Academy of Sciences-National Research Council Reports, Journal of Documentation 20, 170-174 (September 1964). [I-9]

Shaw, R.R., Information Retrieval, Science 140, 606-609 (May 10, 1963), AD 407 864.

Sherrington, Andrew M. and Richard H. Orr, Recurrent Bibliographies as Current Awareness Tools for Problem-Oriented Research: Evaluation of an Experimental Product of MEDLARS, in Abstracts (of papers submitted to International Federation for Documentation 1965 Congress, Washington, D.C., October 10-15, 1965), Washington, D.C., Secretariat, 1965 FID Congress, p. 87.

Shilling, Charles W., Requirements for a Scientific Mission-Oriented Information Center, American Documentation 14, 49-53 (January 1963).

Shilling, Charles W., Page Cost Policy of Biological Journals, Communique 9-63, Biological Sciences Communication Project, Washington, D.C., American Institute of Biological Sciences, March 1963, 1 v.

Shoffner, Ralph M., A Technique for Organization of Large Files, American Documentation 13, 95-103 (January 1962). [IV-32]

Sinnett, Jefferson D., An Evaluation of Links and Roles Used in Information Retrieval, Report No. ML TDR 64-152, Wright-Patterson Air Force Base, Ohio, Air Force Materials Laboratory, July 1964, 139 p., AD 606 192. (Report of M.S. Thesis, AF Institute of Technology, Air University, December 1963, 300 p., AD 432 198) [II-15]

Slamecka, Vladimir, Classificatory, Alphabetical, and Associative Schedules as Aids in Coordinate Indexing, American Documentation 14, 223-228 (July 1963). [II-16]

Slamecka, Vladimir and J. Jacoby, Effect of Indexing Aids on the Reliability of Indexers, Report No. RADC-TDR-63-116, Bethesda, Md., Documentation, Inc., June 1963, AD 410 032.

Snyder, M.B., et al., Methodology for Test and Evaluation of Document Retrieval Systems: A Critical Review and Recommendations, Report No. HSR-RR-66/6-SK, Report to the National Science Foundation, McLean, Va., Human Sciences Research, Inc., January 1966, 75 p. plus appendices.

Sparks, David E., Mark M. Chodrow and Gail M. Walsh, A Methodology for the Analysis of Information Systems, Report No. R-4003-1, Final Report to the National Science Foundation, Wakefield, Mass., Information Dynamics Corp., May 1965, 102 p., PB 169 264.

Sparks, David E., Mark M. Chodrow, Gail M. Walsh and L.L. Laine, A Methodology for the Analysis of Information Systems, in Abstracts (of papers submitted to International Federation for Documentation 1965 Congress, Washington, D.C., October 10-15, 1965), Washington, D.C., Secretariat, 1965 FID Congress, p. 88.

Sparks, David E., Mark M. Chodrow, Gail M. Walsh and L.L. Laine, A Methodology for the Analysis of Information Systems, Report No. R-4003-1, Appendix A to Final Report, Wakefield, Mass., Information Dynamics Corp., May 1965, 520 p., PB 169 265.

Spiro, H.T. and Allan D. Kotin, A Cost Analysis of an Automated System for the Library of Congress, in Automation and the Library of Congress, Washington, D.C., Library of Congress, 1963, p. 27-88.

Sprague, Ralph H., Jr., A Comparison of Systems for Selectively Disseminating Information, Indiana Business Report No. 38, Bloomington, Indiana, Indiana University, Graduate School of Business, 1965, 70 p. [I-29]

Stevens, Mary Elizabeth, A Machine Model of Recall, in Information Processing (Proceedings of International Conference on Information Processing 1959), Paris, UNESCO, 1960, p. 309-315.

Stevens, Mary Elizabeth, Problems of Evaluation, in Automatic Indexing: A State-of-the-Art Report, NBS Monograph 91, Washington, D.C., National Bureau of Standards, March 30, 1965, p. 143-163. [III-25]

Stevens, Mary Elizabeth and Genevie H. Urban, Automatic Indexing Using Cited Titles, in Statistical Association Methods for Mechanized Documentation (Symposium Proceedings), NBS Misc. Publ. 269, Washington, D.C., National Bureau of Standards, December 15, 1965, p. 213-215.

Stevens, Mary Elizabeth and Genevie H. Urban, Training a Computer to Assign Descriptors to Documents: Experiments in Automatic Indexing, in AFIPS Conference Proceedings, Volume 25, 1964 Spring Joint Computer Conference, Baltimore, Md., Spartan Books, 1964, p. 563-575. [II-52]

Stevens, Norman D., A Comparative Study of Three Systems of Information Retrieval, Ph. D. Thesis, New Brunswick, N. J., Rutgers, The State University Press, 1960. [I-30]

Stevens, Norman D., A Comparative Study of Three Systems of Information Retrieval: A Summary, American Documentation 12, 243-246 (October 1961).

Stiassny, Simon, Mathematical Analysis of Various Superimposed Coding Methods, American Documentation 11, 155-169 (April 1960). [I-31]

Stiles, H. Edmund, The Association Factor in Informational Retrieval, Journal of the ACM 8, 271-279 (April 1961).

Stiles, H. Edmund, Progress in the Use of the Association Factor in Information Retrieval, in Symposium on Materials Information Retrieval (Proceedings), Tech. Doc. Report No. ASD-TDR-63-445, Dayton, Ohio, Aeronautical Systems Division, AF Materials Laboratory, May 1963, p. 143-153, AD 407 609.

Stitelman, Joseph, Donald P. Stein and Donald W. King, Search Teses of an Experimental File of Electronic Circuits, in Information Retrieval Among Examining Patent Officers (Proceedings of Fourth Annual Meeting of the Committee for International Cooperation in Information Retrieval Among Examining Patent Offices, Washington, D.C., October 1964), Ed. by H. Pfeffer, Washington, D.C., Spartan Books, 1965, p. 413-423.

Swanson, Don R., An Experiment in Automatic Text Searching, in Word Correlation and Automatic Indexing, Phase I, Final Report, Report No. C82-OU4, Canoga Park, Calif., Thompson Ramo Wooldridge, Inc., April 30, 1960, 36 p. plus appendix.

Swanson, Don R., Information Retrieval: State-of-the-Art, in Proceedings of the Western Joint Computer Conference 1961, Glendale, Calif., Western Joint Computer Conference, 1961, p. 239-246.

Swanson, Don R., The Evidence Underlying the Cranfield Results, The Library Quarterly 35, 1-20 (January 1965). [I-10]

Swanson, Don R., Searching Natural Language Text by Computer, Science 132, 1099-1104 (October 21, 1960). [II-53]

Swanson, Don R., Interrogating a Computer in Natural Language, in Information Processing 1962, Ed. by C. M. Popplewell, Amsterdam, North-Holland Publishing Co., 1963, p. 124-127. [III-26]

Swanson, Don R., On Indexing Depth and Retrieval Effectiveness, in Second Congress on the Information System Sciences (Proceedings), Ed. by Joseph Spiegels and Donald E. Walker, Washington, D.C., Spartan Books, 1965, p. 311-319.

Swets, John A., Information Retrieval Systems, Science 141, 245-250 (July 19, 1963). [IV-33]

Swets, John A., Measures of Effectiveness of Information Retrieval Systems: A Review and A Proposal, Cambridge, Mass., Bolt, Beranek and Newman, Inc., March 1963, 23 p.

Swid, R. E., Linear vs. Inverted File Searching on Serial Access Machines, in Automation and Scientific Communication. Short Papers, Part 2 (papers contributed to the 26th Annual Meeting of the American Documentation Institute, Chicago, Ill., October 1963), Ed. by H. P. Luhn, Washington, D.C., American Documentation Institute, 1963, p. 321-322.

Tague, Jean, Effectiveness of a Pilot Information Service for Educational Research Materials, Cooperative Research Project No. 1743, Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, 1963, 56 p. [II-54]

Tague, Jean, Matching of Question and Answer Terminology in an Educational Research File, American Documentation 16, 26-32 (January 1965).

Tague, Jean, et al., Case Histories: Phase Report, in Information Retrieval in Action (papers presented at 1962 Conference at the Center for Documentation and Communication Research), Cleveland, Ohio, The Press of Western Reserve University, 1963, p. 241-317.

Taube, Mortimer, An Evaluation of "Use Studies" of Scientific Information, in Emerging Solutions for Mechanizing the Storage and Retrieval of Information. Studies in Coordinate Indexing, Vol. V, Washington, D.C., Documentation, Inc., 1959, p. 46-71.

Taube, Mortimer, Evaluation of Information Systems for Report Utilization, in Studies in Coordinate Indexing, Vol. I, Washington, D.C., Documentation, Inc., 1953, p. 96-110. [I-32]

Taube, Mortimer, A Note on the Pseudo-Mathematics of Relevance, American Documentation 16, 69-72 (April 1965). [III-28]

Taube, Mortimer, A Note on the Evaluation of the WRU Semantic Code as an Example of Generic Coding, American Documentation 13, 185-186 (April 1962).

Taube, Mortimer and Laurence B. Heilprin, The Relation of the Size of the Question to the Work Accomplished by a Storage and Retrieval System, Report No. AFOSR-TN-57-483, Washington, D.C., Documentation, Inc., August 1957, 13 p., AD 136 476.

Taube, Mortimer, et al., Cost as the Measure of Efficiency of Storage and Retrieval Systems, in Studies in Coordinate Indexing, Vol. II, Washington, D.C., Documentation, Inc., 1956, p. 18-33.]III-27]

Taulbee, Orrin E., Bibliography on the Evaluation of Information Storage and Retrieval Systems, Akron, Ohio, Goodyear Aerospace Corp., August 25, 1965.

Tell, Bjorn V., Auditing Procedures for Information Retrieval Systems, in Abstracts (of papers submitted to International Federation for Documentation 1965 Congress, Washington, D.C., October 10-15, 1965), Washington, D.C., Secretariat, 1965 FID Congress, p. 28-29.

Tell, Bjorn V., et al., On Criteria for Evaluating IR Systems, in Trend Report (submitted by Swedish Classification Research Group to the 29th FID Conference in Stockholm, 1963), Bromma, Sweden, Tekniska Litteratursallskapet, 1963, p. T1-T8. [III-29]

Thompson, H.B., Efficiency - A Paramount Need in Information Storage and Retrieval, in Symposium on Materials Information Retrieval (Proceedings), Tech. Doc. Report No. ASD-TDR-63-445, Dayton, Ohio, Aeronautical Systems Division, AF Materials Laboratory, May 1963, p. 67-73, AD 407 609.

Thompson Ramo Wooldridge, Inc., Automatic Abstracting, Report No. C107-3U1, Final Report to U.S. Air Force, Rome Air Development Center, Canoga Park, Calif., Thompson Ramo Wooldridge, Inc., February 2, 1963, 53 p. [II-55]

Thorne, R.G., The Efficiency of Subject Catalogues and the Cost of Information Searches, Journal of Documentation 11, 130-148 (September 1955). [I-33]

Thrall, R.M., W.V. Caldwell, C.H. Coombs and M.S. Schoeffler, A Model for Evaluating the Output of Intelligence Systems, Naval Research Logistics Quarterly 8, 25-40 (March 1961).

Thrall, R.M., C.H. Coombs and W.V. Caldwell, Linear Model for Evaluating Complex Systems, Naval Research Logistics Quarterly 5, 347-361 (December 1958).

Timms, H.L., Case Study: Increasing the Efficiency of the Use of Information, Aerospace Research Applications Center, in Abstracts (of papers submitted to International Federation for Documentation 1965 Congress, Washington, D.C., October 10-15, 1965), Washington, D.C., Secretariat, 1965 FID Congress, p. 40-41.

Tinker, John F., Imprecision in Meaning Measured by Inconsistency of Indexing, American Documentation 17, 96-102 (April 1966).

Trachtenberg, Alfred, Automatic Document Classification using Information Theoretical Methods, in Automation and Scientific Communication. Short Papers, Part 2 (papers contributed to the 26th Annual Meeting of the American Documentation Institute, Chicago, Ill., October 1963), Ed. by H.P. Luhn, Washington, D.C., American Documentation Institute, 1963, p. 349-350.

U.S. Congress, Senate, Committee on Government Operations, Documentation, Indexing, and Retrieval of Scientific Information. A Study of Federal and Non-Federal Science Information Processing and Retrieval Programs, Document No. 113, 86th Congress, 2nd Session, Washington, D.C., Government Printing Office, 1960, p. 103.

Valvoda, M.A., A Comparison of Manual and Machine Searching Techniques, in Information Retrieval in Action (papers presented at 1962 Conference at Center for Documentation and Communication Research), Cleveland, Ohio, The Press of Western Reserve University, 1963, p. 51-72.

Van Dijk, Marcel, Essai sur le Cout de la Recherche Documentaire par les Methodes d'Indexation Coordonnée, Revue de la Documentation 30, 143-145 (November 1963).

Van Oat, James G., J.L. Schultz, R.E. McFarlane, F.H. Kvalnes and A.W. Riester, Links and Roles in Coordinate Indexing and Searching: An Economic Study of Their Use and An Evaluation of Their Effect on Relevance and Recall, Journal of Chemical Documentation 6, 95-101 (May 1966).

Verhoeff, Jacobus, William Goffman and Jack Belzer, Inefficiency of the Use of Boolean Functions for Information Retrieval, Communications of the ACM 4, 557-558, 594 (December 1961). [III-30]

Vickery, Brian C., On Retrieval System Theory, Washington, D.C., Butterworths, 1965, 191 p.

Vickery, Brian C., The Structure of Information Retrieval Systems, in Proceedings of International Conference on Scientific Information (Washington, D.C., November 1958), Vol. II, Washington, D.C., National Academy of Sciences-National Research Council, 1959, p. 1275-1289.

Vickery, Brian C., The Statistical Method in Indexing, Revue de la Documentation 28, 56-62 (May 1961).

Votaw, David F., Jr., Statistical Science and Information Technology, in Second Congress on the Information Systems Sciences (Proceedings), Ed. by Joseph Spiegels and Donald E. Walker, Washington, D.C., Spartan Books, 1965, p. 219-228.

Wadsworth, Harrison M. and Robert E. Booth, Reliability Distributions of Documentation Systems, Technical Note No. 15, Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, April 1960, 39 p.

Wadsworth, Harrison M. and Robert E. Booth, The Reliability of Documentation Processes and Systems, Technical Note No. 16, Cleveland, Ohio, Western Reserve University, Center for Documentation and Communication Research, April 1960, 16 p.

Wall, Eugene, Further Implications of the Distribution of Index Term Usage, in Parameters of Information Science (Proceedings of the American Documentation Institute Annual Meeting, Philadelphia, Pa., October 5-8, 1964), Washington, D.C., Spartan Books, 1964, p. 457-466.

Wallace, Everett M., Rank Order Patterns of Common Words as Discriminators of Subject Content in Scientific and Technical Prose, in Statistical Association Methods for Mechanized Documentation (Symposium Proceedings), NBS Misc. Publ. 269, Washington, D.C., National Bureau of Standards, December 15, 1965, p. 225-229.

Wallace, Everett M., User Requirements, Personal Indexes, and Computer Support, Report No. SP-2535, Santa Monica, Calif., System Development Corp., July 25, 1966, 7 p., AD 636 833.

Wayne, Ivor, A Survey of Users of the American Society for Metals-Western Reserve University Searching Service, Report No. BSSR:352 to the National Science Foundation, Washington, D.C., Bureau of Social Science Research, Inc., July 1962, 27 p. plus appendix. [II-56]

Webb, Kenneth W., W.C. Suhler, G.G. Heller and S.P. Todd, Evaluation Models for Information Retrieval and Command Control Systems (EMIR), Report No. TIE4-0887, Bethesda, Md., IBM Corp., August 18, 1964, 15 p.

Weik, Martin H. and V.J. Confer, Survey of Scientific and Technical Information Retrieval Schemes Within the Department of the Army, Report No. BRL 1169, Aberdeen Proving Ground, Md., Ballistic Research Laboratories, July 1962, 89 p. plus appendix.

Westbrook, J., How to Measure Quality, Research Laboratory Bulletin (General Electric Co.), Fall 1961, p. 20-22.

White, D.R.J., D.L. Scott and R.N. Schultz, POED - A Method of Evaluating System Performance, IEEE Transactions on Engineering Management EM-10, 177-182 (December 1963).

Williams, Ann S. and Isaac D. Welt, Analysis and Indexing of Psychopharmacological Literature, Journal of Chemical Documentation 5, 176-179 (August 1965).

Williams, John H., Jr., A Discriminant Method for Automatically Classifying Documents, in AFIPS Conference Proceedings, Vol. 24, 1963 Fall Joint Computer Conference, Baltimore, Md., Spartan Books, 1963, p. 161-166. [III-31]

Williams, John H., Jr., Results of Classifying Documents with Multiple Discriminant Functions, in Statistical Association Methods for Mechanized Documentation (Symposium Proceedings), NBS Misc. Publ. 269, Washington, D.C., National Bureau of Standards, December 15, 1965, p. 217-224.

Williams, Thyllis M., Language Engineering, in Documentation in Action (based on 1956 Conference on Documentation at Western Reserve University), Ed. by J.H. Shera, A. Kent and J.W. Perry, New York, Reinhold Publishing Corp., 1956, p. 330-337.

Wright, R.C. and C.W.J. Wilson, Classification With Peek-a-boo for Indexing Documents on Aerodynamics: An Experiment in Retrieval, in Proceedings of International Conference on Scientific Information (Washington, D.C., November 1958) Vol. I, Washington, D.C., National Academy of Sciences-National Research Council, 1959, p. 771-801.

Wuest, Francis J., Studies in the Methodology of Measuring Information Requirements and Use Patterns. Report No. 1: Questionnaire and Appendix, Bethlehem, Pa., Lehigh University, May 1965, 38 p.

Wurm, Bengt R., The Relation between Number of Documents and Number of Terms and Their Discriminatory Power in Information Retrieval for U.S. Pharmaceutical Patents, in Information Retrieval Among Examining Patent Offices (Proceedings of Foruth Annual Meeting of the Committee for International Cooperation Among Examining Patent Offices, Washington, D.C., October 1964), Ed. by Harold Pfeffer, Washington, D.C., Spartan Books, 1966, p. 349-360.

Wyllys, Ronald E., Document Searches and Condensed Representations, in Joint Man-Computer Indexing and Abstracting, Report No. Mitre SS-13, Bedford, Mass., The Mitre Corp., 1962, p. 37-60. (Also available as Report No. SP-804, Santa Monica, Calif., System Development Corp., May 1, 1962.) [IV-34]

Yow, Bobby Gene, A Comparison between a Coordinate Indexing System and an Algorithmic Associative Indexing System for Information Retrieval, M.S. Thesis, University of Pittsburgh, 1966.

U. S. GOVERNMENT PRINTING OFFICE : 1967—30[illegible]

NBS TECHNICAL PUBLICATIONS

PERIODICALS

JOURNAL OF RESEARCH reports National Bureau of Standards research and development in physics, mathematics, chemistry, and engineering. Comprehensive scientific papers give complete details of the work, including laboratory data, experimental procedures, and theoretical and mathematical analyses. Illustrated with photographs, drawings, and charts.

Published in three sections, available separately:

• Physics and Chemistry

Papers of interest primarily to scientists working in these fields. This section covers a broad range of physical and chemical research, with major emphasis on standards of physical measurement, fundamental constants, and properties of matter. Issued six times a year. Annual subscription: Domestic, $5.00; foreign, $6.00*.

• Mathematics and Mathematical Physics

Studies and compilations designed mainly for the mathematician and theoretical physicist. Topics in mathematical statistics, theory of experiment design, numerical analysis, theoretical physics and chemistry, logical design and programming of computers and computer systems. Short numerical tables. Issued quarterly. Annual subscription: Domestic, $2.25; foreign, $2.75*.

• Engineering and Instrumentation

Reporting results of interest chiefly to the engineer and the applied scientist. This section includes many of the new developments in instrumentation resulting from the Bureau's work in physical measurement, data processing, and development of test methods. It will also cover some of the work in acoustics, applied mechanics, building research, and cryogenic engineering. Issued quarterly. Annual subscription: Domestic, $2.75; foreign, $3.50*.

TECHNICAL NEWS BULLETIN

The best single source of information concerning the Bureau's research, developmental, cooperative and publication activities, this monthly publication is designed for the industry-oriented individual whose daily work involves intimate contact with science and technology—*for engineers, chemists, physicists, research managers, product-development managers, and company executives*. Annual subscription: Domestic, $1.50; foreign, $2.25*.

*Difference in price is due to extra cost of foreign mailing.

NONPERIODICALS

Applied Mathematics Series. Mathematical tables, manuals, and studies.

Building Science Series. Research results, test methods, and performance criteria of building materials, components, systems, and structures.

Handbooks. Recommended codes of engineering and industrial practice (including safety codes) developed in cooperation with interested industries, professional organizations, and regulatory bodies.

Miscellaneous Publications. Charts, administrative pamphlets, Annual reports of the Bureau, conference reports, bibliographies, etc.

Monographs. Major contributions to the technical literature on various subjects related to the Bureau's scientific and technical activities.

National Standard Reference Data Series. NSRDS provides quantitative data on the physical and chemical properties of materials, compiled from the world's literature and critically evaluated.

Product Standards. Provide requirements for sizes, types, quality and methods for testing various industrial products. These standards are developed cooperatively with interested Government and industry groups and provide the basis for common understanding of product characteristics for both buyers and sellers. Their use is voluntary.

Technical Notes. This series consists of communications and reports (covering both other agency and NBS-sponsored work) of limited or transitory interest.

CLEARINGHOUSE

The Clearinghouse for Federal Scientific and Technical Information, operated by NBS, supplies unclassified information related to Government-generated science and technology in defense, space, atomic energy, and other national programs. For further information on Clearinghouse services, write:

Clearinghouse
U.S. Department of Commerce
Springfield, Virginia 22151

Order NBS publications from:
Superintendent of Documents
Government Printing Office
Washington, D.C. 20402

U.S. DEPARTMENT OF COMMERCE
WASHINGTON, D.C. 20230

OFFICIAL BUSINESS

POSTAGE AND FEES PAID
U.S. DEPARTMENT OF COMMERCE