DOCUMENT RESUME

ED 061 979 24 LI 003 649

AUTHOR Bracken, Paula
TITLE OTIS Basic Index Access System (OBIAS); A System for Retrieval of Information From the ERIC and CIJE Data Bases Utilizing a Direct Access Inverted Index of Descriptors and a Reformatted Direct Access ERIC-CIJE File.
INSTITUTION Oregon Total Information System, Eugene.
SPONS AGENCY National Center for Educational Communication (DHEW/OE), Washington, D.C.
REPORT NO Project-0-0764
PUB DATE Jan 72
NOTE 18p.;(0 References)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Computer Programs; *Data Bases; Dial Access Information Systems; Electronic Data Processing; *Information Retrieval; *Information Systems; *Search Strategies
IDENTIFIERS CIJE Data Base; *ERIC Data Base; OBIAS; OTIS Basic Index Access System

ABSTRACT

The OTIS Basic Index Access System (OBIAS) for searching the ERIC data base is described. This system offers two advantages over the previous system. First, search time has been halved, reducing the cost per search to an estimated $10 on a batch basis. Second, the "OTIS ERIC Descripter Catalog" which contains all descriptors used in the ERIC records is updated each quarter. This frequency of updating guarantees a high level of descriptor current awareness. The OBIAS search system is limited to searching against descriptors, published and unpublished, in the current file, thus, history file searches (1968 or before) must be searched with the earlier system. The OBIAS description includes: an overview of the system, system procedures, program documentation, machine configuration, comments and alternatives to the basic system, an example of the search card for OBIAS and statistics of run times. (Author/NH)

OTIS BASIC INDEX ACCESS SYSTEM (OBIAS)

A system for retrieval of information from the ERIC and CIJE data bases utilizing a direct access inverted index of descriptors and a reformatted direct access ERIC-CIJE file.

Developed under contract to: Retrieval Dissemination Project

George Katagiri, Director

Paula Bracken, Analyst/Programmer
OREGON TOTAL INFORMATION SYSTEM
354 East 40th Street
Eugene, Oregon 97405
503/342-5361

January 1972

CONTENTS

INTRODUCTION

by

Benjamin L. Jones
OTIS Automated Library System (OALS) Supervisor

The ERIC search capability at OTIS began processing searches in the last quarter of 1970.

The original system as it was installed revealed a search cost slightly in excess of $55 per search. Experience with this system uncovered ways to cut search time. System modifications then reduced search costs by two thirds, or to between $15 - $20 per search on a batch basis of 10 to 15 searches.

Under the direction of the Retrieval Dessemination Project, directed by George Katagiri, OTIS programmer/analyst Paula Bracken developed the ERIC search system, OBIAS, described on the followin pages. This system offers two advantages over our previous systems.

First, search time has been halved. At this point in time per search cost figures are not firm, but our expectation is that per search cost should not exceed $10 on a batch basis.

Second, an updated OTIS ERIC Descriptor Catalog is created each quarter upon receipt of the update tapes. The catalog contains all descriptors used in the ERIC records. The quarterly issuance of this catalog guarantees a high level of descriptor current awareness. The descriptors appear in alphabetic sequence showing frequency quotients in both the current file (all documents published in 1969 or later) and history file (all documents published in 1968 or before). A four digit identifying number also appears with each descriptor. It is this number, and not the descriptor itself, that is input for a search.

A distinction between OBIAS and our earlier systems should be made. Our earlier systems were multipurpose in nature. That is, they could search against nearly all fields within the ERIC records. These systems are similar to the systems used to search ERIC tapes across the nation. Our newest search system can only search against descriptors, published and unpublished, in the current file. History file searches (1968 or before) must be searched with the earlier system.

Questions concerning this system and its availability may be directed to Benjamin L. Jones at OTIS.

OVERVIEW OF THE SYSTEM

Two sets of programs are the basic elements of OBIAS.

The first set builds and updates these three files:

1. A direct access inverted index of the ERIC and CIJE descriptors with keys, ED and EJ numbers, to the data file.

2. A direct access file of reformatted "current" ERIC and CIJE records.

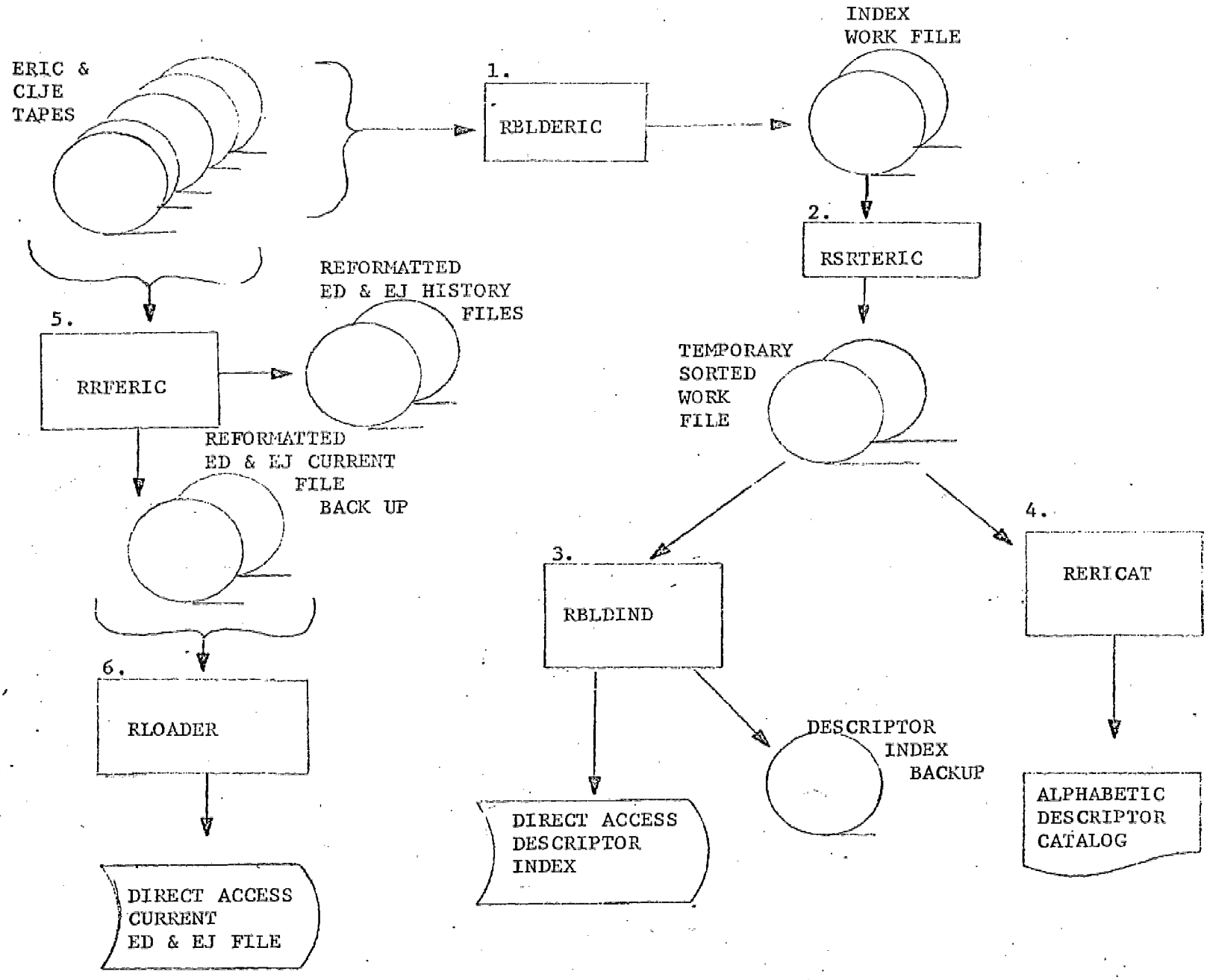
3. A sequential file of reformatted "historical" ERIC and CIJE records that can be accessed by the existing modified QUERY program.

An alphabetized listing of all descriptors and frequency counts of their appearances in the current and history files is produced by one program in the first set.

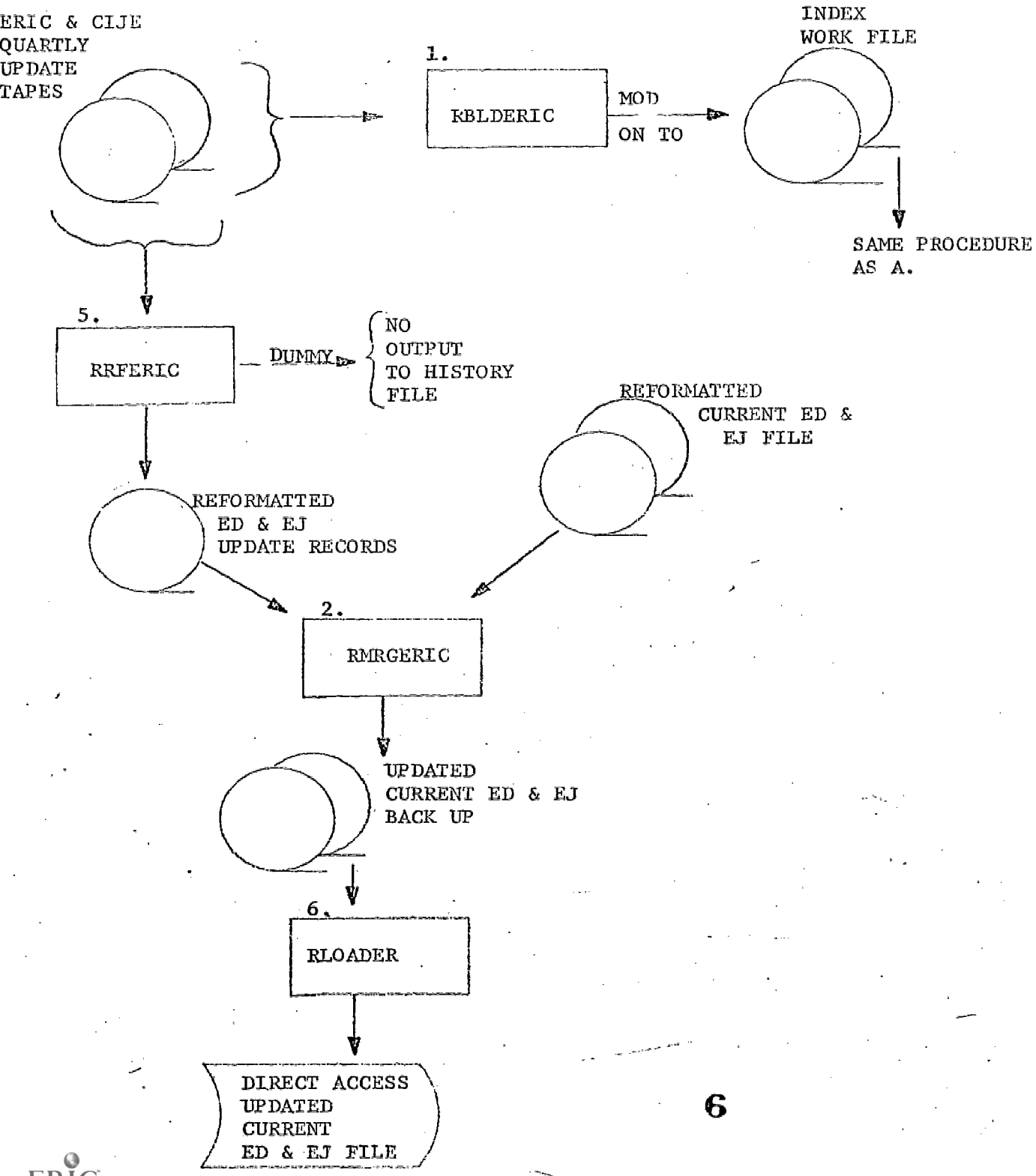
The second set of programs processes searches of the inverted index from descriptors and logical operators input as punched cards, then selects the ED and EJ records determined by the processing and writes them to a print file. The print file is printed by the existing QUERY print program.

SYSTEM PROCEDURES

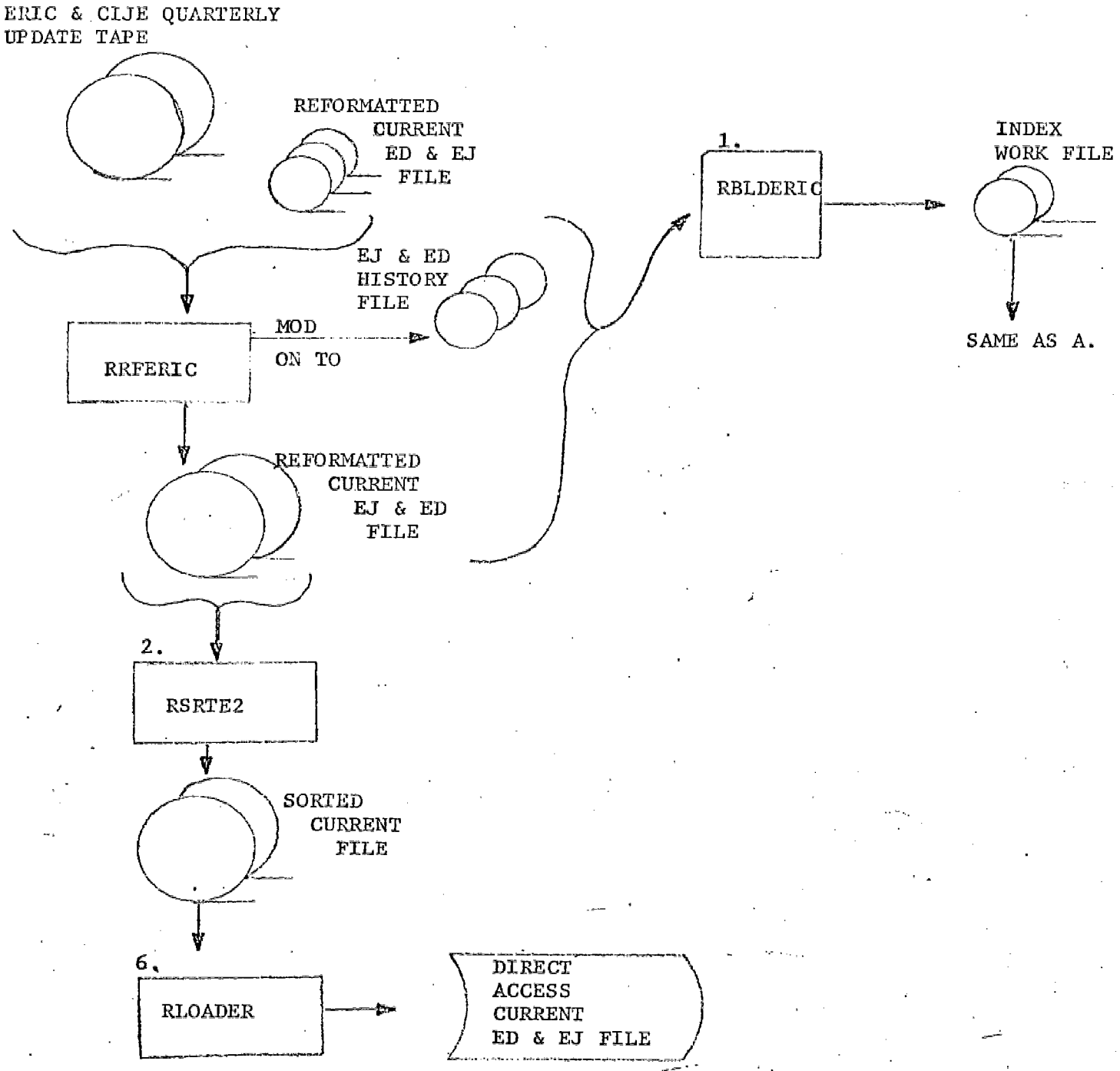
I. FILE BUILDING PROCEDURES
   A. BUILD ORIGINAL FILES



D. QUARTERLY UPDATE OF FILES



C. ANNUAL HISTORY UPDATE



D. FILE BUILDING AND MAINTENANCE PROGRAMS

1. RBLDERIC - Scans ERIC and/or CIJE records and produces one record for each descriptor in each input record including the descriptor, the ED or EJ number and a currency flag. The asterisk prefix of major descriptors is eliminated.

2. RSRTERIC - Sorts and Merges - IBM utilities
   RMRGERIC - Utility program IERRCOØØ
   RSRTE2 - RSRTERIC: Sort fields - Descriptor (major) and ED or EJ number (minor), ascending order
   RMRGERIC: Merge field - ED or EJ number, ascending order.
   RSRTE2: Sort field - ED or EJ number, ascending order

3. RBLDIND - Builds a direct access inverted index file and a sequential back up file.

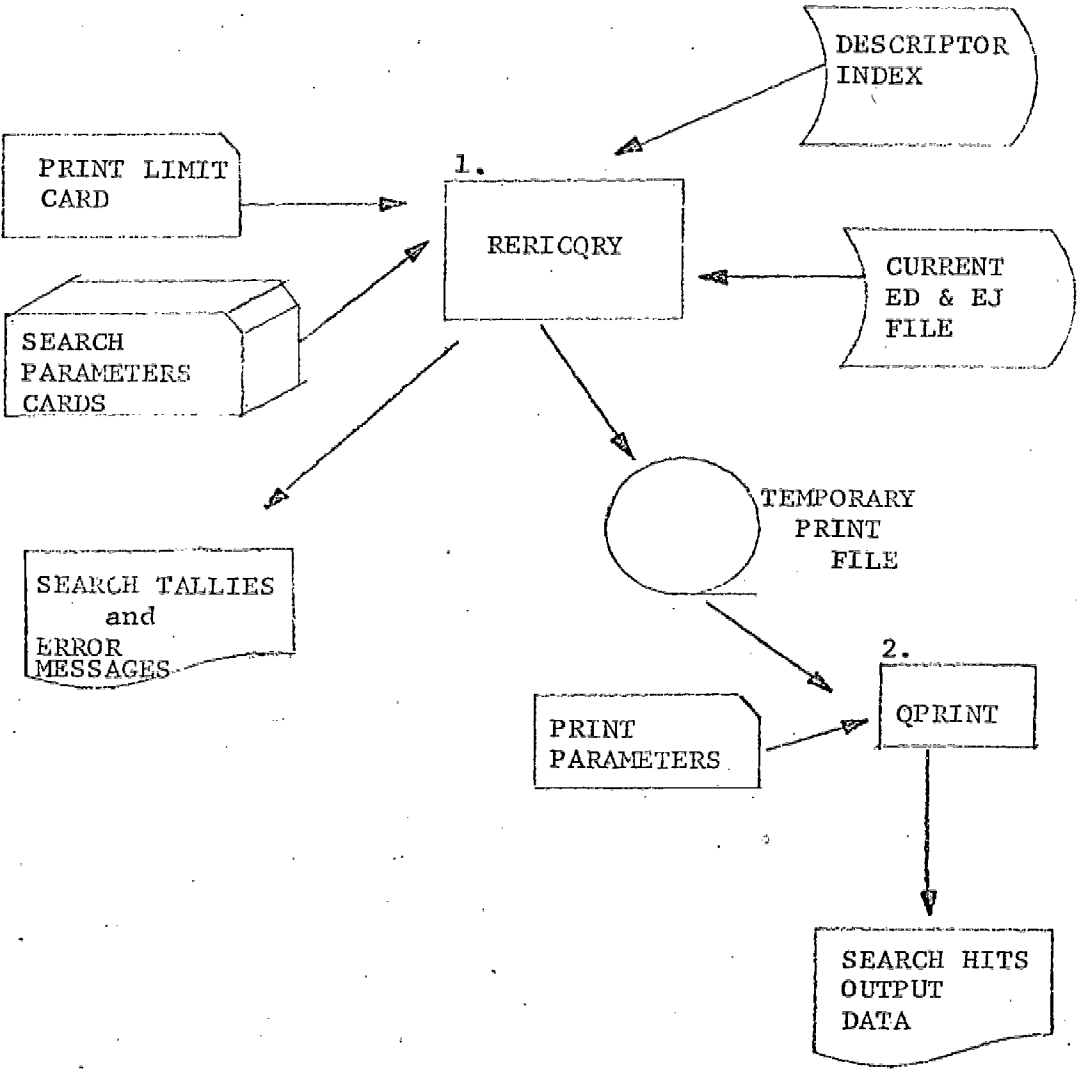
4. RERICAT - Prints an alphabetized listing of the descriptors with frequency counts of the appearance of each descriptor in the current file and in the history file.

5. RRFERIC - Reformats ERIC and CIJE tape records by eliminating fields deemed unnecessary for the projected uses of the search output, and divides the reformatted records into two sequential files, a historical file and a current file.

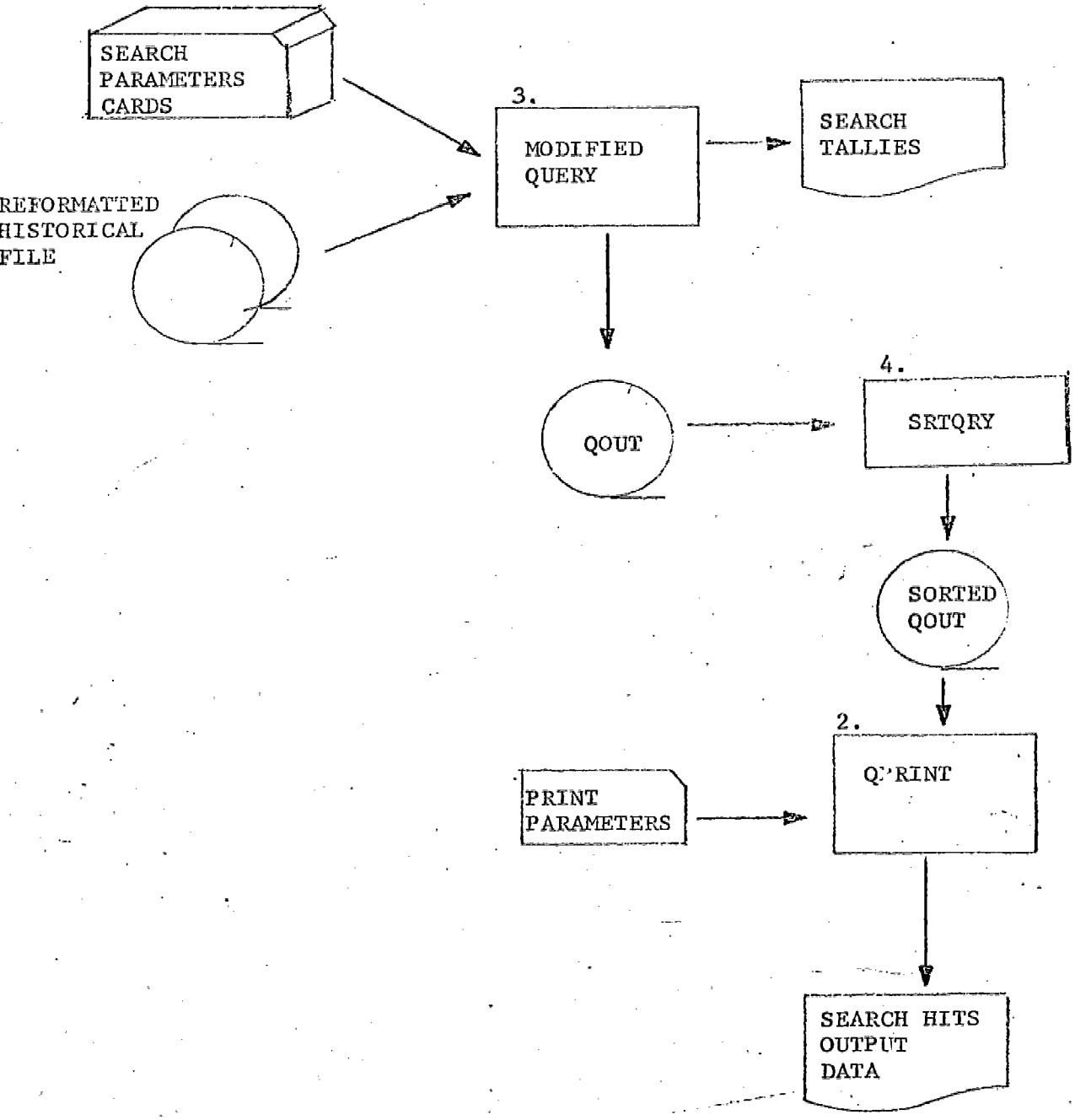
6. RELOADER - Loads the current sequential file as a direct access file. IBM utility - LYNETTE

II. SEARCH PROCEDURES

A. SEARCH OF CURRENT FILE



B. SEARCH OF HISTORICAL FILE



C. FILE SEARCH PROGRAMS

1. RERICQRY - Processes input search parameters by directly accessing the descriptor inverted index and producing an array of keys to current records that fall within these parameters. Then the program directly accesses the current file, appends a search identifier and writes the records to a sequential print file.

2. QPRINT - Existing QUERY print program.

3. QUERY - Modification of existing QUERY program. Modified program runs in approximately 1/3 the time originally required by the program and is better suited to the multi task enviro-ment it must run in.

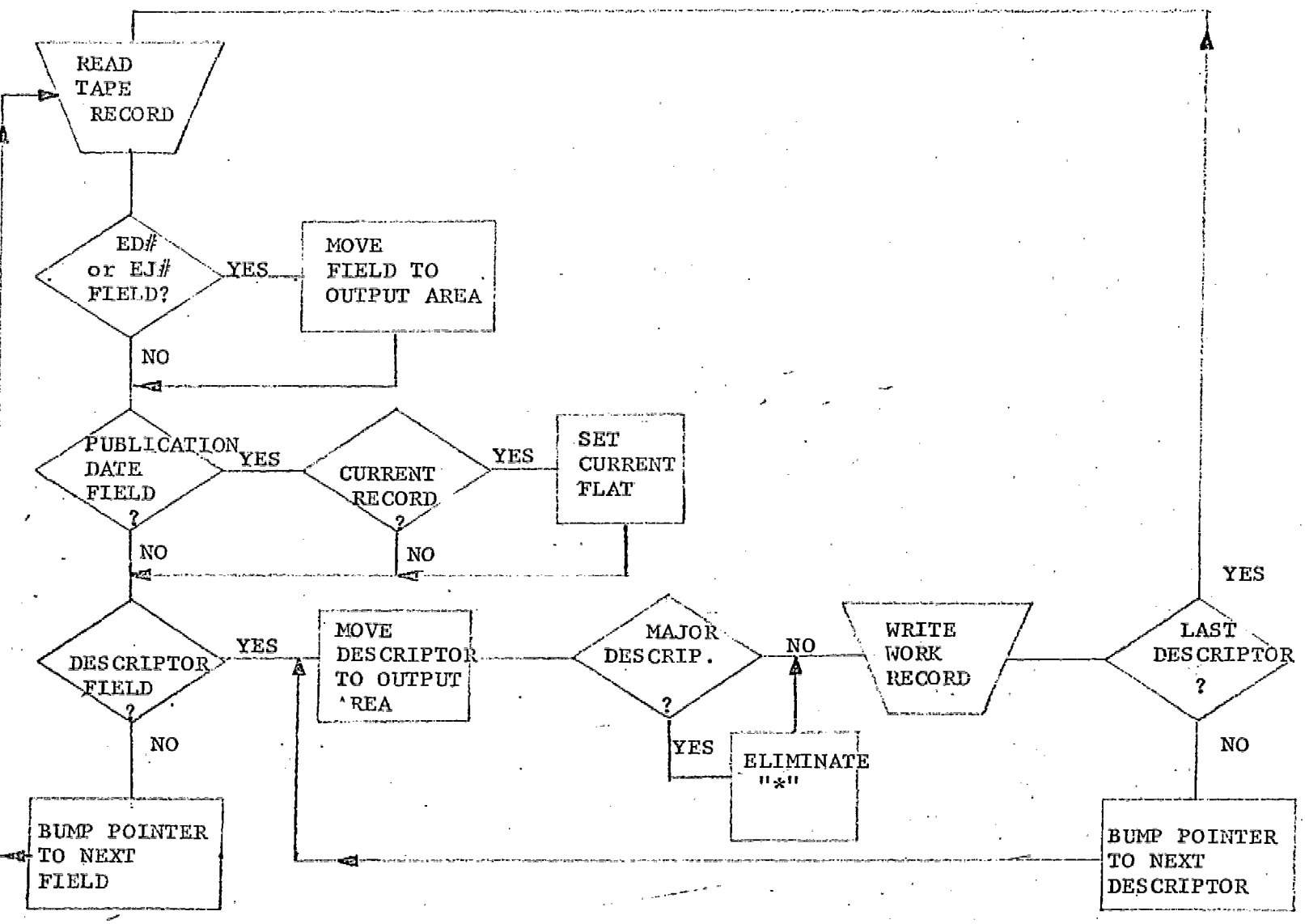
4. SRTQRY - IBM utility sort, IERRCOØØ.
   Sort field: Search Identifier (major), ED or EJ number (minor), ascending order.

PROGRAM DOCUMENTATION

1. FILE BUILDING AND MAINTENANCE PROGRAMS

A. RBLDERIC

1. Language: PL1 Level F
2. Input: ERIC and/or CIJE tapes
3. Output: Descriptor work file
4. Flow chart:



OTIS SYSTEM SOFTWARE & HARDWARE CONFIGURATION

Operating System:

OTIS operates under the full OS/360-MVT Control Program which provides for the concurrent scheduling of a variable number of jobs in a dynamic environment. Jobs in an OS/360 multiprogramming environment are protected from alteration by other user programs, systems programs, or input/output operations through OS/360 support of the storage protection feature. Our full multiprogramming support is designed to increase system through-put and decrease job turn-around time while optimizing the use of the system resources.

Language Processors:

A wide range of language compilers and assemblers that contain powerful diagnostic capabilities is available on the OTIS system. These include PL/1, COBOL, FORTRAN, RPG, ALGOL, and Assembler Language.

Teleprocessing:

Remote terminal controls and data transmission requirements can be handled simultaneously with other processing through the multiprogramming facilities of OS/360. Our use of the Telecommunications Access Method (TCAM) to construct telecommunications service provides the bonus facilities to perform error recovery, error logging, on-line terminal tests, and other functions that increase system availability.

Hardware Configuration:

An IBM S/360 Model 50 with attached 2361 Core Storage, Model 1, provides 1,179,648 bytes of directly addressable core memory. This is one of the largest in the Northwest and allows us to run jobs requiring extremely large amounts of core storage. Input/output devices include:

4 - 2400 Series Magnetic Tape Units - (9 track R/W heads, 800 BPI recording density.)

2 - 2314 Direct Access Storage Facilities - (Each stores 233,400,000 bytes or 466,800,000 packed digits on 8 removable disk packs.)

1 - 2321 Data Cell Drive, Model 1 - (Capacity of 400 million bytes or 800 million packed digits on-line.)

1 - 2540 Card Read Punch - (Reads 1000 cards per minute, punches 300 cards per minute.)

OTIS System Software & Hardware Configuration (contd.)

1 - 1403 Printer, Model N1 - (Line width 132 print positions, print rate 1,100 lines per minute.)

For telecommunications, users can choose either IBM 2740 Communications Terminals or TTY-33 Teletype Terminals as their needs dictate.

COMMENTS AND ALTERNATIVES TO THE BASIC SYSTEM

1. Index Building

After the descriptors were extracted from the tape files, sorted and printed, an unexpected problem was detected. Approximately 5,000 key punch errors were in the descriptors. Inverted letters, extra letters, missing letters, descriptors separated by commas instead of semi-colons, or by blanks — all were there. Also "used for" terms instead of the correct descriptors were there. In terms of the 775,000 descriptors extracted, 5,000 is a small percentage, but too many for the inverted index.

An unplanned for program, RFIXI, was developed to correct these errors in the inverted index. No attempt was made to correct the tape files, however, thus searches of the history tapes will miss those records with errors. The great majority of these errors were in the Journal tapes.

Some of the errors were so appropriate that they were corrected with regret. "CURRRICULUM" must indicate a strong belief in the 3 R's; and "EMOTIONALLY DISTRUBED CHILDREN" even disturbed the word. But when "SEX EDUCATION" is fouled up, that's too much.

2. Reformating Records:

A conference with the search negotiators familiar with the needs of the users determined that the following fields were the most useful. They are the only ones retained in the records of the search files at OTIS.

Accession Number	(key to direct access current file)
Publication Date	
Title	
Personal Author	
Descriptors	
Identifiers	
EDRS Price	
Descriptive Note	
Abstract	
Availability	
Journal Citation	

3. File Division:

The search negotiators also determined that the current file would contain all records that indicated a publication date of 1969 to the present. The history file consists of publication dates of 1968 and before. This will be adjusted on an annual basis to keep the current file from growing to unmanageable proportions.

Comments and Alternatives to the Basic System (contd.)

4. Index Format:

The choice of a fixed length index record and trailer records was chosen with possible future developments in mind. This type of record format would interface with the teleprocessing system used by OTIS. With table descriptions of the index added, any of the 120 terminals over the state could display the ED numbers for any descriptor. With the insertion into the system of a modification of the processing program, a search could be entered by terminal. A tally of the hits and/or the selected ED numbers could be displayed. Then, at the request of the terminal operator, the search hits keys could be stored for later selection from the data file and printing at a specified time.

5. Alternatives to the Basic System:

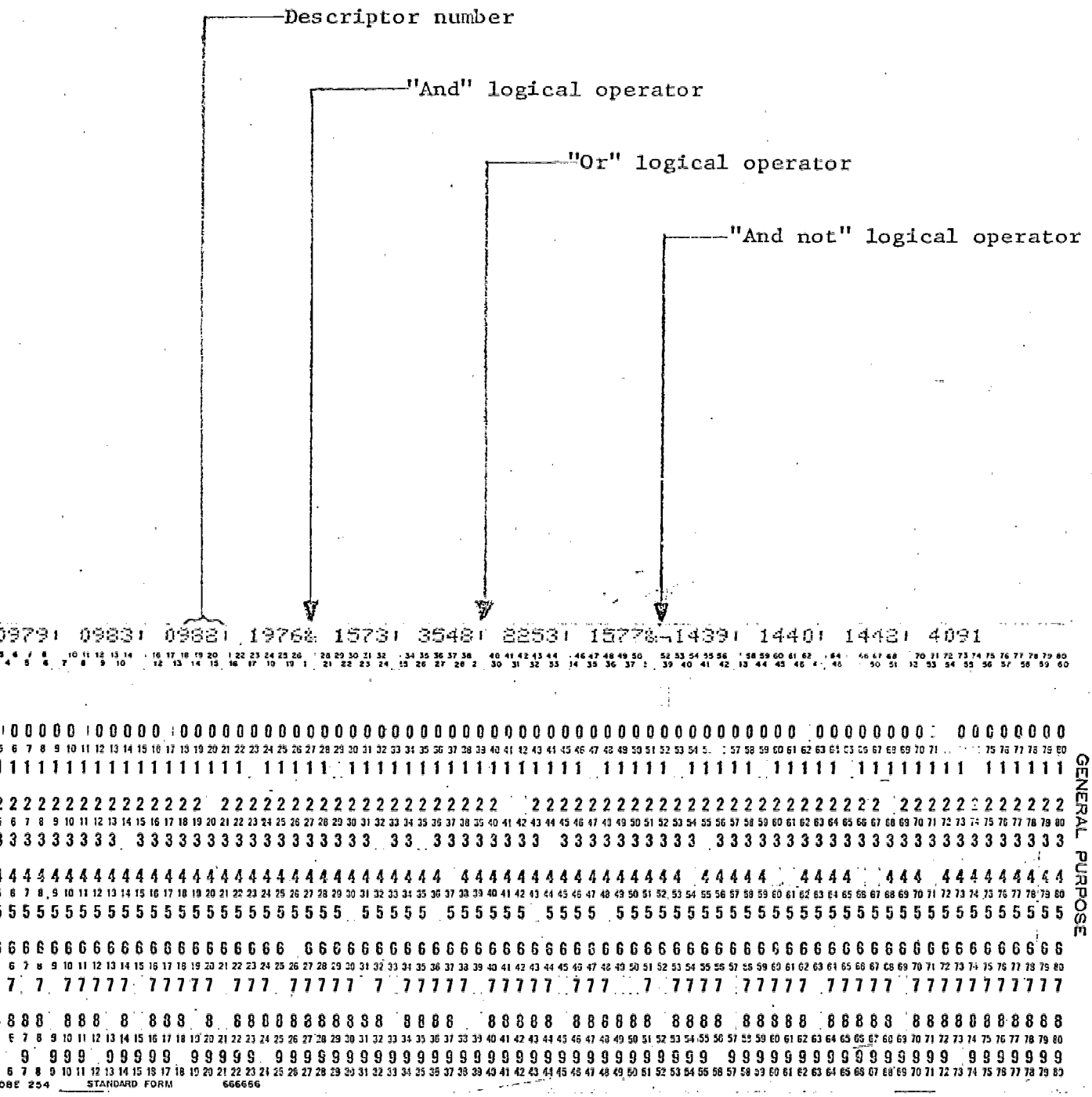
For shops with limited direct access storage a variation of this retrieval system may be useful. The direct access processing of the inverted index could still be employed. At the point where the first array contains the hits, instead of accessing a data file, a file of records consisting of search identifiers and ED and EJ numbers could be output. This file could then be sorted by ED numbers. A quick run of the tape file could pick off the desired records and attach the search identifiers to the selected records.

The selected records could then be sorted by search identifiers and printed.

Even with the extra steps, this procedure should require less than half of the CPU time the modified QUERY program takes. Or about 1/6 the time used by the original program.

An adaptation of the search program is possible for installations with limited core for execution. The program is constructed in seven logical blocks. These blocks could be handled as sub-tasks with a small control program to roll in the appropriate block as it is needed. A degradation of the time factor is involved, but the gain over sequential searching is still massive.

Example of search card for OBIAS



STATISTICS OF RUN TIMES

The following statistics refer to actual production runs, as opposed to test runs, for the office of Instructional Technology, Oregon Board of Education, Salem, Oregon, and the ERIC Clearinghouse of Educational Management, U. of O., Eugene, Oregon.

Production Run 1:

Number of searches	=	1
Number of descriptors	=	9
Number of "or's"	=	6
Number of "and's"	=	1
Number of "and not's"	=	1
Total Hits Produced	=	33
Total Search Time	=	1 min. 46.35 sec.
Main CORE Required	=	192 K

Production Run 2:

Number of searches	=	12
Number of descriptors	=	73 (6 per search)
Number of "or's"	=	51 (4.25 per search)
Number of "and's"	=	10
Number of "and not's"	=	Ø
Total Hits Produced	=	989 (82.5 per search)
Total Search Time	=	8 min. 37.45 sec. (43 sec. per search)
Main CORE Required	=	206K

Production Run 3:

Number of searches	=	40
Number of descriptors	=	307 (7.7 per search)
Number of "or's"	=	218 (5.4 per search)
Number of "and's"	=	40
Number of "and not's"	=	6
Total Hits Produced	=	2180 (54 per search)
Total Search Time	=	90 min. 00.53 sec. (2 min. 30 sec. per search)
Main CORE Required	=	222K

"Search Time" includes only actual search time; print time is not included. Based on experience with these and other production runs, search and print time should rarely exceed 2 min. 30 sec. Using a current figure of $180 per CPU hour, CPU cost per average search comes to $7. A handling charge of $5 per search must be added to this figure, bringing a total per search cost to $12.