DOCUMENT RESUME

ED 090 920 IR 000 529

AUTHOR Spuck, Dennis W.; And Others
TITLE Information Retrieval: Presentation and Demonstration of an Interactive Computer-Based Search Program.
INSTITUTION Wisconsin Univ., Madison. Wisconsin Information Systems for Education.
PUB DATE Apr 74
NOTE 46p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$1.85 PLUS POSTAGE
DESCRIPTORS *Algorithms; *Bibliographies; Computer Oriented Programs; *Computers; Educational Research; *Information Retrieval; *Search Strategies; Symposia
IDENTIFIERS Educational Resources Information Center; ERIC; Hash Coding; *Wisconsin Information System For Education; WISE ONE

ABSTRACT

A symposium with four major presentations centering on the topic of computer-based information retrieval. Also highlighted are several features of the Wisconsin Information System for Education (WISE-ONE) and the Educational Resources Information Center (ERIC) system. The first paper in the series discusses the development, current capabilities and future directions of the WISE-ONE system and the second analyzes the history and present status of ERIC. The third paper focuses on file search strategies for the WISE-ONE information retrieval system, comparing various search algorithms used in information retrieval programs and leading to an in-depth consideration of the hash coding scheme used in WISE-ONE. The final paper treats the development of search strategies for effective computerized literature searches of the ERIC data base. The use of the ERIC thesaurus and of the logical operators AND, OR, and NAND is stressed and details are provided relating to a demonstration search supported by a UNIVAC 1108 computer. (PB)

2603

ED 090920

INFORMATION RETRIEVAL: PRESENTATION AND DEMONSTRATION OF AN INTERACTIVE COMPUTER-BASED SEARCH PROGRAM

A Symposium Presented

at the

Annual Meeting of the American Educational Research Association

Chicago, Illinois

April 19, 1974

Contributors

Dennis W. Spuck, Chairman
Donald N. McIsaac
Tom Olson
Roy Tally

THE WISE-ONE BIBLIOGRAPHIC RETRIEVAL PROGRAM

A University of Wisconsin Research Tool

Don McIsaac

It is a pleasure to introduce the WISE-ONE search program today. Computer retrieval of information is the wave of the future and it's been satisfying to be a part of this development. I will take a short time to discuss some of the historical antecedents of the project development, current status of the system, and the prospects for future application. Dennis Spuck's paper is a discussion of the ERIC system and its implications for educational research. Tom Olson will present some of the technical aspects of the WISE-ONE speed and responsiveness. Roy Tally, a major user of the system, will discuss the dissemination and use characteristics of the program. Tom and Roy will then provide an on-line demonstration of the system, illustrating the highlights of the program's use.

Wisconsin's School of Education's first involvement in the information retrieval game began with the ERIC Clearinghouse on school house facilities. The project, under the direction of Howard Wakefield, took a computer based direction offering a batch oriented search of the existing ERIC files on the University's Control Data 3600. At the suggestion of Professor Wakefield the School of Education invested in the collection of ERIC microfiche presently located in the School of Education IMC. The clearinghouse on the Madison campus marked the beginning of a long and productive relationship with ERIC. We, however, were not the only people to see the vision of the future.

Alternatives for Bibliographic Retrieval

A number of alternatives for computerized search of the education literature exist, and the number of organizations offering search capability increases each year. Many are associated with universities—or research organizations with a high motivation for making the specifics of our growing literature base available in an effective and efficient manner. Moreover, there is an increasing number of data bases prepared for the computerized search of materials. It is the collection, abstracting, coding and processing of the citations which makes rapid retrieval a realistic and feasable possibility. The educational ERIC clearinghouses offer a superb abstract service which today enjoys a wide and varied distribution. Investments in the tools for use of these materials will continue to increase as the cost of computing continues to drop and the available tapes of information demonstrate their usefullness. The experiment at Wisconsin will be reproduced on other campuses, further extending the utility of the ERIC tapes and other machine readable files.

A winter issue of the ERIC Newsletter (Vol. 7, Number 1) specified ten machine readable files offered through seven nationwide computer-search services in education. The files included were:

AIM--Abstracts of Instructional Materials in Vocational and Technical Education.

ARM--Abstracts of Research Materials in Vocational and Technical Education.

CPE--Current Projects in Education.

ExCHILD--Exceptional Children Abstracts.

FI4C--Fugitive Information Collection of Contra Costa County.

FIDO--Fugitive Information Data Organizer.

FRC--Field Reader Catalog

PACE--Pacesetters in Innovation.

PCL--Professional Curriculum Library

Psych Ab--Psychological Abstracts

CIJE--Current Index to Journals of Education

RIE--Research in Education

The summary of nationwide computer-search services is reproduced in this document for your interest and information.

NATIONWIDE COMPUTER-SEARCH SERVICES IN EDUCATION

Description	SMIRS (Clearinghouse/ Lane IED)	PROBE (Indiana)	SMERC (San Mateo County)	SDC/ERIC (System Development Corporation)	RISE (Pennsylvania)	SRIS (Phi Delta Kappa)	LIRS (Lockheed)
Sources	RIE, CIJE, AIM, ARM	RIE, CIJE	RIE, CIJE, FIDO, PCL, special collections	RIE, CIJE	RIE, CIJE, Psych-Ab, AIM, ARM, Ex-Child, special collections, fugitive data sources	RIE, CIJE	RIE, CIJE, CPE, PACE, AIM, ARM, FRC, ExChild, PsychAb, F14C
Number of Descriptors	up to 20	up to 12	negotiable	up to 129 related terms	10 or more	up to 12	up to 10
Output	130 abstracts &/or bibliographic citations	100 abstracts, up to 800 author/title citations	50 abstracts, 50 bibliographic citations, 150 identification number/title citations	50 abstracts, complete with bibliographic data	combination of abstracts & bibliographic citations for first 100 items; numerical hit list for additional 100	first 100 articles include abstracts; subsequent items furnish complete documentation, no abstracts	50 abstracts
Cost	$16 per search ($17 for foreign orders)	$12 per search of RIE *or* CIJE; $18 per combined search	average of $15 to $20 per search	$25 per search	$10 per searchable file for 10 descriptors, $25 for 20 or more descriptors	$5 per search	$25 per question
Turnaround Time	1 to 2 weeks	5 days (delivery by UPS)	10 days to 2 weeks	2 days to 1 week	1 week	3 weeks	1 week
Address & Phone Number	School Management Information Retrieval Service, ERIC Clearinghouse on Educational Management, University of Oregon, Eugene, OR 97403 Ph. (503) 686-5943	PROBE, Education Library, Rm 30, School of Education, Indiana University, Bloomington, IN 47401 Ph. (812) 337-5718 *Attn:* Robert N. Benninghoff	San Mateo Educational Resources Center, 333 Main St., Redwood City, CA 94063 Ph. (415) 360-1441 *Attn:* Frank W. Mattas	System Development Corporation, SDC/ERIC Search Service *West coast:* 2500 Colorado Ave., Santa Monica, CA 90406 Ph. (213) 393-9411 *East coast:* 5827 Columbia Pike, Falls Church, VA 22041 Ph. (703) 820-2220	Research and Information Services for Education, 198 Allendale Rd., King of Prussia, PA 19406 Ph. (215) 265-6056	Phi Delta Kappa, School Research Information Service, 8th & Union, Bloomington, IN 47401 Ph. (812) 339-1156	Lockheed Information Retrieval Program Office. *West coast:* 3251 Hanover St., Dept. 52-08, Bldg. 201 Palo Alto, CA 94304 Ph. (415) 493-4411 ext. 45034 *East coast:* 405 Lexington Ave., New York, NY 10017 Ph. (212) 697-7171

A survey of the scientific-technical tape services compiled by (Carroll 1970) is available as ERIC Document 044165. The report presents a survey of commercially available tape services which can provide libraries and information centers with data bases of scientific and technical literature.

History of the Wisconsin Effort

The availability of automated search capability is a well documented fact. The cost of the services is reasonable, ranging from $10.00-25.00 depending on the nature of the search and the extent of service required. The range in turnaround for the services varies from two days to three weeks. With so much activity for high quality retrieval capability, why would the University of Wisconsin-School of Education embark on a similar kind of program? There are several reasons.

1. The availability of the ERIC microfiche called for a quick and easy way to search for the references of a complicated search formula.
2. An interactive search capability permits the on-line development of a search formula providing the user with instant feedback on the progress of the search. It was important to develop a program which would conveniently offer this search capability.
3. The feedback to the user could be a list of ERIC documents. The proximity of the computer terminal to the microfiche facilitates the search and reduces the need to produce hard copy abstracts from the computer tape file. (While this was an initial motivation, history has not supported the premise.

There is very little direct access to the microfiche files when abstracts are so easy to produce on campus.)

4. Many research projects funded on campus have existing accounts with the computing center. A bibliographic retrieval system on the University computer simplifies the procedures for procuring a search. The catalog is as far away as the nearest terminal.
5. Bibliographic retrieval implemented on the local computer would reduce the cost of computer searches. Current methods of hash coding and organization of mass storage produces a rapid return product on a low cost basis. In time, we felt we could make the search capability available to every staff member and student on the University campus on a convenient and low cost basis.
6. Other agencies of the state can benefit from the search and retrieval capability.

With the above rationale in mind, the initial development began in 1970 as a project of the Wisconsin Information System for Education (WISE). We were fortunate to attract a computer science major, Tom Olson, to our organization. For a period of a year he lived and breathed the technical aspects of information science. The project became a class project in computer science, where the basic search routine was conceptualized and tested. The initial version of the program produced lists of accession numbers. Since the system was originally designed to search the ERIC RIE files, and since the ERIC microfiche were so handy, the initial version served a most valuable purpose.

It was only a short time, however, before the system was trained to also produce the author, title, and citation information on command. The search capability was generalized to accomodate multiple files so that the information contained in the CIJE, AIM and ARM files could be easily accessed. At this time, 1971, we recognized that the system had a general attraction for researchers on the Madison campus. A number of workshops were scheduled for staff and students within the School of Education.

The fortuitous result of our advertising was to attract the attention of Roy Tally, employed by the Department of Public Instruction in the information services group. His background in the field of information science produced our most valued critic. His insights and understanding helped produce the product to be demonstrated for you today. While he is still employed by the Department of Public Instruction, he is also one of us. He continues his partnership, contributing to the many enhancements of the system.

Present WISE-ONE Capabilities

In addition to the initial capabilities of boolean search operators, AND, OR, and NAND, the WISE-ONE program will permit paranthetic formulas nested 15 deep. Each inquiry to the system provides an immediate response on the status of the search. The number of citations referenced by the current entry and the status of the search queue are immediately known by the user. The user may SAVE a complex search formula for later use or application to an alternate file. A search queue which has been carefully derived may also be SAVEd for subsequent intersection investigations with alternate queues. Either the queue or the formula may be recalled through the use of the ADD command.

The current University computer has terminals located in Green Bay, Milwaukee, Parkside, Sheboygan, and the Physical Sciences Laboratory located in Oregon, Wisconsin. In addition, there are fifteen high speed terminals located on the Madison campus. The high speed output may be sent to any of these sites from the central computer. The listing commands permit the imposition of a publishing window based upon publishing dates or file serial number limits. In addition to the citation information, the user may request that abstracts be forwarded to the high speed printers at any of the sites connected to the central computer.

Dissemination Activities

The WISE-ONE program enjoys an active schedule of demonstrations around the campus and the state. The school districts in Wisconsin participate in the information game through an information service offered in the Department of Public Instruction. Participants in this aspect of the dissemination contact the DPI with their information request. Roy or his staff field the questions filling out a request form. The requests are then processed through the WISE-ONE program and the resulting lists and abstracts are forwarded to the DPI for final edit checks and distribution.

The Vocational Board is currently supporting a dissemination program for the vocational districts in the state. Each of the district offices has a terminal linked to the Madison computer. Each has a series of demonstrations and instructional workshops scheduled with a field consultant in information retrieval. The system now supports the AIM and ARM files, in addition to the ERIC RIE and CIJE files.

Continuing Technical Support

The project was transferred to the Madison Academic Computing Center (MACC) in September, 1973. It has become a library program of the Center, supported by the Center's technical staff. The Center works with an advisory staff comprised of School of Education, Department of Public Instruction and Vocational Board personnel. The purpose of this group is to assist in the continued development and dissemination of the system. Technical manuals, user documentation and consulting assistance on the system are offered under the auspices of the computing center.

The Future of the WISE-ONE System

The key to the future of the system lies in the marketing concepts employed to further the dissemination of computer based bibliographic retrieval. An Information Science publication (Kuel, 1972) on the marketing perspectives for "ERIC-like" information systems outlines three dimensions of marketing thought and technique. Marketing is a social process which focuses on the concept of "needs satisfaction" to consumers. The bibliographic search concept, of which WISE-ONE is an example, requires a marketing emphasis. In order to satisfy this market responsibility, the manager of information systems should engage in product definitions which are based upon assessment of user needs. The information system manager must be in a position to identify new product opportunities and move to develop them as an on-going part of the system. The WISE-ONE program now has an answer-back capability permitting users to express their desires, demands and problems through the terminal. This mechanism keeps the technical staff in touch with the users.

We must define more efficient patterns of dissemination. It is good to extend the use of the system to the vocational schools, but what other publics exist which can benefit from the rapid retrieval of bibliographic information. It is imperative that we continue to identify more efficient means of using the system resources. The computer industry breeds obsolescence through invention. New hardware and software techniques should be exploited when consumers can benefit. While the development costs of the WISE-ONE system involved only a small amount of money, fifteen thousand dollars is a reasonable estimate of the resources deployed to the effort. The public investment in information systems of this type requires further dissemination. More outlets for the kind of service available on the Madison campus are required. The incredible federal investment in the information base of ERIC is a valuable public service. The real benefit of the ERIC documents and services will be realized when the system trickles into the public school curriculum and research areas. One outlet for the research and information requirements of public education is through ERIC-like information systems. Marketing, in its best sense, offers the best possibility for making the information available to continue the improvement of education.

With this brief review of the WISE-ONE history, I look forward to my colleagues discussion of the ERIC system and the WISE-ONE bibliographic search and retrieval program.

The ERIC System: History and Analysis

by

Dennis W. Spuck
Assistant Professor
Department of Educational Administration
University of Wisconsin-Madison

A primary data base of interest to educational researchers and practitioners is that developed by the Educational Resources Information Centers (ERIC). The ERIC system was conceived in the early and mid-sixties and grew to maturity through the late sixties and early seventies. This paper traces the development of the ERIC system through these periods. Following this historical account, ERIC process and products are analyzed leading to an awareness of important limitations in ERIC search outcomes, and suggestions are made for resolving some of these limitations.

Historical Overview

According to Burchinal (1969, p. 56), ERIC was designed to accomplish three important objectives:

1) To guarantee ready access to the world's English-language literature relevant to education. In information science terminology this is the documentation function of the program.
2) To generate new information products by reviewing, summarizing, and interpreting current information on priority topics. This is the information analysis function of the system. Products include bibliographies, state-of-knowledge papers, critical reviews, and interpretive summaries.

3) Infuse information about educational developments, research findings, and outcomes of exemplary programs into educational planning and operations.

The need for the ERIC system germinated with the awareness of growing quantities of information generated as a result of research thrusts directed toward technological and social concerns of the post-Sputnik era. The United States Office of Education was inundated with research reports literally piling up in the halls. Further, reports received by the U.S.O.E. were narrowly disseminated when they were received, but thereafter became hard to find and later even difficult to identify. Much important research was simply not available to professionals in the field of education after a very short while.

ERIC development began in 1959 with a federally supported feasibility study of an information system to meet the in-house needs of the Office of Education. At this time too, Western Reserve University was contracted to develop a thesaurus of terms useful in indexing educational documents.

The ERIC idea was soon expanded from an in-house utility concept to one of generalized value for the field of education. The basic organizational plan was accepted in 1964, marking the founding of ERIC and was moved toward implementation with the funding of ERIC under ESEA in 1965. The centralized model of such information centers as the Defense Documentation Center (DDC), the National Technical Information Service (NTIS) and the NASA Scientific and Technical Information Facility were rejected for a more decentralized plan involving subject and topic oriented clearinghouses located at appropriate points throughout the nation. This decentralized model was consistent with the decentralized view of education in the United States. In order to coordinate the efforts of the clearinghouses, a central coordinating office was established within the Office of Education. (It is now located in the National Institute of Education.) ERIC central was charged with the technical coordination and evaluation of the clearinghouses, as well as with the formulation of operating procedures, and policies of the ERIC system. The clearinghouses were allowed considerable autonomy in the interpretation and implementation of central policy.

In September 1965, a Panel on Educational Terminology was formed by the Office of Education with James L. Eller as chairman, to advise and lead the ERIC Thesaurus development. The Panel report, "Guidelines for the Development of a Thesaurus of Educational Terms," (United States Department of Health, Education and Welfare, 1966, p. 13) defined a Thesaurus as "a term association list structured to enable indexers and subject analysts to describe the subject information of a document to a desired level of specificity at input, and to permit searchers to describe in mutually precise terms the information required at output." The Thesaurus authority list thus forms a key element in the ERIC network.

Also during 1965 the first twelve clearinghouses were established; within two years there were a total of 18 clearinghouses. At present there are 16 clearinghouses, although the actual number reached a peak of 20 at one point.

The subject or topical orientations of the present 16 clearinghouses are listed below.

I. Subject-Oriented Clearinghouses:

Languages and Linguistics

Science, Mathematics and Environmental Education

Reading and Communication Skills

Social Studies/Social Science Education

II. Consumer Groups Clearinghouses:

Career Education

Disadvantaged

Early Childhood Education

Handicapped and Gifted Children

Rural and Small Schools

III. Functional Groups Clearinghouses:

Educational Management

Teacher Education

Counseling and Personnel Services

VI. Level Oriented Clearinghouses:

Junior Colleges

Higher Education

V. Technically Oriented Clearinghouses:

Tests, Measurement and Evaluation

Information Resources

The clearinghouses were charged with the acquisition screening, abstracting and cataloging of fugitive documents, those documents not normally available through formal publishing channels. The documents collected were processed and compiled centrally. The first collection of these document references and abstracts was published as Research In Education (RIE) in November 1965. A private contractor was employed by ERIC to coordinate the central collection of the manu[illegible]ts, references and abstracts received from the individual clearinghouses. The first such contract was awarded in June 1965 to North American Rockwell; they organized the basic information which was printed by the Government Printing Office as RIE. This contract was awarded to the LEASCO corporation in January 1970.

Also charged to a private contractor was the responsibility for reproducing and disseminating upon request the documents catalogued in the ERIC system. Two primary forms of reproduction are available: hard copy and microfiche. The first contract for the Educational Documents Reproduction Service (EDRS) was granted to the Bell and Howell Company, in November 1965. This service was transferred to the National Cash Register Company in January 1968 and to LIPCO in 1971.

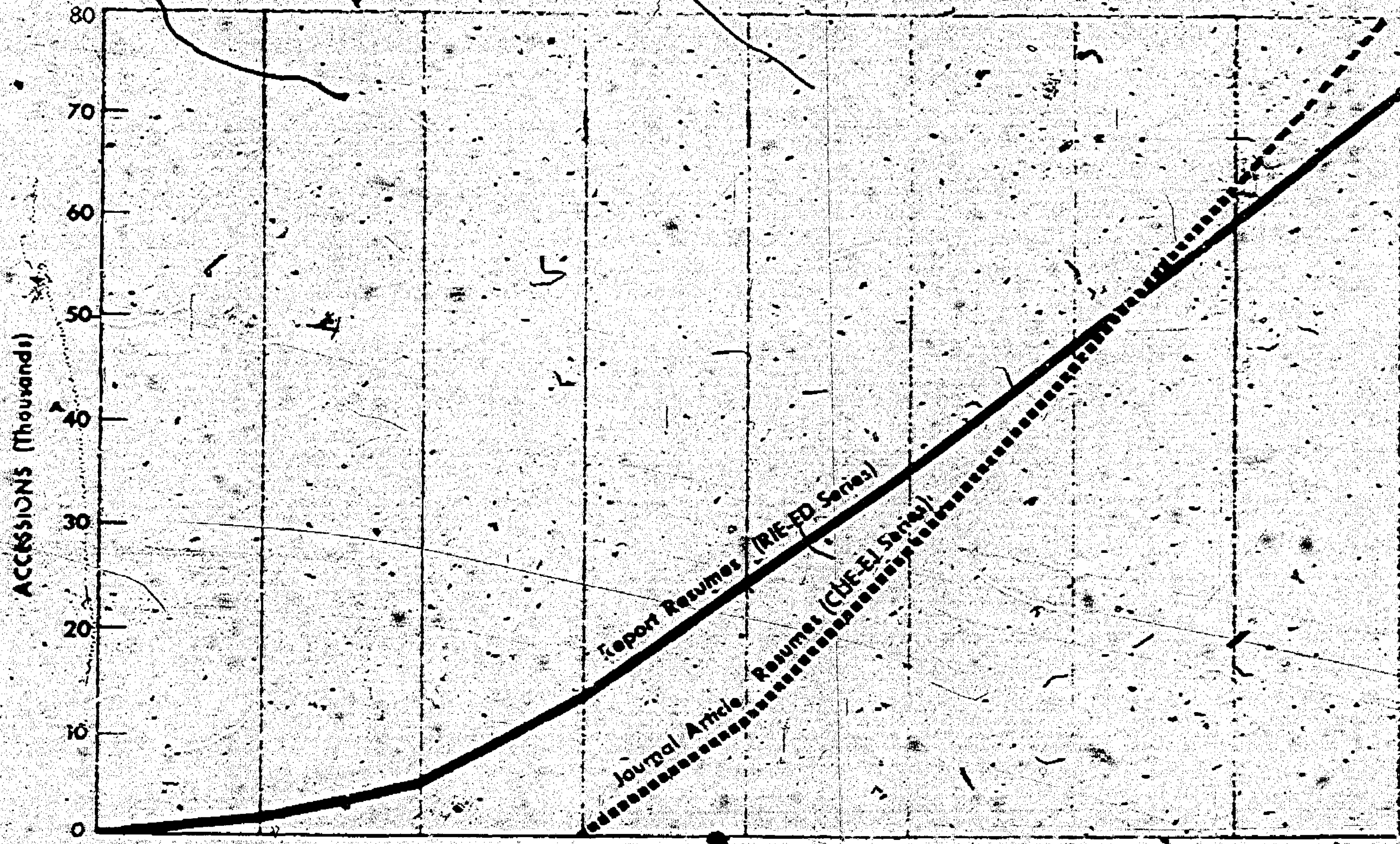
In 1969 the scope of the ERIC system was extended beyond the fugitive document collection to include the indexing of journal articles of interest to professionals in the field of education. The collection of these references was first published as the Current Index to Journals in Education (CIJE) in 1969. This index, like the ERIC Thesaurus is published by Crowell, Collier and Macmillian Information Sciences (CCM) Incorporated.

The two ERIC document collections, RIE and CIJE continue to grow at a rapid pace. Figure 1 depicts numerically and graphically the

Figure 1

BEST COPY AVAILABLE

ERIC DATA BASE-FILE GROWTH



		1966	1967	1968	1969	1970	[illegible]	1972	1973
Reports	Accessions Added:	1834	3469	8803	10,453	10,456	12,330	12,230	[illegible]
	Cumulative Total:		5303	14,106	24,559	35,015	47,345	[illegible]	[illegible]
Articles	Accessions Added:				11,707	15,892	17,672	[illegible]	[illegible]
	Cumulative Total:				11,707	27,599	45,271	62,751	[illegible]

ERIC Data Base Users Interchange, February, 1973, Attachment 73.1

growth of these two data base files. While the RIE collection began three years earlier than the CIJE collection, the CIJE collection is, at present, larger than the document file, due to the more rapid growth of the journal file. CIJE is growing at about 18,000 to 19,000 accessions a year, while the document file is growing at approximately 13,000 to 14,000 accessions a year. Both of these collections should surpass the 100,000 mark within the next two years. The journal file may reach this point during 1974.

The ERIC Thesaurus is also growing, as new descriptors are added. Figure 2 depicts the growth of this file between 1967 and 1972. The growth of this file is leveling off at less than 100 additional main descriptor terms a year, with a total at present of approximately 5000 main terms.

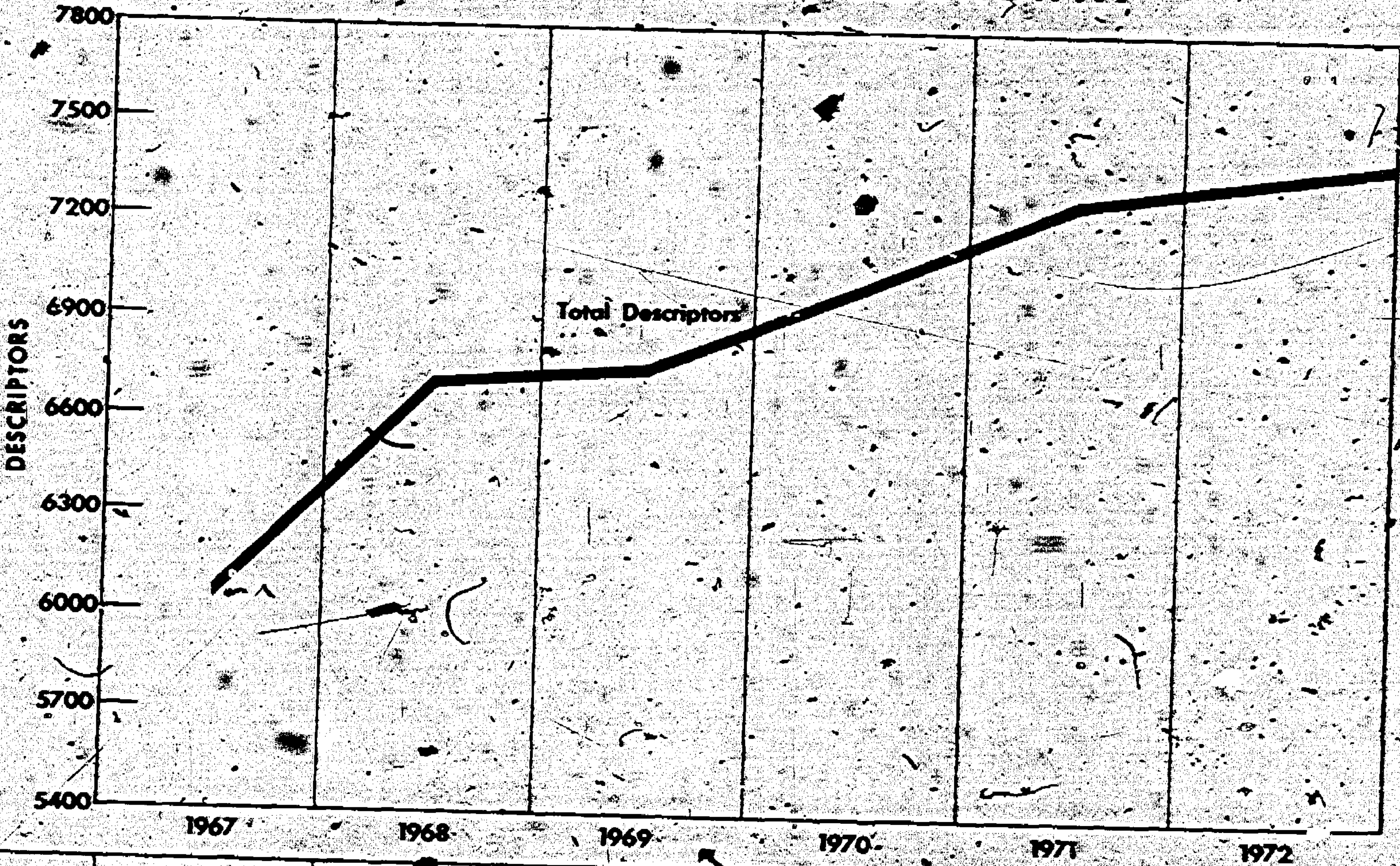
The ERIC System: An Analysis

The completeness and quality of the information cataloged in the ERIC collections may be examined through consideration of the organization of and processes involved in the ERIC system. The ERIC system has been described as having four levels (Brandhorst, 1972, p. 4):

1. ERIC central located in NIE,
2. the 16 clearinghouses located in universities, professional societies, associations, councils, etc.,
3. the commercial level, including facilities for managing the data bases, and disseminating ERIC products and documents,
4. the users who receive the benefits of these activities

Figure 2

ERIC THESAURUS-FILE GROWTH



	1967	1968	1969	1970	1971	1972
Use References	1126	1847	2140	2308	2395	2443
Main Terms	3411	4239	4562	4761	4877	4935
Total Descriptors	4537	6086	6702	7069	7272	7378

ERIC Data Base Users Interchange, February, 1973, Attachment 73-1

Analysis of the organizational structure linking these levels and of the processes within each of these levels allow us to identify critical points in the ERIC system which potentially affect the quality of ERIC output. Of particular interest in this discussion are the existing relationships between the first and second, and third and fourth levels identified by Brandhorst, as well as the processes existing within the second level. The effect of these relationships and processes on the objective of comprehensive information retrieval by educational researchers and practitioners will be analyzed. As will be seen the quality of information retrieved is highly depended upon the manual processes involved.

As indicated earlier, the ERIC system is conceptualized as a decentralized set of fairly autonomous clearinghouses whose efforts are coordinated by the central ERIC facility in NIE. The 16 clearinghouses identify with a specific educational subject or topic. As new content areas emerge, they may not fit nicely into the mission established for any one of the clearinghouses, calling for a change in contract to legitimize the acquisition of documents by the clearinghouse in the new area. Such a formal adjustment was made in the area of environmental education, as it was added to the focus of the clearinghouse on science and mathematics. A problem arises in that while the mission of the clearinghouses is changing, many important early documents may not become a part of the information collection. Also, there exist content areas which are of limited scope, but none-the-less important to some consumers of educational research, which do not fall neatly within the scope of any particular clearinghouse and are therefore not systematically included.

Major responsibilities of the various clearinghouses focus on the acquisition, screening, coding and abstracting of documents in the clearinghouse's areas of concentration; see Figure 3. A fundamental limiting influence on the entire information retrieval capability resides in the acquisition of documents to be added to the system. Each of the clearinghouses is responsible for establishing a systematic and comprehensive document collection network. Acquisitions result from requests made by the center of known researchers in the field or from the submission of unsolicited documents by researchers. In either case the comprehensiveness of clearinghouse's acquisitions is a function of the thoroughness of the collection network and the visibility and status which the clearinghouse holds in the subject or topical area.

Greenwood and Weiler (1972, p. 62) pointed out that the autonomy of clearinghouse directors has led to inconsistencies in the acquisition of certain types of documents across the ERIC network. They indicated specifically the collection of dissertations and foreign documents. They also reported the conclusion that members of the educational community, both within and outside ERIC indicate that much fugitive literature of importance to education is missing from the ERIC collection.

Upon receipt of the documents, the clearinghouses screen them for appropriateness for entry into the ERIC collection. Initial screening is for reproducability, the existence of copyrights which would preclude dissemination, and current availability of the document within the network. Next, the documents are screened for appropriateness of content relative to the field of education and more specifically, the

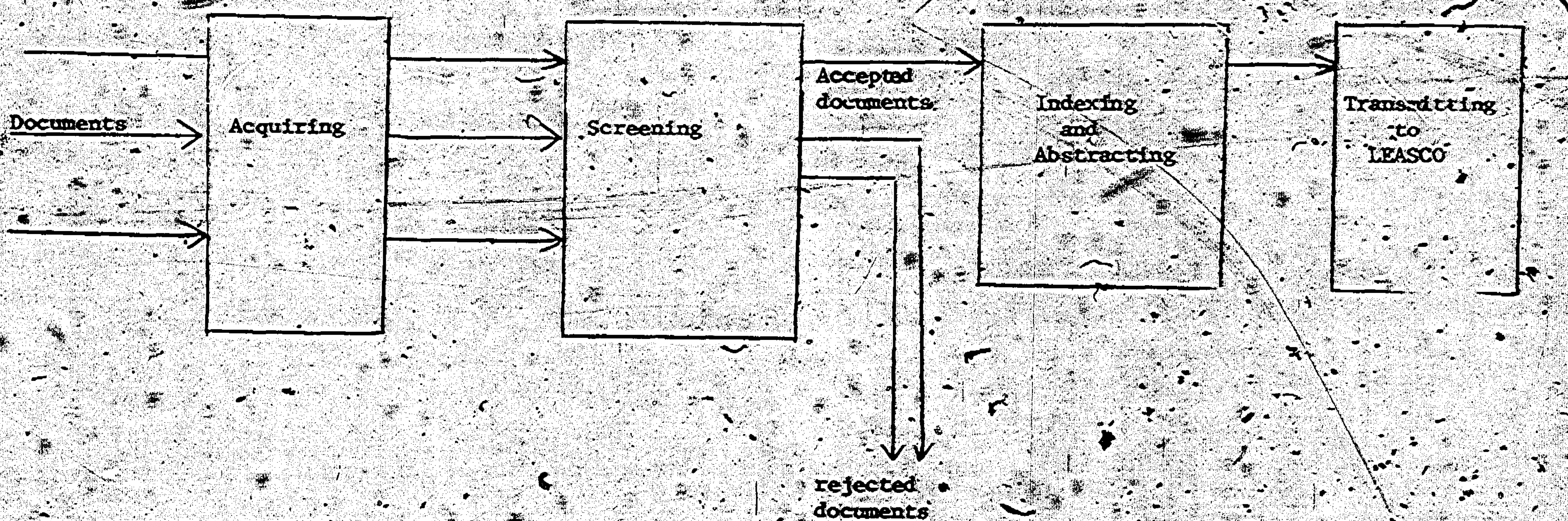


Figure 3

ERIC Clearinghouse Document Processing Functions mission of the particular clearinghouse. The screening process and the subsequent indexing and abstracting functions are clearly highly dependent upon the quality of clearinghouse staff. These functions, like that of document acquisition, place absolute limitations on the quality of later search outcomes.

In the process of document indexing, the Thesaurus of Educational Descriptors (CCM Information Corporation, 1972) is very important. This publication provides the vehicle whereby documents and articles are entered into the ERIC collection and whereby they are retrieved by the ERIC consumer groups. The Thesaurus has been criticized on the basis of it being too large, too confusing and difficult to change. (Greenwood and Weiber, 1972, p. 64). Suggestions for additions to the Thesaurus originate with the clearinghouse, but final lexicorographic control is retained by ERIC control.

Pressure, primarily from information specialists, has led to the establishment of an ERIC Vocabulary Improvement Program (VIP) which is working toward the elimination of ambiguities, and outmoded and unused indexing terms (ERIC Processing and Reference Facility, 1974, Attachment 1). This program will also consider the proliferation of identifiers, that is indexing terms not found in the Thesaurus, but which are none-the-less useful in identifying a document. An example of such proliferation and inconsistency is afforded by the classification of documents related to ESEA Title III. Seventeen different identifiers are used to index such documents. This program also, hopefully, will deal with the development and elaboration of non-subject descriptors to facilitate classification and retrieval. Two important descriptors of this type are level (.e.g.

secondary, elementary, higher education) and document type (e.g. resource lists, evaluation report, research). While these terms are used at present, they are not used consistently (Hull and Wagner, 1972).

Another problem noted in the use of the Thesaurus is that descriptor terms are not used consistently across clearinghouses. This is particularly true of keywords which are specific to the concerns of one clearinghouse and only peripheral to the concerns of other clearinghouses (Greenwood and Weiler, 1972, p. 12).

The problems of differences in indexing and abstracting procedures existing between clearinghouses led Fry (1972), in his Evaluation Study of ERIC Products and Services, to recommend that the indexing and abstracting tasks should be centralized (p. 2-21). Clearinghouses would still play the primary role in document acquisition and perhaps an increased role in synthesis and dissemination under this recommendation. This suggestion might result in increased efficiency leading to a reduction in total time from document capture to its ultimate appearance in the ERIC collection.

An important link in the ERIC systems which is too frequently ignored or at least under considered, is the link between the formal ERIC network, the acquisition, processing and reporting functions, and the consumer, the fourth level of the Brandhorst model. This link exists in the form of the information search specialist, a person who can translate the needs of the research consumer into the exact lexicographic and boolean syntax of the search formula. This person's knowledge of and experience with the Thesaurus is critical to the conduct of a comprehensive search. It is tempting to speculate that the most severe limitation to the comprehensiveness of the ERIC search is that imposed by the untrained consumer conducting the actual search. Knowing how to use the Thesaurus to organize a search strategy is a rare but growing specialty. Dissemination and training programs such as the Information Retrieval Demonstration and Research Project sponsored by the Center for Studies in Vocational and Technical Education at the University of Wisconsin(Lambert, 1974), allow for information specialists such as Roy Tally to train additional specialists who, in turn, bring the ERIC data base into direct contact with the educational consumer. In this case through the media specialists located in Vocational, Technical and Adult Education Institutions in the State of Wisconsin.

While most of the foregoing comments have focused on potential and actual limitations of the ERIC network, it is only fair to mention that the several recent major evaluations of the ERIC system have established the general level of utility of and consumer satisfaction with the ERIC system (Fry, 1972; Greenwood and Weiler, 1972; Hall and Wagner, 1972; and Wagner, 1972), Such studies typify the potential impact that indepth analysis and evaluation can have on the improvement of the ERIC system. Additional studies of ERIC processes and products are to be encouraged, especially those which involve detailed consideration of research generators' and research consumers' assessements of the completeness, utility and timeliness of the ERIC data bases. This information can provide direct feedback to clearinghouses and ERIC central and lead to improved acquisition and processing procedures.

References

Brandhorst, W. T., MANAGING THE ERIC DATA BASE. A paper presented at the Fall Joint Computer Conference. Anaheim, California, 1972 (ED069303).

Burchinal, L. G., "The Educational Resources Information Center: An Emergent National System." JOURNAL OF EDUCATIONAL DATA PROCESSING, 7, 2, 1970, pp. 55-67.

CCM Information Corporation, THESAURUS OF ERIC DESCRIPTORS. New York: CCM Information Corporation, 1972.

ERIC Processing and Reference Facility, "ERIC Vocabulary Improvement Program." ERIC DATA BASE USERS INTERCHANGE, March, 1974, Attachment #1, pp. 1-2.

Fry, B. M., EVALUATION STUDY OF ERIC PRODUCTS AND SERVICES. Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1972, Volumes 1-4 (ED060923).

Greenwood, P. W. and Weiler, D. M., ALTERNATIVE MODELS FOR THE ERIC CLEARINGHOUSE NETWORK. Santa Monica, California: Rand Corporation, 1972 (ED058509).

Hull, C. C. and Wagner, J., Educational Resources Information Center (ERIC) File Partition Study: Final Report. Santa Monica, California: System Development Corporation, 1972 (ED067520).

Lambert, R. H., INFORMATION DEMONSTRATION AND RESEARCH PROJECT, Proposal. The Center For Studies in Vocational and Technical Education, University of Wisconsin, Madison, 1974.

U.S. Department of Health Education and Welfare, GUIDELINES FOR THE DEVELOPMENT OF A THESAURUS OF EDUCATIONAL TERMS. Washington, D.C.: Office of Education, 1966.

Wagner, J., EVALUATION STUDY OF NEC INFORMATION ANALYSIS PRODUCTS: FINAL REPORT. Santa Monica, California: System Development Corporation, 1972.

File Search Strategies for the WISE-ONE Information Retrieval System

by

Tom Olson
Madison Academic Computing Center
University of Wisconsin-Madison

WISE-ONE is designed to provide fast, and efficient access to computer-based information files. It was designed specifically to meet the needs of researchers requiring access to the ERIC system, but the logic is sufficiently general to accommodate any of the many computer-based library systems. In order to provide for the on-line computer access capability, it is necessary to restructure the data-base. In paper a variety of alternatives are discussed and the WISE-ONE system and logic is explained.

The Binary Tree

One possible data structure is a binary tree which is often employed for processing natural language and computer compilers. The binary tree relies upon a carefully contrived data base where entry is always at a common point followed by selective branching until the correct key is found. The easiest way to conceptualize the tree structure is to trace the creation of the data base because the search and creation logic are identical. Let's start with a coded word. It is stored with the associated data record awaiting comparison with the second keyword. A numeric compare with the second keyword will produce a negative, positive or zero result. When the result is equality, the data record is expanded with the text associated with the new keyword. A negative result will cause a left node pointer to be selected from an available space list. The keyword and associated data record is then stored in the location specified by that pointer. Similarly, a right-node position is selected when the compare result is positive. Thus, two pointers may be identified with each node or keyword in the data structure. As each new keyword and data record is entered, it follows the logical path of left or right node pointers until an empty node is encountered. The data record and associated pointer is then added to the tree. See Figure 1. When a search is desired, the search follows a path through the node pointers until a compare on the keyword produces a zero result. The data record associated with that node is the desired information. The path is dictated by plus or minus results for a compare between the search keyword and the node keyword.

Several advantages and disadvantages accrue from this method.

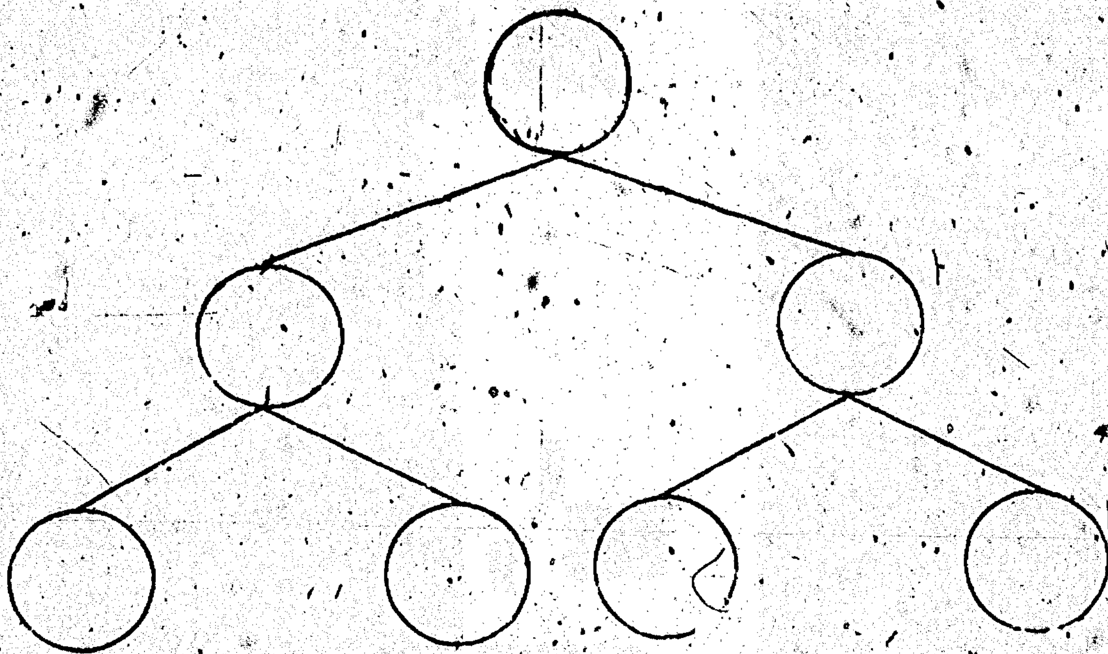


FIGURE 1

Illustration of a Binary Tree

Dynamic creation of the data base is possible. Updates to the system are quite simple. Deletions are not easily accomodated as they interrupt the flow of the tree. A balanced tree, while not essential, provides a more efficient search. The logic for balancing a tree structure is extremely complex. Search trees are relatively short but increase as a log function of data size.

The major drawback to this structure occurs when the structure is being updated or corrected. To keep the search times optional, it is important that the tree be fully balanced, such that all the nodes on one level are filled before the nodes on the next level are used. If this balancing is not done, the tree becomes so unbalanced that search time is significantly degenerated. Algorithms to balance the tree structure exist but they are inefficient when working on mass storeage because they require a large number of I/O requests. Another problem is the high overhead associated with placing new information into the file. For example, to insert a new node on a fully balanced tree of 16,000 nodes requires 16 I/O requests before the proper parent is found. This overhead becomes substantial when a large number of new nodes are to be added or the existing tree is unbalanced. In addition, a large amount of space is consumed by the linkage information when a large number of nodes are stored in the tree. The problems associated with the tree structure indicate it is only a highly desirable procedure when used with small or static data bases.

Index Sequential Files

A second approach is commonly called index sequential. In this frequently employed method, the file is sequentially ordered and an index of references is developed from the file. Thus, a search may be limited to the index, locating the approximate search entry point into the main file. An index sequential approach cuts the search-time significantly from a sequential search by locating specific search entry points and eliminating the need to examine each record. However, large files require large indexes and the method only delays the need to consider more efficient procedures. It is simple to employ and therefore is highly desirable when operating with relatively small files or when the search response time is not critical. Updates are relatively simple because of the sequential order of the file. Update information may be merged into the sequential file. This does require that the file be recopied, which may be a costly method when the file is very large. Many variations have been developed in the interest of optimizing the update procedures of index sequential files. However, for the application to an on-line inquiry to a large bibliographic data base, the search time is generally not sufficiently responsive. For this reason, the WISE-ONE staff settled on a hash coding scheme for citation identification.

Hash Coding for Data Base Entry

The hashing method may be employed as a variation of the Index Sequential Approach in which the index is hashed using a mathematical permutation of the keys to determine the approximate location of the citation in the file. This method is efficient but may lead to slow response time when employed in an interactive mode.

The structure of the WISE-ONE data-base is a linked table scheme and is an adaptation of a direct chain, hashing scheme employing a linked list structure.* There are three types of tables developed in this process, a base-table, collision tables and citation tables. The heart of the system is the hash coding scheme which is incorporated into the data-base structure. A hash code is a method of computing the storage location of a record based on some mathematical permutation of the search key. The hash

*The ERIC search program - WISE-ONE - was funded by the School of Education, Department of Educational Administration, Wisconsin Information Systems for Education (WISE). Mr. S. C. Yang and Professor Venezky contributed to the development of the hashing scheme. The program was also a class project in Computer Science - CS 638 taught by Professor Travis. These contributions are acknowledged and appreciated.

algorithm WISE-ONE employs generates two numbers: the hash address and the virtual key or residue. These two number correspond to the remainder and quotient of the division of the keyword bit pattern by the size of the base table. The role of the hashing scheme and the collision tables in the structure of the data-base is best explained by tracing the search process. See Figure 2.

When a search key is entered, it is hashed by multiplying successive s... of six character computer words until the entire keyword is stored as a 72 bit product. The middle 13 bits are selected as the hash address for entry into the collision table. This hashing approach is an approximation of the middle squares approach. The method produces a random bit pattern, therefore reducing the probability of collisions. It is obvious that at some time two or more keywords may result in the same hash address. The collision table is designed to resolve these conflicts. When the hash address is computed, the surrounding 36 bits are selected as the residue.

The collision tables are too large to store in memory. Therefore, we need a mechanism to convert the hash code to a mass storeage address for the collision table. The hash address is used to point into a base table. The base table is a core resident list which contains the address of all collision tables which reside on secondary storeage.

The collision table is entered at the hash address and the stored residues are compared. The residue is stored as a psuedo keyword in the interest of storeage efficiency. An equality compare on the residue associated with a given hash address points to a citation table in which all references to the given key word are stored. These references constitute a search queue which may interact through boolean operators with a prior search queue to produce a resultant list. The process is repeated with additional keywords until the search logic is completed.

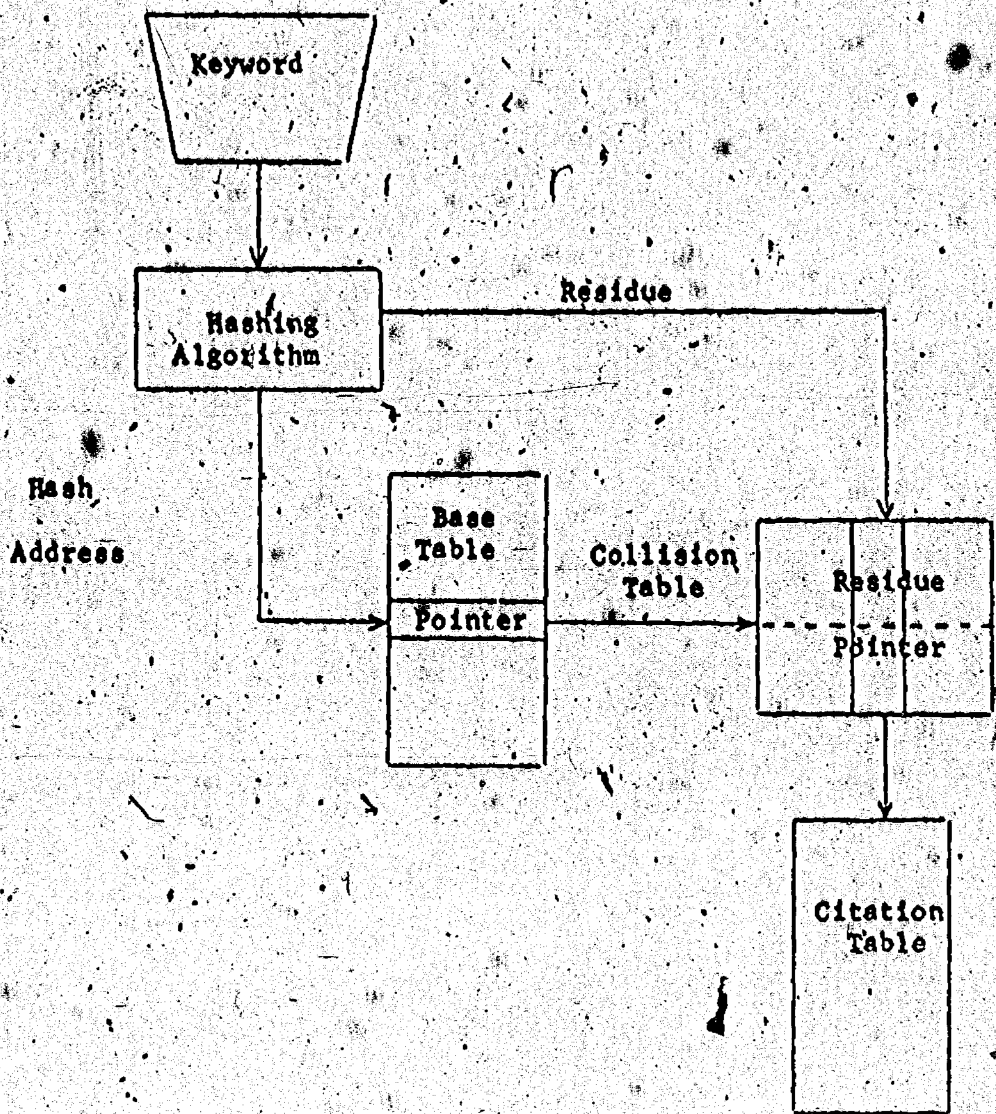


FIGURE 2

Hash Coding Scheme

When the constructed search queue is complete, the citation numbers are hashed in the same manner as the keywords. The hash address is used as a pointer to another base-table to obtain a link to a collision table on secondary storeage. The collision table is then searched for the matching virtual key or residue and the associated link is then followed to obtain the title, author and journal citation of the ERIC number. This process is repeated for each citation number in the search queue until all citations have been printed.

Boolean Operators

It is convenient to think of three types of queues. The secondary queue is the result of a keyword search. It contains a list of the citations associated with a given keyword. The primary queue is a prior list of citations which interact with the secondary queue through a boolean operator. It may be empty. The resultant queue is that which is produced by the interaction of the primary and secondary queues through a boolean operator.

The boolean operators for WISE-ONE include AND, OR, and NAND. The AND operator generates the resultant queue which contains citations common to both the secondary and primary queues.

Primary Queue		Secondary Queue		Resultant Queue
Cit. 1		Cit. 2		Cit. 6
Cit. 3	.AND.	Cit. 6	=	Cit. 9
Cit. 6		Cit. 8		
Cit. 9				

The OR operator generates a resultant queue which contains citations unique to both the primary and secondary queue. The OR operator is therefore additive in nature.

Primary Queue		Secondary Queue		Resultant Queue
Cit. 1		Cit. 2		Cit. 1
Cit. 3	.OR.	Cit. 3	=	Cit. 2
		Cit. 4		Cit. 3
				Cit. 4

The NAND operator is a BUTNOT operator which reduces the primary queue by all matches within the secondary queue. It is helpful for systematic reduction of the primary queue.

Primary Queue		Secondary Queue		Resultant Queue
Cit. 1		Cit. 1		
Cit. 2		Cit. 3		Cit. 2
Cit. 3	.NAND	Cit. 5	=	Cit. 4
Cit. 4				
Cit. 5				

The boolean operators permit the dynamic construction of sear formulas. Each keyword entry will involve the hash algorithm and wi produce a resultant queue as prescribed by the selected operator. In order to optimize the building of search formulas, it is useful to employ parenthetic logic.

(COLLEGES.OR.UNIVERSITIES.OR.HIGHER EDUCATION).AND.(FISCAL SUPPORT.OR.FINANCE)

This approach expands the utility of searching large and complex data-bases. The nature of the ERIC files requires that such a capability be available.

Updating the File

The creation and update of the data-base follow a different line of development than the search process. The keywords in the form of descriptors, identifiers and author's last names are abstracted from the ERIC tapes along with the title, author and date of the citation. Each keyword is hashed and the hash address residue and keyword are written into a file along with the ERIC citation number. The title, author and citation numbers are written into another file. The keyword file is then sorted on citation number within residue within hash address. This file is then merged with the existing master file to create a new master file. The master file contains all the information in the proper order for easy generation of the table structure. The data-base search files are then generated from master file and the title and author file. See Figure 3.

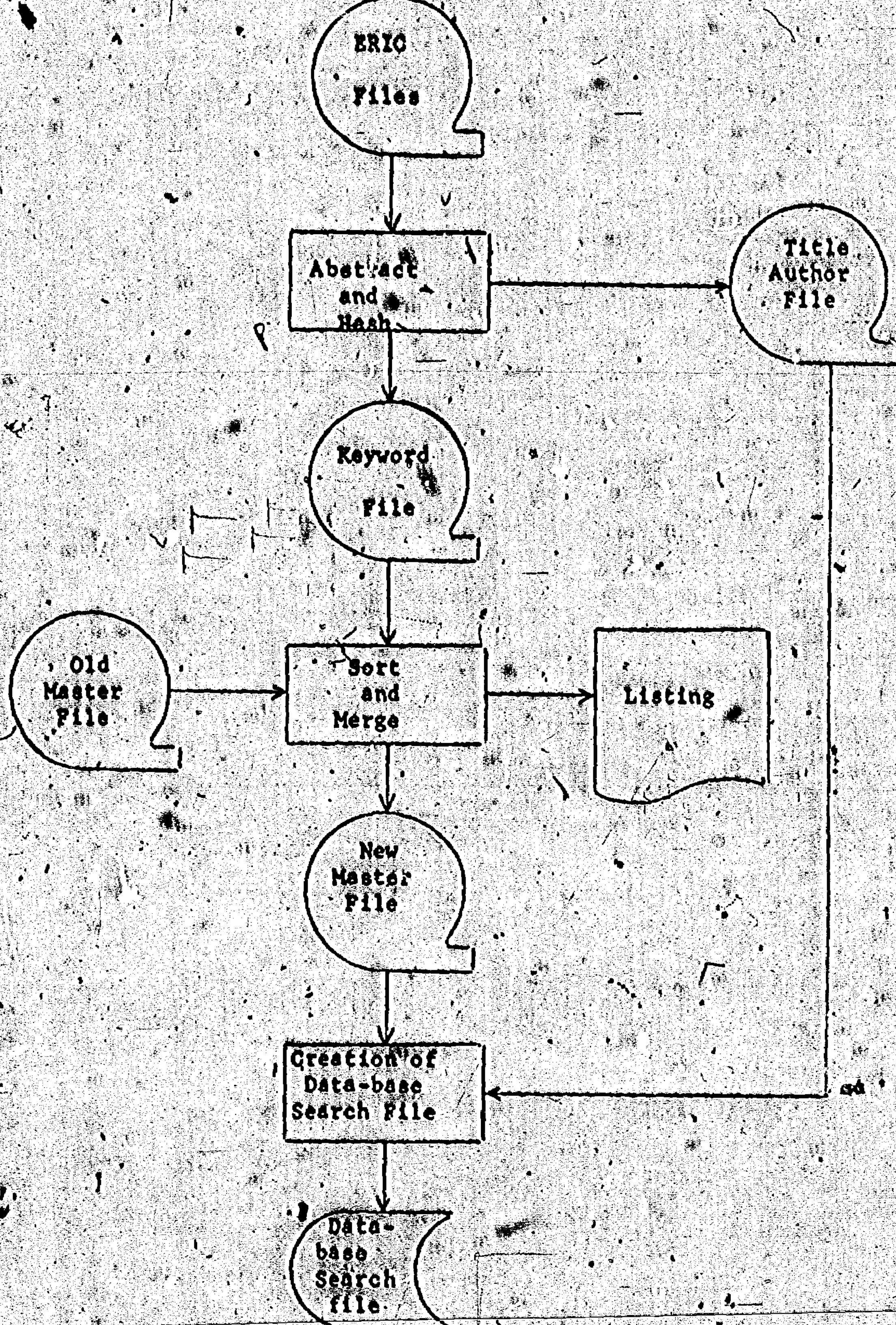
There are a number of advantages to this method of storeage and retrieval, the most notable being its extremely fast search time. The CPU time on the Univac 1108 per keyword is in the order of hundreths of seconds. The overall search time is less than a tenth of a second per keyword.

Another important feature of this search method is that search time will not increase significantly as the data-base grows in size. This is because the number of probes to the disk to search for any keyword is two, one to read the collision table and one to read the citation table. The only portions of search-time that will increase are those associated with the collision table search for the residue and the time required for boolean process of the longer lists.

WISE-ONE currently runs on the Univac 1108 at the computing center on the University of Wisconsin-Madison campus. It is written in 1108 assembler and Fortran V. It uses about 31k 36 bit words of core storeage and about 1500k words of disk storeage for each file. (For IBM types, this can be translated to 124k bytes of memory and 6 megabytes of mass storeage.) The nature of the hashing scheme forces the code to be machine dependent and it would require considerable reprogramming to run this system on computers other than UNIVAC 1100 series machines.

FIGURE 3

File Generation and Update Procedures



WRITING SEARCH STRATEGIES FOR EFFECTIVE COMPUTERIZED LITERATURE SEARCHES ON THE ERIC DATA BASE

Roy Tally
Supervisor - Information Retrieval Center
Wisconsin Department of Public Instruction

Introduction

A variety of computer based information retrieval systems are now available in education, medicine, psychology and a host of hard sciences. Searches of these data bases offer the researcher or practitioner powerful tools for rapid screening of vast quantities of material. New expertise is required to make efficient use of these resources.

This paper discusses techniques for preparing a query for the ERIC data base. The WISE*ONE (1) search system developed at the University of Wisconsin-Madison is used in the search examples. Topics include: 1) writing the initial information need statement; 2) identification and structuring of major concepts in the statement; 3) using the Thesaurus of ERIC Descriptors (2), and; 4) grouping terms in the search strategy using logical connectives OR, AND, NAND (NOT) and auxiliary commands OPEN and CLOSE.

The Search Statement

The most important factor in obtaining relevant information through computer searches is adequate communication of the request parameters. Those who use search services through libraries or center will most often be required to fill out a request form. The user is asked to qualify his request with any constraints such as date or total number of references desired. The most important portion of the form is that which asks for a statement of the information need. One could enter a cryptic note like "...all of the references on pollution and environment." A more helpful statement would be:

> The high school is developing a program in environmental quality. One aspect of the program will be an investigation of pollution effects in our local urban area. Students will develop projects under teacher guidance. Materials on environmental quality, pollution, or pollution effects would be helpful.

In the above statement, factors of background, scope and purpose and ultimate use of the information suggest many additional search keys to experienced searchers. The practice of writing complete sentences is most helpful. Even more detail could be included with the request.

Attachments such as a key article or reference or an abstract of a research proposal are invaluable.

If you do your own search, it is advisable to write out the search statement in detail. Once you are working with a list of key words or on a terminal, it is easy to become confused. Having the search statement at hand helps to maintain a focus for selection of terms and judgments on document relevance.

Structuring the Search Statement

A useful technique for selecting search terms involves breaking down a statement into a concept term matrix. Identify the significant words in the statement representing concepts and rewrite them into a horizontal array. Then, consult the Thesaurus for each concept and record related or narrower terms in a vertical array beneath each concept.

	CONCEPTS: environmental pollution	CONCEPTS: cities/urban areas
TERMS	pollution	city problems
	air pollution control	city improvement
	water pollution control	city planning
	ecological factors	urban areas
		urban environment

This example will be referred to again at the end of the section on logical operators.

Using the Thesaurus

Descriptor terms are listed in the Thesaurus in a variety of formats. The bulk of the Thesaurus is the hierarchical listing. All terms are listed alphabetically in bold face type. Beneath each term is a hierarchy of terms in lighter type. Abbreviations preceed each section of the hierarchies and are interpreted as follows:

UF - used for; the bold face term is preferred usage. The UF term appears elsewhere in the Thesaurus in its alphabetical order in a smaller bold face type. It is followed by a "use" reference.

BT - broader term; this term is more general than the main term. It appears elsewhere in alphabetical order, in bold face type, and is followed by its own hierarchy.

NT - narrower term; there can be more than one term here. Narrower terms also appear elsewhere followed by a similar hierarchy.

RT - related term; these terms may be considered as being on the same level with the main term. They are not necessarily synonyms. Again, they appear elsewhere as main terms.

SN - scope note; this is a short note on the use of the term rather than a definition.

The user can confidently enter the Thesaurus at any point and, using the term hierarchies, chain through the listings for all of the available descriptor terms of interest.

A second helpful listing is found near the end of the Thesaurus. It is called the Rotated Descriptor Display. Here, terms are listed by every component word. For example:

POLITICS
POLLUTION
AIR POLLUTION CONTROL
WATER POLLUTION CONTROL
POLYGRAPHS

The Thesaurus also groups descriptors under fifty-two general headings. These listings appear in the Descriptor Group Display. Each group is assigned a three digit number. The number is posted to the bold face entry for each term in the Hierarchical Listing. Review of the group listings can suggest additional related descriptors.

Logical Operators

The WISE-ONE system employs Boolean algebraic functions to process information search strategies. Command words controlling the functions are AND, OR, and NAND.

The results of these commands can best be illustrated with a short explanation of basic set theory which is analogous to the way the computer processes search terms. If one enters the term POLLUTION: (computer response-lower case; user response-upper case. Numbers retrieved are illustrations only and are not accurate search results.)

proceed

→ POLLUTION

250 documents in data base

250 documents in search queue at level Ø

→ END

a set would be constructed of all reports that have POLLUTION as a key word.

The computer response gives the number of items retrieved for the term and begins to accumulate results in a holding area called the search queue.

If one adds to the logic as follows:

proceed

→POLLUTION

250 documents in data base

250 documents in search queue at level Ø

→AND

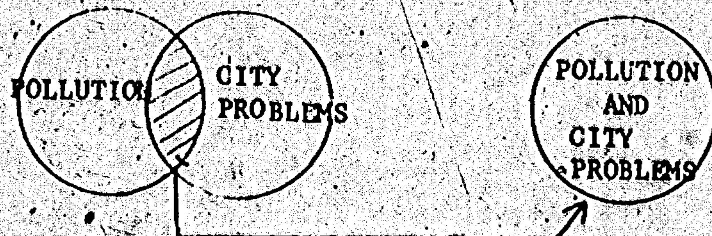
→CITY PROBLEMS

387 documents in data base

15 documents in search queue at level Ø

→END

the set for POLLUTION would be constructed as before and a second set constructed of items which have CITY PROBLEMS as a key word. The intersection of these two sets is the final results of the logic processing, and each of the items in this set has both POLLUTION and CITY PROBLEMS as key words.



The result of the intersection is contained in the search queue following the last term.

If the logic is changed as follows: (computer responses are deleted for clarity).

proceed

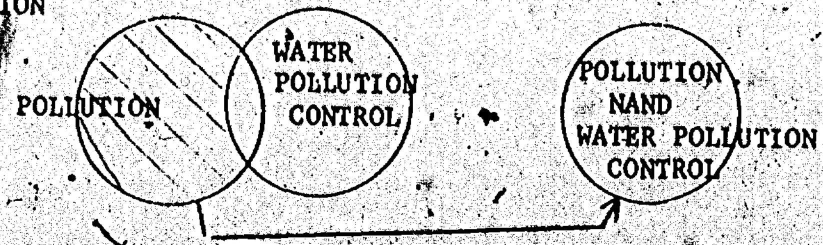
POLLUTION

NAND

WATER POLLUTION CONTROL

END

the set WATER POLLUTION CONTROL is deleted from the set POLLUTION



Using both the AND and NAND functions together one could write the following logic.

proceed

POLLUTION

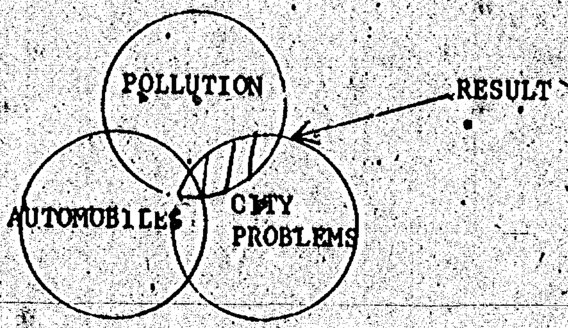
AND

CITY PROBLEMS

NAND

AUTOMOBILES

END



We have taken the intersection of POLLUTION and CITY PROBLEMS and deleted from it the intersection with the set of report numbers that have AUTOMOBILES as a key word.

To this point the discussion includes single terms separated by command words. In application, it is necessary to select related descriptor terms and combine their search results into a groups for further logic manipulations. Using the example of POLLUTION, one adds the search term AIR POLLUTION CONTROL to the set. This is accomplished with the OR command:

proceed

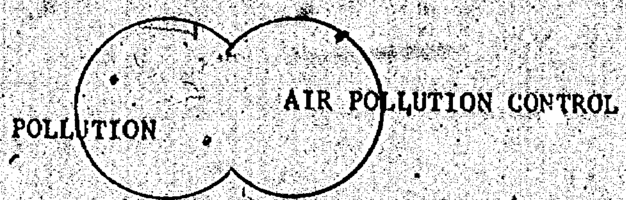
POLLUTION

OR

AIR POLLUTION CONTROL

END

A new set is produced which contains either the term POLLUTION or AIR POLLUTION CONTROL.



The function of combining related terms can be extended to any number desired so long as each term is followed by the OR operator.

By analogy, we may wish to extend the second concept in the sample, CITY PROBLEMS. NOTE: The auxiliary commands of OPEN and CLOSE are required for this operation and fit into the sample logic as follows:

proceed

POLLUTION

AND

OPEN

CITY PROBLEMS

OR

CITY IMPROVEMENT

CLOSE

END

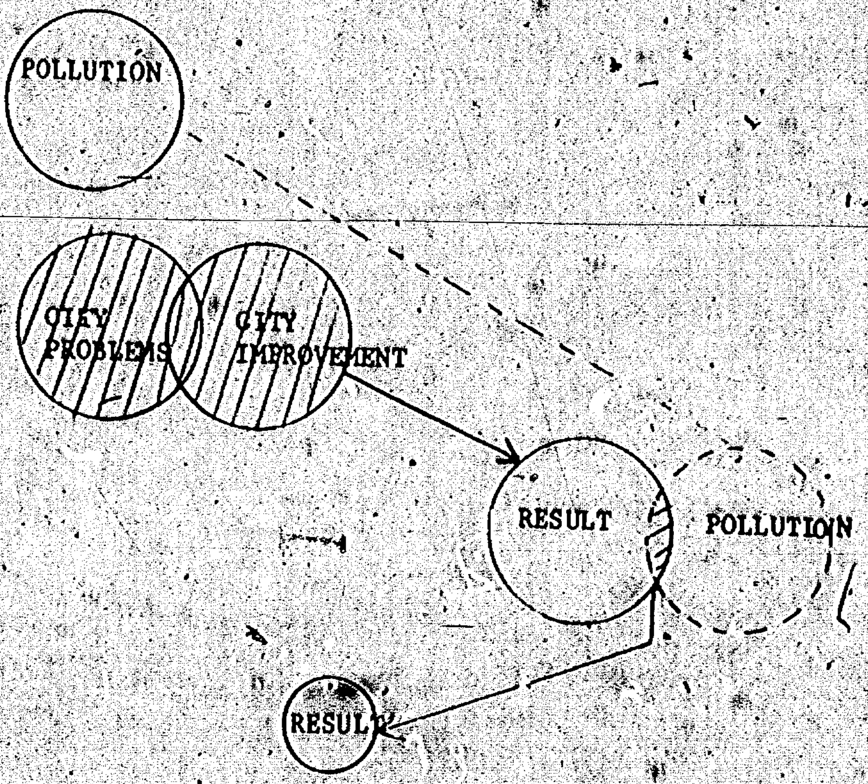
The OPEN command has the effect of isolating the first set POLLUTION while the set CITY PROBLEMS or CITY IMPROVEMENT is accumulated. The command CLOSE resolves the preceding combination of sets. In effect, the OPEN and CLOSE commands are equivalent to parentheses. The term for such an enclosed set is "nested" set. Using diagrams:

STEP 1

STEP 2, 3

STEP 4

RESULT



Returning to the concept term matrix, logical operators can be inserted by a simple convention. All terms in a vertical array must be connected by OR operators while terms or groups in the horizontal array must be connected by AND or NAND operators.

Concepts

environmental pollution AND	cities/urban areas
. pollution	. OPEN
. OR	. city problems
. air pollution control	. OR
. OR	. city improvement
. water pollution control	. OR
. OR	. city planning
. ecological factors	. OR
	. urban areas
	. OR
	. urban environment
	. CLOSE

The OPEN and CLOSE operators allow the user to accumulate terms in the second concept. The AND operator takes effect after the close statement.

Conclusion

This paper offers a fundamental review of the techniques for search preparation. The examples are admittedly limited. The discussion of logic manipulation is only barely introduced. In the future, searchers can expect to see more detailed discussions on strategy building, but it is felt that the fundamental problem of question refinement will remain the major determinant in what constitutes a successful search.

References

1. Olson, Tom and Others; USER DOCUMENTATION, WISE*ONE; Madison Academic Computing Center, 1210 West Dayton St., Madison, WI 53706

2. _____________, THESAURUS OF ERIC DESCRIPTORS; Macmillan Information Corporation, New York, 1972.