THE SHARED CATALOGING SYSTEM OF THE OHIO COLLEGE LIBRARY CENTER

Frederick G. KILGOUR, Philip L. LONG, Alan L. LANDGRAF, and John A. WYCKOFF: Ohio College Library Center, Columbus, Ohio

Development and implementation of an off-line catalog card production system and an on-line shared cataloging system are described. In off-line production, average cost per card for 529,893 catalog cards in finished form and alphabetized for filing was 6.57¢. An account is given of system design and equipment selection for the on-line system. File organization and programs are described, and the on-line cataloging system is discussed. The system is easy to use, efficient, reliable, and cost beneficial.

The Ohio College Library Center (OCLC) is a not-for-profit corporation chartered by the State of Ohio on 6 July 1967. Ohio colleges and universities may become members of the center; forty-nine institutions are participating in 1971/72. The center may also work with other regional centers that may "become a part of any national electronic network for bibliographic communication."

The objectives of OCLC are to increase the availability to individual students and faculty of resources in Ohio's academic libraries, and at the same time to decrease the rate of rise of library costs per student.

The OCLC system complies with national and international standards and has been designed to operate as a node in a future national network as well as to attain the more immediate target of providing computer support to Ohio academic libraries. The system is based on a central computer with a large, random access, secondary memory, and cathode ray tube terminals which are connected to the central computer by a network of telephone circuits. The large secondary memory contains a file of bibliographic records and indexes to the bibliographic record file. Access to this central file from

the remote terminals located in member libraries requires fewer than five seconds.

OCLC will eventually have five on-line subsystems: 1) shared cataloging; 2) serials control; 3) technical processing; 4) remote catalog access and circulation control; and 5) access by subject and title. This paper concentrates on cataloging; the other subsystems are not operational at the present time.

Figure 1 presents the general file design of the system. The shared cataloging system has been the first on-line subsystem to be activated, and the files and indexes it employs are depicted in Figure 1 by the heavy black lines and arrows. As can be seen in the figure, much of the system required for shared cataloging is common with the other four subsystems.

The three main goals of shared cataloging are: 1) catalog cards printed to meet varying requirements of members; 2) an on-line union catalog; and 3) a communications system for requesting interlibrary loans. In addition, the bibliographic and location information in the system can be used for other purposes such as book selection and purchasing.

The only description of an on-line cataloging system that had appeared in the literature during the development of the OCLC system is that of the Shawnee Mission (Kansas) Public Schools (1). The Shawnee Mission cataloging system produces uniform cards from a fixed-length, non-MARC record. The OCLC system uses a variable-length MARC record and has great flexibility for production of cards in various formats. There are a number of reports describing off-line catalog card production systems, including systems at the Georgia Institute of Technology (2), the New

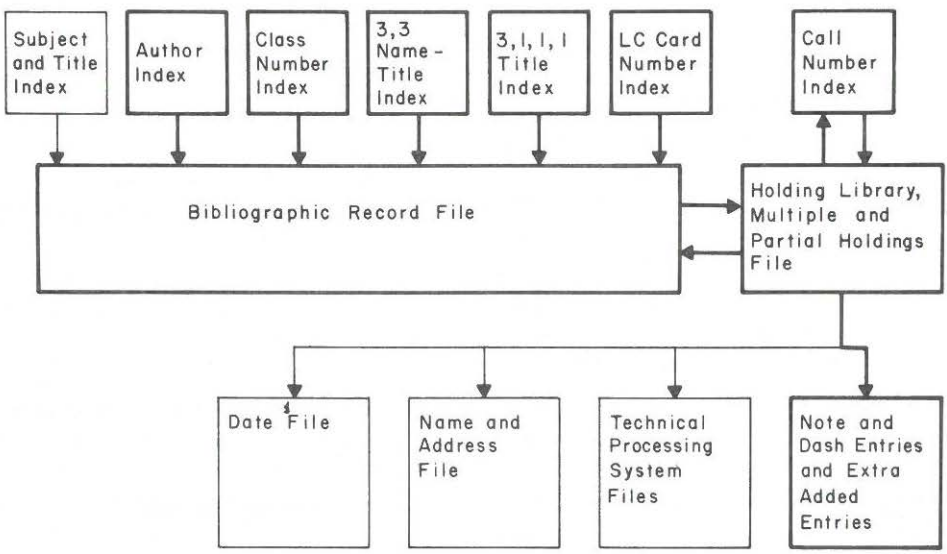

Fig. 1. General File Design; Shared Cataloging Subsystem in Heavy Lines.

England Library Information Network (NELINET) (3), and the University of Chicago (4). The flexibility of the OCLC system distinguishes it from these three systems as well.

CATALOG CARD PRODUCTION—OFF-LINE

An off-line catalog card production system based on a file of MARC II records was activated a year before the on-line system (5). OCLC supplied member libraries with request cards (punch cards prepunched with symbols for each holding library within an institution). For each title for which catalog cards were needed, members transcribed Library of Congress (LC) card numbers onto a request card. Members sent batches of cards to OCLC at least once a week. At OCLC, the LC card numbers were keypunched into the cards and new requests were combined with unfilled requests to be searched against the MARC II file. By the spring of 1971, over 70 percent of titles requested were found the first time they were searched.

The selected MARC II records were then submitted to a formatting program that produced print images on magnetic tape for all cards required by a member library. The number of cards to be printed was determined by the number of tracings on the catalog record and the number of catalogs into which cards were to go including a regional union catalog (the Cleveland Regional Union Catalog) and the National Union Catalog. Individual cards were formatted according to options originally selected by the member library. These options included: 1) presence or absence of tracings and holdings information on each of nine different types of cards; 2) three different indentions for added entries and subject headings; 3) a choice of upper-case or upper- and lower-case characters for each type of added entry and subject heading; and 4) many formats for call numbers. OCLC returned cards to members in finished form, alphabetized within packs for filing in specific local catalogs.

The primary objective of off-line operation was the production of catalog cards at a lower cost than manual methods in OCLC member libraries. Early activation of off-line catalog card production did reduce costs and gave some members an opportunity to take advantage of normal staff turnover by not filling vacated positions in anticipation of further savings after activation of the on-line system.

Other objectives of off-line operation were the automated simulation of on-line activity in member libraries and development and implementation of catalog card production in preparation for card production in an on-line operation. The number of catalog card variations required by members, even after members had reviewed and accepted detailed designs of card products, proved to be higher than anticipated. More than one man-year was expended after activation of the off-line system in further development and implementation to take care of the formats and card dissemination variations requested by specific libraries. The one year advance start on

catalog production made possible by using MARC II records in the off-line mode proved to be a far greater blessing than anticipated, for it would have been literally impossible to have activated on-line operation and catalog card production simultaneously.

A major goal of OCLC card production is elimination of uniformity required by standardized procedures. The OCLC goal is to facilitate cooperative cataloging without imposing on the cooperators. The cost to attain this goal is slight, for although there is a single expense to establish a decision point in a computer program, the cost of selection among three or thirty alternatives during program execution is infinitesimal.

Design of catalog cards and format options began four months before off-line activities. Two general meetings of the OCLC membership were held at which card formats were reviewed and agreed upon in a general sense. Next, the OCLC staff published a description of catalog card production and procedures for participation (6). This publication was reviewed by the membership and format variations were reported for inclusions in the procedure. Members reported few variations at this time, but when implementation for individual members was undertaken, it was necessary to build many additional options into the computer programs. To assist the OCLC staff in defining options for off-line catalog products and on-line procedures, an Advisory Committee on Cataloging was established. This committee met several times and provided much needed guidance and counsel.

The catalog card format options that members could select were extensive. For example, although the position of the call number was fixed in the upper left-hand corner of the card, there were 24 basic formats for LC call numbers, and libraries using the Dewey Decimal Classification could format their call numbers as they wished. In general, the greatest number of format options are associated with call numbers, probably because there has never been a standard procedure for call number construction.

Programs

Because designing, writing, coding, and debugging of catalog card production programs can cost tens of thousands of dollars, OCLC sought existing card production programs that could run on computers at Ohio State University, which is the generous host of the Ohio College Library Center. Only two programs were located that could both produce cards in the manner required by OCLC and run on OSU computers. Card production costs were not available for one of the programs, but because analysis suggested that the design of the program would create very high card costs, this program was not selected. The other program had been written and used at the Yale University Library, and although the card production costs were high, it was known that changes could be made to increase efficiency. Thus, arrangements were made to obtain and run the Yale programs at OSU.

Members were free to choose a variety of format options and submitted on a Catalog Profile Questionnaire (Figure 2) their specifications for each

catalog. Holdings information and tracings could be printed on any or all of nine types of cards: 1) shelf list; 2) main entry; 3) topical subject; 4) name as subject; 5) geographic subject; 6) personal and corporate added entries; 7) title added entry; 8) author-type series added entry; and 9) title-type series added entry. Subject headings and added entries could have top-of-card or bottom-of-card placement and could be printed in all upper-case or in upper- and lower-case characters. Any type of subject heading and added entry could begin at the left edge of the card or at the first, second, or third indention. Other options are described in the *Manual for OCLC Catalog Card Production* (5).

The data received on Catalog Profile Questionnaires were transferred to punch cards and a computer program written in SNOBOL IV embedded the information in the form of a Pack Definition Table (PDT) in one of the principal catalog production programs named CONVERT (CNVT). Each PDT defined the cards to go into the catalogs of one holding library, a holding library being a collection with its own catalog.

The first major program in the processing sequence was PREPROS, which was written in IBM 360 Basic Assembler Language (BAL) and run on an IBM 360/75. PREPROS converted records from the weekly MARC II tapes to an OCLC internal processing format, including conversion of MARC II characters from ASCII to EBCDIC code. This program also parsed LC call numbers and partially formatted them. It also checked for end-of-field and end-of-record characters and verified the length of record. Finally, it wrote the output records in LC card number sequence into huge variable format blocks of 20,644 characters. The large blocks reduced computer costs since the pricing algorithm employed on the IBM 360/75 imposed a charge for each physical read and write operation.

The magnetic tape output weekly by PREPROS was then submitted to CNVT together with the old master file of bibliographic records in LC card number order and a file of request cards that had been sorted in LC card number order. CNVT merged the records on the weekly tape with the master file and then matched the requests by LC card number. When a match was obtained, CNVT deleted some fields from the bibliographic record and formatted the call number according to the specifications of the library that had originated the request. It then wrote the modified record and associated PDT's onto an output tape in external IBM 7094 binary-coded-decimal (BCD) character code with the record format converted to that of the Yale Bibliographic System. The second principal product of CNVT was the new master tape of bibliographic records that would become the old master for the next week's run. CNVT also punched out a card bearing the LC card number for each request card for which there was a match. These punch cards were used to withdraw cards from the request card file so that they would not be submitted again. CNVT was first run on an IBM 360/50.

The tape file of modified records and PDT's was then submitted to

OHIO COLLEGE LIBRARY CENTER

Catalog Profile Questionnaire

I. To define the pack of a receiving catalog, the Member should complete the following table. Directions for completing the table are in the Instruction Manual, pp. 2-3. Leave blank rows for types of entry not to be included in this pack.

II. 1. What is the name of the holding library or collection for which this pack contains cards? Juvenile AKJ

2. What is the name of the receiving catalog into which this pack will go? Union shelflist AKRsl

3. If this receiving catalog is not in the holding library or collection, put in the following box the stamp to appear above the call number (see Instruction Manual).

Type of Entry	Holdings Information		Tracings		Subject Headings Position		Indention of Headings at Top of Cards (first line only)				Capitalization of Headings at Top of Cards.	
	Yes	No	Yes	No	Top of Card	Bottom of Card	Left edge	First indention	Second indention	Third indention	Upper case	Upper and lower case
Main Entry to be Arranged by Call Number (Shelf List)	✓		✓									
Main Entry												
Topical Subject Entry												
Name as Subject Entry												
Geographic Subject Entry												
Personal or Corporate Added Entry												
Title Added Entry												
Author-Type Series Added Entry												
Title-Type Series Added Entry												

Institution: University of Akron

Fig. 2. Catalog Profile Questionnaire.

EXPAND, a modified Yale program written in MAD and run on an IBM 7094. By combining the number of tracings and PDT requirements, EXPAND developed a card image for each catalog card required by the requesting library. It also prepared a sort tag for each image so that the image could be subsequently sorted by library into packs and alphabetized within each pack. EXPAND essentially did the formatting of catalog cards except for the complex LC call number formatting carried out by CNVT.

The file of card images was passed to a program named Build Print Tape (BLDPT) written in BAL and run on the IBM 360/75. BLDPT first converted the external IBM 7094 BCD characters to EBCDIC. Next BLDPT sorted the images, and finally, it arranged the images on a single tape to allow printing on continuous, two-up catalog card forms—the first half of the sorted file was printed on the left-hand cards and the second half on the right.

The PRINT program was also written in BAL but run on an IBM 360/50. It was designed so that either the entire file or a segment as small as four cards could be printed; the latter feature was of greatest use in reprinting cards that for one of several reasons were not satisfactorily printed during the first run. Cards were printed six lines to an inch and the print train used was a modified version of the train designed by the University of Chicago which in turn was a modified version of the IBM TN train.

The printer attached to the IBM 360/50 was an IBM 1403 N1 printer. This printer appears to be superior to any other high-speed printer currently available, but to obtain a product of high quality, it was necessary to fine-tune the printer, to use a mylar ribbon from which the ink does not flake off, and to experiment with various mechanical settings to determine the best setting for tension on the card forms and for forms thickness. Above all, patience in large amounts was required during initial weeks when it seemed as though a messy appearance would never be eliminated.

OCLC off-line catalog card production programs were written in assembler language and higher level languages. Use of higher level languages for character manipulation incurs unnecessarily high costs. Therefore, for a large production system like OCLC, it is absolutely required that processing programs and subroutines that manipulate all characters, character by character, be written in an assembler language to obtain efficient programs that run at low cost. Programs that do not manipulate characters, such as the OCLC program for embedding PDT's in CNVT, may well be written in a higher level language.

Materials and Equipment—A Summary

Off-line catalog production was based on availability of MARC II records on magnetic tapes disseminated weekly by the Library of Congress. Without the MARC II tapes, the off-line procedure could not have operated. Each week, the new MARC II records were added to the previous cumulated master file also on magnetic tape, and previously unfilled and new requests were run against the updated file.

OSU computers employed were an IBM 360/75, an IBM 360/50, an IBM 7094, and an IBM 1620. The run procedure was complex and therefore somewhat inefficient, but this inefficiency was traded off against a predictably high expense to write a new card formatting program.

Members submitted a request for card production on a punch card on which the member had written an LC card number. Members could specify a recycling period of from one to thirty-six weeks for running their request cards against the MARC II file before unfulfilled requests would be returned. In general, request cards bore LC card numbers for that section of the MARC II file that was complete; at first, the file was inclusive for only "7" series numbers, but in early 1971 the RECON file for "69" numbers was added. Request cards often numbered several thousand a week.

Catalog card forms are the now-familiar two-up, continuous forms with tractor holes along each side for mechanical driving. The card stock is Permalife, one of the longest-lived paper stocks available. A thin slit of about one thirty-second of an inch in height converts each three-inch vertical section of card stock to 75 mm. The lowest price paid in a lot of a half million cards has been $8.065 per thousand.

After having been printed, the card forms are trimmed on a modified UARCO Forms Trimmer, model number 1721-1. This trimmer makes four

continuous cuts in the forms and produces cards with horizontal dimensions of 125 mm. Cards are stacked in their original order as printed and are therefore in filing order. The trimmer operates at quoted speeds of 115 and 170 feet per minute or 920 and 1,360 cards per minute. Measurements of speeds of operations confirmed these ratings.

Results

The off-line catalog production system produced 529,893 catalog cards from July 1970 through August 1971 at an average cost of 6.57 cents per card. This cost includes over twenty separate cost elements plus a three-quarter cent charge for overhead. The firm of Haskins & Sells, Certified Public Accountants, reviewed the costing procedures that OCLC employs, found that all direct costs were being included, and recommended the three-quarter cent overhead charge.

The number of extension cards varies from library to library depending almost entirely on the types of cards on which libraries have elected to print tracings. However, one university library with a half-dozen department libraries and requiring tracings on only shelf list and main entry cards averages approximately six cards per title.

Cataloging using the OCLC off-line system results in a decrease of staff requirements, and some libraries that used the system during most of the year found that they needed less staff in cataloging. Reduction of staff by taking advantage of normal staff turnover facilitated financial preparation for the OCLC on-line system in these libraries.

Evaluation

Despite the obvious inefficiences generated by running production computer programs on four different computers in two different locations and despite inefficiencies in the programs themselves, computer costs to process MARC II tapes and to format catalog cards, but not to print them, was 2.27 cents per card. As will be shown later, newer and more efficient programs have halved this cost, but even at 2.27 cents per card for formatting and .33 cents per card for printing, the cost of OCLC off-line card production is less than half the cost of more traditional card production methods (7).

Two features originally designed into the system were never implemented, somewhat diminishing the usefulness of the system for some libraries. One of the incompleted features was a technique for deleting, changing, or adding a field to a MARC record (this capability exists in the on-line system). Absence of this procedure meant that libraries had to accept LC cataloging without modification except to call numbers. The second missing feature was the ability to print multiple holding locations on cards (this capability also exists in the on-line system) although it was possible to print multiple holdings in one location. This deficiency limited the usefulness of the system for large libraries processing duplicates into

two or more collections. Both of these features could have been activated, but shortage of available time prior to activation of the on-line system prevented their implementation.

Figure 3 shows the high quality of the catalog cards produced. Subsequent to attainment of this level of quality, there have been no complaints from members except in cases where a piece of chaff from the card forms went through the printer and caused omission of characters. OCLC continues to vary the design of its continuous forms to achieve completely chaff-free stock.

The shortest possible time in which cards could be received by the member library after submitting a request card was ten days, but it is doubtful that this response time was often achieved. The minimum average response time for the three-quarters of requests for which a MARC record was located on the first run was two weeks. Delays at a computer center or incorrect submission of a run could extend this delay to three and four weeks, and unfortunately such delays were cumulative for subsequent requests until the "weekly" runs were made sufficiently more often than weekly to catch up. If another delay occurred during a catch-up period, the response time further degraded. During the fourteen months of operation, there were two serious delays.

The amount of normal turnover that occurred in OCLC libraries during the fourteen months and that was taken advantage of by not filling positions was too small to reduce the financial burden incurred in starting up the on-line system. A few libraries demonstrated that it was possible to take advantage of such attrition. However, 20 percent of the libraries did not participate in the on-line system and perhaps half of those who did participate were uncertain as to whether the on-line cataloging system would operate or would operate at a saving.

When feasibility of on-line shared cataloging has been substantiated and other centers begin to implement similar systems, it should be possible to activate off-line catalog production sufficiently in advance of on-line implementation to enable participants to take adequate advantage of normal attrition to minimize, or nearly eliminate, additional expenditures. Experience such as that of OCLC will enable new centers to calculate the number of months necessary for off-line production required to reduce salary expenditures by an amount needed to finance the on-line system.

SHARED CATALOGING—ON-LINE

The cataloging objectives of the on-line shared cataloging system are to supply a cataloger with cataloging information when and where the cataloger needs the information and to reduce the per-unit cost of cataloging. Catalog products of the system are the same as the off-line system—catalog cards in final form alphabetized for filing in specific catalogs; the on-line system is not limited to MARC II records but also allows cataloging input by member libraries. The shared cataloging system, which accommo-

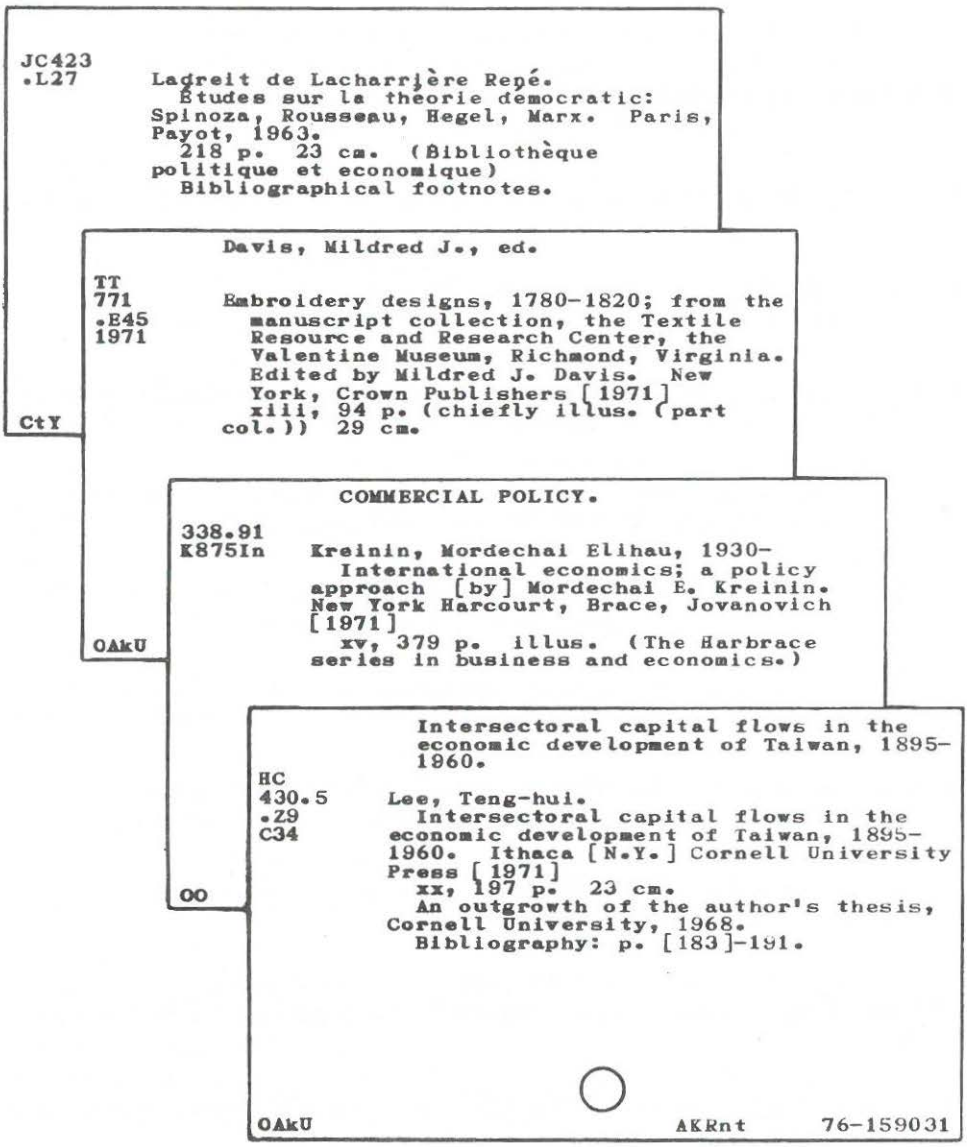

Fig. 3. Computer-Produced Catalog Cards. (FIGURE REDUCED 25%)

dates all cataloging done in modern European alphabets, builds a union catalog of holdings in OCLC member libraries as cataloging is done. One library, Wright State University, is converting its entire catalog to machine-readable form in the OCLC on-line catalog. The third major goal is a communications system for transacting interlibrary loans.

System Design and Equipment Selection

Figure 4 depicts the basic design of computer and communication com-

ponents for the comprehensive system comprised of the five subsystems described in the introduction. The machine system for shared cataloging was designed to be a subsystem of the total system so that subsequent modules could be added with minimal disruption. Similarly, the logical design of the shared cataloging subsystem was constructed so that the modules of shared cataloging would be common to the remaining file requirements as shown in Figure 1.

Design of the on-line shared cataloging system began with a redefinition of the catalog products of off-line catalog production (5). In this exercise, the Advisory Committee on Cataloging, comprised of members from seven libraries, contributed valuable assistance. The committee was also most helpful in designing the formats of displays to appear on terminal screens.

Important decisions in the design of the computer, communications, and terminal systems were those involving mass storage devices and terminals. Random access storage was the only type feasible for achieving the objective of supplying a user with bibliographic information when and where he needed it. Hence, random access memory devices were selected for the comprehensive system and ipso facto for shared cataloging.

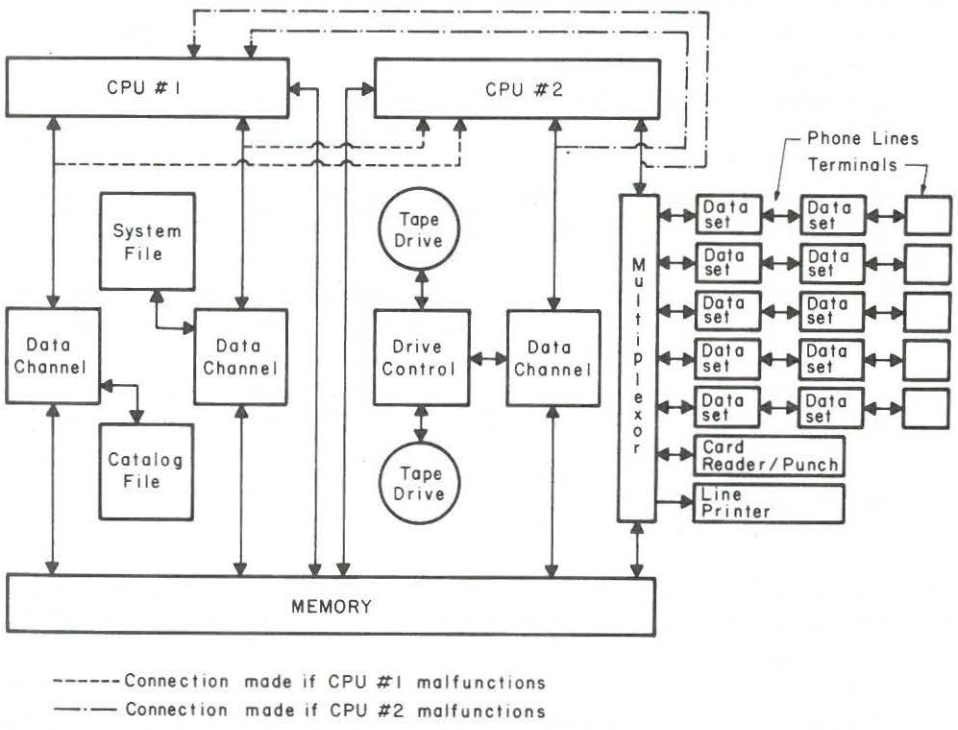

Fig. 4. Computer and Communication System.

The cathode ray tube (CRT) type of terminal was selected primarily because of its speed and ease of use by a cataloger. CRT terminals are far more flexible in operation than are typewriter terminals from the viewpoint of both the user and machine system designer. For these reasons, CRT terminals can enhance the amount of work done by the system as a whole.

It was originally planned to select a computer without the assistance of computerized simulation, but in the course of time, it became clear that it was impossible to cope with the interaction among the large number of variable computer characteristics without computerized simulation. Therefore, a contract was let to Comress, a firm well known for its work in computer simulation. Ten computer manufacturers made proposals to OCLC for equipment to operate the five subsystems at peak loading (an average five requests per second over the period of an hour).

All ten proposed computer systems failed because simulation revealed inefficiencies in their operating systems for OCLC requirements. OCLC and Comress staff then proposed a modification in operating systems, which the manufacturers accepted. The next series of trials revealed that more than half of the computers or secondary memory files would have to be utilized over 100 percent of the time to process the projected traffic. As a result of these findings, one computer manufacturer withdrew its proposal, and five others changed proposals by upgrading their systems. On the final simulation runs, the percent of simulated computer utilization ranged from 19.70 percent to 114.31 percent.

A subsequent investigation of predictable delays due to queuing in such a system showed that unacceptable delays could arise if computer utilization rose above 30 percent at peak traffic. Three manufacturers proposed computer systems that were under 30 percent utilization and, for these, a trade-off study was made that included such characteristics as cost, reliability, time to install the applications system, and simplicity of program design. The findings of the simulation and trade-off studies provided the basis of the decision to select a Xerox Data Systems Sigma 5 computer.

Major components of the OCLC Sigma 5 are the central processing unit (CPU), three banks of core memory with a total capacity of 48 thousand 32-bit words or 192 thousand 8-bit bytes, a high speed disk secondary memory, 10 disk-pack spindles with total capacity of 250,000,000 bytes plus two spare spindles, two magnetic tape drives, two multiplexor channels, five communications controllers, a card reader, card punch, and printer. The character code is EBCDIC. Figure 5 illustrates the Sigma 5 configuration at OCLC. In this configuration, the burden of operating communications does not fall on the CPU so that there is no requirement for "cycle stealing" that slows processing by a CPU.

The lease cost to OCLC of the equipment represented in Figure 5 is $16,317 monthly. The listed monthly lease of the equipment is $21,421 from which an educational discount of 10 percent is deducted. (The remaining difference is due to a rebate because the original order included secondary

memory units that XDS was to obtain from another manufacturer who proved incapable of supplying units that fulfilled specifications. Hence, XDS was forced to supply other memory units having a higher list price but has done so at a cost per bit of the units originally ordered.)

The printer furnished with the Sigma 5 does not provide the high-quality printing required for library use. At the present time, OCLC prints catalog cards on an OSU IBM 1403-N1 printer that without doubt provides the highest quality printing currently available from a line printer. However, OCLC is designing an interface between a Sigma 5 and an IBM 1403

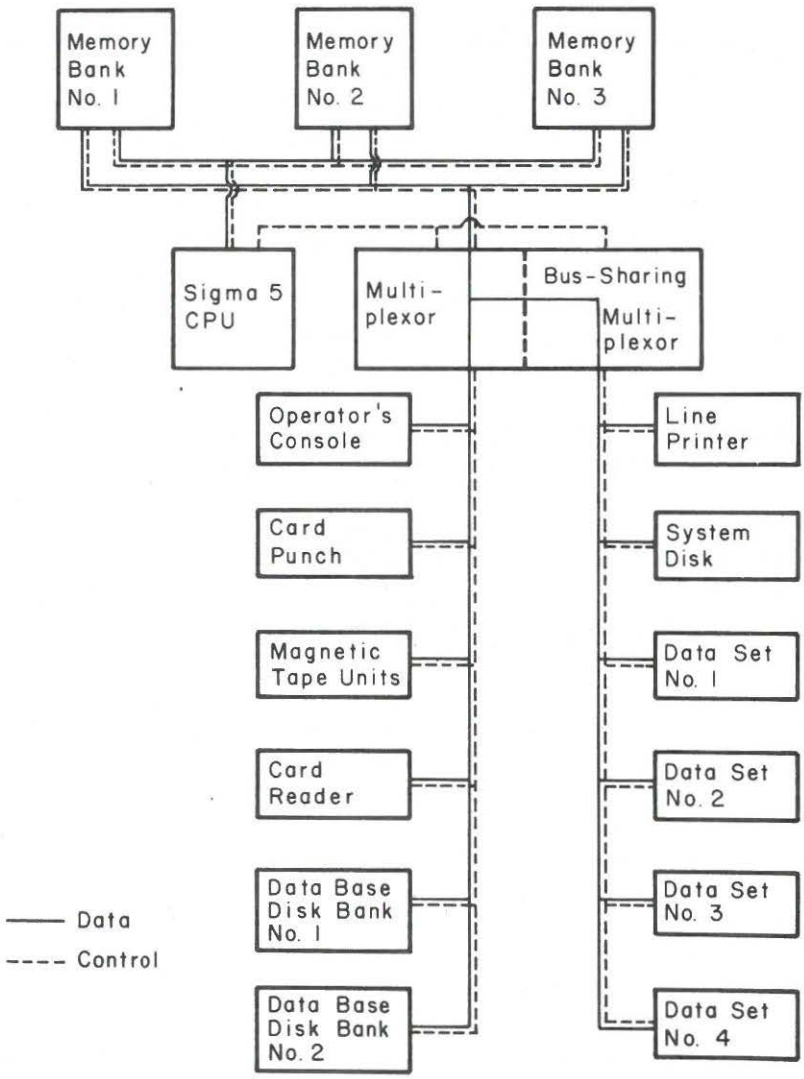

Fig. 5. XDS Sigma 5 Configuration.

printer; XDS is also developing a new type of printer that will provide high quality output. When the Sigma 5 can produce quality printing, it will be fully qualified to be used for nodes in national networks.

As has already been stated, the CRT-type terminal was selected because of its ease of use. Moreover, the simulation study confirmed that CRT terminals would place far less burden on the central computer and therefore, for the OCLC system, would make possible selection of a less expensive computer than would be required to drive typewriter terminals. Although typewriter terminals cost less, the total cost could be higher for a system employing typewriter terminals than for one using CRT's because of greater central computer expense.

Library requirements for a CRT terminal are: 1) that the terminals have the capability of displaying upper- and lower-case characters and diacritical marks; 2) that the image on the screen be highly legible and visible; 3) that the terminal possess a large repertoire of editing capabilities; and 4) that interaction with the central computer and files be simple and swift. System requirements were: 1) that the terminal accept and generate ASCII code; 2) that it make minimal demands for message transmissions from and to the central site; 3) that it have the capability of operating with at least a score of other terminals on the same dedicated line; and 4) that its cost, including service at remote sites, be about $150 per month.

Data were collected on CRT's produced by fifteen manufacturers, and three machines were identified as being prime candidates for selection. OCLC carried out a trade-off study in which thirty-three characteristics were assessed for these three machines. One of the thirty-three (reliability) could not be judged for any of the three because none had yet reached the market. For the remaining characteristics, the Irascope LTE excelled or equaled the other two terminals for twenty-eight characteristics including all nineteen characteristics of importance to the OCLC user. Moreover, the Irascope was outstandingly superior in its ability to perform continuous insertion of characters, line wrap-around during insertion of characters, repositioning of characters so that each line ends in a complete word, and full use of its memory. However, the Irascope was the most expensive—$175 a month as compared with $153 and $166. Nevertheless, the Irascope was selected because of its obvious superiority. Pilot operation by library staffs has not produced complaints concerning visibility or operability; complaints during pilot operation have sprung from failures caused by a variety of bugs in telephone systems and a couple of bugs in the terminals that were subsequently exterminated.

The number of terminals needed by a member library for shared cataloging was calculated on the assumption that six titles could be processed per terminal-hour. It was also assumed that a library might have only one staff member to use the terminal throughout the year. It was further assumed that as much as three months of the terminal operator's time would be lost to vacations, sick leave, and breaks. At the rate of six titles per terminal-hour

and with 2,000 working hours in a year, 12,000 titles would be processed annually assuming full-time use. Since only nine months was assumed to be available, it was estimated that 9,000 titles would be processed at each terminal.

In large libraries where there would be more than one staff member to operate a terminal, there would be three months of time available to do input cataloging, and since only a few libraries will be obtaining less than 75 percent of cataloging from the central system, it appears that a formula of one terminal for every 9,000 titles or fraction thereof cataloged annually would give each library sufficient terminal-hours. In actual operation, operators have been able to work at twice the assumed rate of six titles per terminal-hour so that there is reason to believe that these guidelines will provide adequate terminal capability.

File Organization

The primary data that will enter the total system are bibliographic records, and since the system is being designed to conform to standards, the National Standard for Bibliographic Interchange on Magnetic Tape (8) has been complied with in file design. In other words, the system can produce MARC records from records in the OCLC file format; more specifically, the system can regenerate MARC II records from OCLC records derived originally from MARC II records, although an OCLC record contains only 78 percent of the number of characters in the original MARC II record. Similarly, the system can generate MARC II records from original cataloging input by member libraries.

The simulation study clearly showed that bibliographic data would have to be accessed in the shortest possible time if the system were to avoid generating frustrating delays at the terminal. Imitation of library manual files or of standard computer techniques for file searching would not provide sufficient efficiency. OCLC, therefore, set about developing a file organization and an access method that would take advantage of the computation speeds of computers.

OCLC research on access methods has produced several reports (9,10,11) and has developed a technique for deriving truncated search keys that is efficient for retrieval of single entries from large files. These findings have been employed in the present system that contained over 600,000 catalog records in April 1973, arranged in a sequential file on disks, and indexed by a Library of Congress card-number index, author-title index, and a title index. The research program on access methods did not, however, investigate methods for storing and retrieving records.

Research on file organization included experiments directed toward development of a file organization that would minimize processing time for retrieval of entries or for the discovery that an entry is not in the file. Since the OCLC system is designed for on-line entry of data into the data base, it was not possible to consider a physically sequential file for the index files.

The indexed sequential method of file organization obviates the data-entry obstacle posed by physical sequential organization, but is inefficient in operation. Consequently, scatter storage was determined to be the best method for meeting the efficient file organization requirements of the system.

The findings of the investigation have shown that very large files of bibliographic index entries organized by a scatter-store technique in which search keys are derived from the main entry can be made to operate very efficiently for on-line retrieval and at the same time be sparing of machine time even in those cases where requests are for entries not in the file (12). This research also produced two powerful mathematical tools for predicting retrieval behavior of such files, and a design technique for optimizing record blocking in such files so that, on the average, only one to two physical accesses to the file storage device are needed to retrieve the desired information.

The files displayed in Figure 1 are constructed by a single file-building program designed so that additional modules can be embedded in the program. The program accepts a bibliographic record, assigns an address for it in the main sequential file, and places the record at that address. Having determined the bibliographic record address, the program next derives the author-title search key and constructs an author-title index file entry which contains the pointer to the bibliographic record. Then the program produces an LC card number index entry and a title index entry, each of which contains the same pointer to the bibliographic record.

When a bibliographic record is used for catalog card production, an entry is made in the holdings file. When the first holdings entry is made for a bibliographic record, a pointer to the holdings entry is placed in that record; the pointer to each subsequent holdings entry is placed in the previous holdings entry. An entry is made at the same time in the call number index containing a pointer to the holdings entry.

This file organization operates with efficiency and economy. The files containing the large bibliographic records and their associated holdings information are sequential, and hence, are highly economical in disk space. The technique used ensures that only a low percentage of available disk area need be reserved for growth of these large sequential files. Disk units can be added as needed. Each fixed-length record in the scatter-store files is less than 3 percent of the size of an average bibliographic record, and since 25 percent to 50 percent of these files are unoccupied, the empty disk area is small because of the small record lengths.

Sequential Files

The bibliographic record file and holdings file are sequential files, the holdings file being a logical extension of the bibliographic record file. A record is loaded into a free position made available by deletion of a record or into the position following the last record. Whenever a new version of a

record updates the version already in the file, the new record is placed in the same location as the old if it will fit; otherwise, it is placed at the end of the file and pointers in the indexes are changed. There is a third, small sequential file containing unique notes for specific copies, dash entries, and extra added entries.

Each bibliographic record contains the information in a MARC II record. Each record also contains a 128-bit subrecord capable of listing up to 128 institutions that could hold the item described by the record. At the present time, only 49 of the 128 bits are used since there are 49 institutions participating in OCLC. The record also includes pointers to entries in index files, so that the data base may be readily updated, and a pointer to the beginning of the list of holdings for the record. In addition, each record has a small directory for the construction of truncated author-title-date entries, which are displayed to allow a user to make a choice whenever a search key indexes two or more records.

Although each bibliographic record includes all information in a standard MARC II record, records in the bibliographic record file have been reduced to 78 percent of the size of the communication record largely by reducing redundancy in structural information. OCLC intends to compress bibliographic records further by reducing redundancy in text by employing compression techniques similar to those described in the literature (13,14).

The holdings file contains a string of holdings records for each bibliographic record; individual records are chained with pointers. Information in each record includes identity of the holding institution and the holding library within the institution, a list of each physical item of multiple or partial holdings, the call number and pointers to the next record in the chain, and to the call number index. The last record in the chain also has a back-pointer to the associated bibliographic record. Whenever there is a unique note, dash entry, or extra added entry coupled to a holding, that holding has a pointer to a location in the third sequential file in which the note or entry resides.

Index Files

Indexes include an author-title index, a title index, and an LC card number index. Research and development are under way leading to implementation of an author and added author index and a call number index. A class number index will be developed and implemented in the future.

With the exception of the class number index, which by its nature is required to be a sequentially accessible file, the OCLC indexes are scatter storage files. The construction of and access to a scatter storage file involves the calculation of a home address for the record and the resolution of the collisions that occur when two or more records have the same home address. The calculation of a home address comprises derivation of a search key from the record to be stored or retrieved and the hashing or randomizing of the key to obtain an integer, relative record address that is converted to a

storage home address. The findings of OCLC research on search keys has been reported (9,10,11).

The hashing procedure employs a pseudo-random number generator of the multiplicative type:

$$\text{Home Address} = \text{rem}(K x_n/m)$$

where K is the multiplier 65539, x_n is the binary numerical value of the search key, and m is the modulus which is set equal to the size of the index file; 'rem' denotes that only the remainder of the division on the right-hand side is used. Philip L. Long and his associates have shown that efficiency of a scatter storage file is rapidly degraded when the loading of the file exceeds 75 percent (12); therefore, OCLC initially loads files at 50 percent of physical capacity. Hence, the modulus is chosen to be twice the size of initial number of records to be loaded. When 75 percent occupancy is reached a new modulus is chosen and the file is regenerated.

Collisions are resolved using the quadratic residue search method proposed by A. C. Day (15) and shown to be efficient (12). In this method, a new location is calculated when the home address is full; the first new location has the value (home address − 2), the second (home address − 6), the third (home address − 12) and so on until an empty location is found if a record is being placed in the file, or the end of the entry chain is found if records are being retrieved. When the file size m is a prime having the form $4n + 3$, where n is an integer, the entire file may be examined by m searches.

Retrieval Techniques

The retrieval of a record or records from the OCLC data base is achieved in fractions of a second when a single request is put to the file, and rarely exceeds a second when queuing delays are introduced by simultaneous operation of upwards of 50 terminals. Response time at the terminal is greater than these figures because of the low communication line data rate, but terminal response time rarely exceeds five seconds.

Figure 6 shows the map of a record in the author-title index file and the title file. In the author-title file, the search key is a 3,3 key with the first trigram being the first three characters of the author entry and the second being the first three characters of the first word of the title that is not an English article (9). For example, "Str,Cha" is the search key for B. H. Streeter's *The Chained Library.* However, any or all of the characters in the trigrams may be all in lower case. The author-title index also indexes title-only entries, but the title index provides a more efficient access to this type of entry.

The pointer in the record map in Figure 6 is the address of the bibliographic record from which the search key was derived. The Entry Chain Indicator Bit is set to 0 (zero) if there is another record in the entry chain and to 1 if the record is last in the chain. When this bit is 0, the search skips to the next record as calculated by Day's skip algorithm. The

Bibliographic Record Presence Indicator Bit is set to 0 (zero) to indicate that the bibliographic record associated with this index entry has been deleted; it is set to 1 to indicate that the bibliographic record is present.

An author-title search of the data base is initiated by transmission of a 3,3 key from a terminal. A message parser analyzes the message and identifies it as a 3,3 author-title search key by the presence of the comma and by there not being more than three characters on either side of that comma. Next, the hashing algorithm calculates the home address and the location is checked for the presence of a record. If no record is present, a message is sent to the terminal stating that there is no entry for the key submitted and suggesting other action to be taken. If a record is present and its key matches the key submitted and if the entry-chain indicator bit signifies that the record at the home address is the only record in the chain, the bibliographic record which matches the key submitted is displayed on the terminal screen.

If the entry-chain bit signifies that there are additional records in the chain, those records are located by use of the skip algorithm. If more than one record possesses the same key as that submitted, truncated author-title-date entries derived from the matching bibliographic records are displayed with consecutive numbering on the terminal screen. The user then indicates by number the entry containing information pertaining to the desired work, and the program displays the full bibliographic record.

The title-index record has the same map as the author-title record and is depicted in Figure 6. The title index is also constructed and searched in

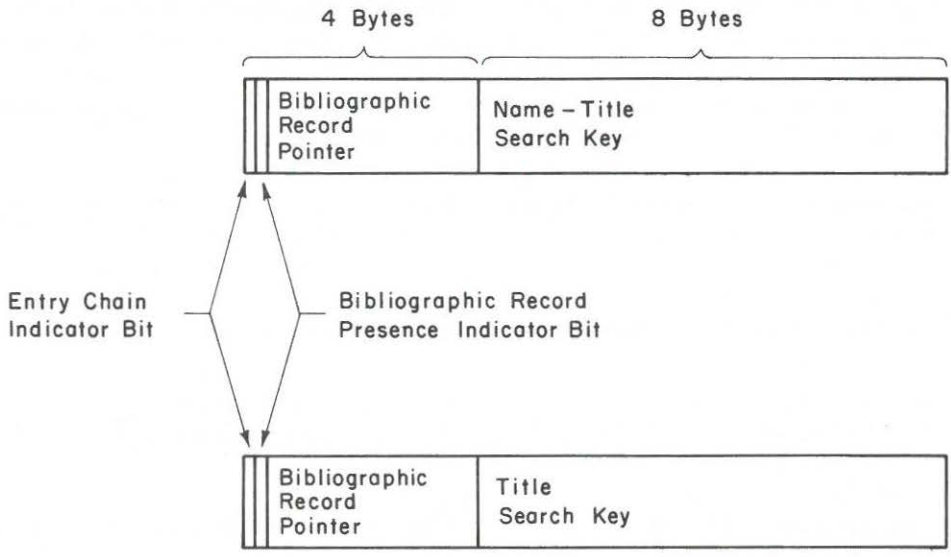

Fig. 6. Author-Title and Title Index Records.

the same manner as the author-title index. The title search key (3,1,1,1) consists of the first three characters of the first word of the title that is not an English article plus the initial character of each of the next three words. Commas separate the characters derived from each word. The title search key is "Cha,L,," for B. H. Streeter's *The Chained Library*, the three commas signifying that the message is a title search key. The bibliographic record pointer and the two indicator bits have the same function as in the author-title record.

Figure 7 exhibits the map for a record in the LC card number index. The three left-most bytes in the LC card number section contain an alphabetic prefix to a number where this is present, or, more usually, three blanks when there is no alphabetic prefix. Similarly the right-most byte contains a supplement number or is blank. The middle four bytes contain eight digits packed two digits to a byte after the digits to the right of the dash have been, when necessary, left-filled with zeroes to a total of six digits. The dash is then discarded. For example, LC card number 68-54216 would be 68054216 before being packed. The pointer and the two indicator bits have the same function as in the author-title index record.

An LC card number search is started with the transmission of an LC card number as the request. The parser identifies the message as an LC card number search by determining that there is a dash in the string of characters and that there are numeric characters in the two positions immediately to the left of the dash. The remainder of the search procedure duplicates that for the author-title index.

On-Line Programs

As is the case with all routinely used OCLC programs, the on-line programs are written in assembly language to achieve the utmost efficiency in processing. In addition, every effort has been made to design programs to run in the fastest possible time. In other words, one of the main goals of the OCLC on-line operation is lowest possible cost.

The simulation study had shown that it was necessary to modify the operating system of the XDS Sigma 5 so that the work area of the operating system would be identical with that of the applications programs. The XDS Real-time Batch Monitor, which is one of the operating systems furnished by XDS for the Sigma 5, has been extensively altered, and one of the alterations is the change to a single work area. Another major change to the operating system was building into it the capability for multi-programming. At the present time, the on-line foreground of the system operates two tasks in that two polling sequences are running simultaneously, and the background runs batch jobs at the same time. This new monitor is called the On-Line Bibliographic Monitor (OBM).

An extension of OBM is named MOTHERHOOD (MH); MH supervises the operation of the on-line programs. MH also keeps track of the activities of these programs and compiles statistics of these activities. In addition, MH

contains some utility programs such as the disk and terminal I/O routines.

The principal on-line application program is CATALOG (CAT); its functions are described in detail in the subsequent sections entitled *Cataloging with Existing Bibliographic Information* and *Input Cataloging*. In general, CAT accepts requests from terminals, parses them to identify the type of request, and then takes appropriate action. If a request is for a bibliographic record, CAT identifies it as such, and if there is only one bibliographic record in the reply, CAT formats the record in one of its work area buffers and sends the formatted record to the terminal for display. If more than one record is in the reply, CAT formats truncated records and puts them out for display. After a single bibliographic record has been displayed, CAT modifies the computer memory image of the record in accordance with update requests from the terminal. For example, fields such as edition statement or subject headings may be deleted or altered, and new fields may be added. When the request is received from the terminal to produce catalog cards from the record as revised or unrevised, CAT writes the current computer memory image of the record onto a tape to be used as input to the catalog card production programs.

The catalog card production programs operate off-line, and the first processing program is CONVERT (CNVT), which formats some of the fields and call numbers. The major activity of CNVT is the latter, for libraries require a vast number of options to set up their call numbers for printing. CNVT also automatically places symbols used to indicate oversized books above, below, or within call numbers as required.

FORMAT is the second program; it receives partially formatted records from CNVT. FORMAT expands each record into the total number of card images corresponding to the total cards required by the requesting library

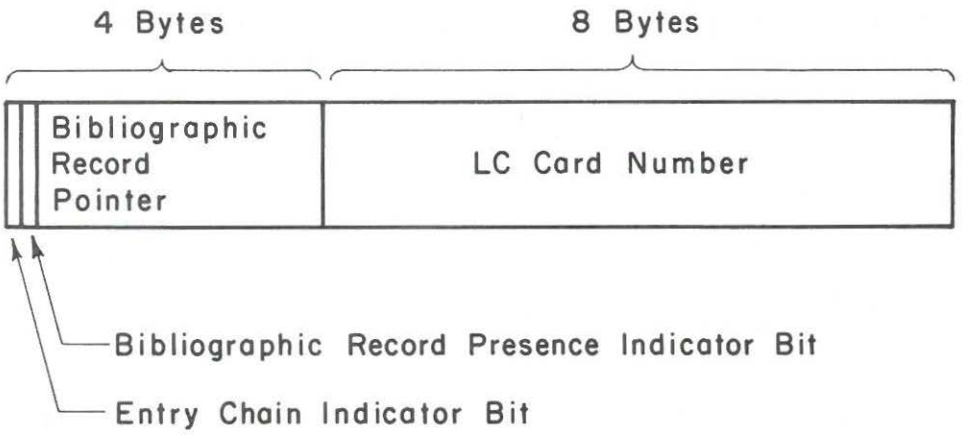

Fig. 7. Library of Congress Card Number Index Record.

for each particular title. FORMAT determines this total from the number of tracings and pack definition tables previously submitted by the library that define the printing of formats of cards to go into each catalog.

FORMAT, which is an extensive revision of EXPAND, contains many options not found in the old off-line catalog card production system. FORMAT can set up a contents note on any particular card, and puts tracings at the bottom of a card when tracings are requested. The author entry normally occurs on the third line, but if a subject heading or added entry is two or more lines long, FORMAT moves the author entry down on the card so that a blank line separates the added entry from the author entry. In other words, each card is formatted individually.

The major benefit of this feature, which allows the body of the catalog data to float up and down the card, is that the text on most cards can start high up on the card, thereby reducing the number of extension cards. The omission of tracings from added entry cards has a similar effect. Table 1 presents the percentage of extension cards in a sample of 126,738 OCLC cards for 18,182 titles produced for twenty-five or more libraries during a seventeen-day period, compared with extension cards in Library of Congress printed cards and in a sample of NELINET cards "for over 1,300 titles" (16). The table shows that the OCLC mixture of cards with and without tracings and with the floating body of text yields about 10.8 percent more extension cards compared to Library of Congress printed cards. Were libraries to restore the original meaning to the phrase "main entry" by printing tracings only on main entry cards, the percentage of extension cards in computer produced catalog cards printed six lines to the inch would probably be less than for LC cards.

FORMAT also sets up a sort key for each record and a sort program sorts the card images by institution, library, catalog, and by entry or call number within each catalog pack. Another program, BUILD-PRINT-TAPE (BPT), arranges the sorted images on tape so that cards are printed in consecutive order in two columns on two-up card stock. Finally, a PRINT program prints the cards on an IBM 1403 N1 Printer attached to an IBM 360/50 computer.

Cataloging With Existing Bibliographic Information

This section describes cataloging using a bibliographic record already in the central file; the next section, entitled *Input Cataloging* describes cataloging when there is no record in the system for the item being cataloged.

The cataloger at the terminal first searches for an existing record, using the LC card number found on the verso of the title page or elsewhere. If the response is negative or if there is no card number available, the cataloger searches by title or by author and title using the 3,1,1,1 or 3,3 search keys respectively. If these searches are unproductive, the cataloger does input cataloging.

When a search does produce a record, the cataloger reviews the record

Table 1. Extension Catalog Card Percentages

Number of Cards	*OCLC MARC II Cards*	*Library of Congress Printed Cards*	*NELINET MARC II Cards*
1	77.2	87.8	79.9
2	18.9	10.0	16.7
3	2.5	1.6	2.5
4	1.1	.3	.6
5	.2	.2	.1
6	—	.1	.2

to see if it correctly describes the book at hand. If it is the correct record and if the library uses Library of Congress call numbers, the cataloger transmits a request for card production by depressing two keys on the keyboard. Cataloging is then complete. If the LC call number is not used, the cataloger constructs and keys in a new number and then transmits the produce-cards request.

If the record does not describe the book as the cataloger wishes, the record may be edited. The cataloger may remove a field or element, such as a subject heading. Information within a field may be changed by replacing existing characters, such as changing an imprint date by overtyping, by inserting characters, or by deleting characters. Finally, a new field such as an additional subject heading may be added. When the editing process is complete, the cataloger can request that the record on the screen be reformatted according to the alterations. Having reviewed the reformatted version, the cataloger may proceed to card production.

When a cataloger has edited a record for card production, the alterations in the record are not made in the record in the bibliographic record file. Rather, the changes are made only in the version of the record that is to be used for card production. The edited version of the record is retained in an archive file after catalog card production so that cards may be produced again from the same record for the same library, should the need arise in the future.

The author index currently under development will enable a cataloger to determine the titles of works in the file by a given author. The call number index, also currently being developed, will make it possible for a cataloger to determine whether or not a call number has been used before in his library. The class number index that will be developed in the future will provide the capability of determining titles that have recently been placed under a given class number or, if none is under the number, the class number and titles on each side of the given number.

Input Cataloging

Input cataloging is undertaken when there is no bibliographic record in the file for the book at hand. To do input cataloging, the cataloger requests

that a work form be displayed on the screen (Figure 8). The cataloger then proceeds to fill in the work form by keyboarding the catalog data, and transmitting the data to the computer field by field as each is completed. As shown in Figure 8, a paragraph mark terminates each field; each dash is to be filled in by the cataloger for each field used. Input cataloging may be original cataloging or may use cataloging data obtained from some source other than the OCLC system.

```
Type:                    Lang:
Form:                    ISBN
Intel lvl:               Card No:
Bibl lvl: ¶
¶
▷   1  1-- --      d ¶
▷   2  24- --      b c ¶
▷   3  250          ¶
▷   4  260 -       b c ¶
▷   5  300         b c ¶
▷   6  4-- --      d ¶
▷   7  5-- -       ¶
▷   8  6-- --      ¶
▷   9  7-- --      d ¶
▷  10  8-- -       ¶
▷  11  092         b ¶
▷  12  049 --      ¶
▷  13  590         ¶
```

Fig. 8. Workform for a Dewey Library.

When the catalog data has been input, revised, and correctly displayed on the terminal screen, the cataloger requests catalog card production. In the case of new cataloging, not only are cards produced, but also the new record is added to the file and indexed so that it is available within seconds to other users. If a MARC II record for the same book is subsequently added to the file, it replaces the input-cataloging record but does not disturb the holdings information.

Union Catalog

Each display of a bibliographic record contains a list of symbols for those member institutions that possess the title. In other words, the central file is also a union catalog of the holdings of OCLC member libraries, although in the early months of operation these holdings data are very incomplete. Nevertheless, they will approach completeness with the passage of time and with retrospective conversion of catalog data. Titles cataloged during the operation of the off-line system have been included in the union catalog.

The union catalog function is an important function of the shared cataloging system, for it makes available to students and faculties, through the increased information available to staff members, the resources of academic institutions throughout Ohio.

Libraries also use the union catalog as a selection tool since they can dispense with expensive purchases of little-used materials residing in a neighboring library. Members also use the file to obtain bibliographic data to be used in ordering.

Assessment

With over nine hundred thousand holdings recorded in the union catalog as of April 1973, it is clear that having this type of information immediately at hand will greatly improve services to students and faculties. Enlargement of holdings recorded will enhance the union-catalog value of the system. Wright State University is in process of converting its holdings using the OCLC system, and the Ohio State University Libraries—the largest collection in the state—has already converted its shelf list in truncated form. The OSU holdings information will soon be available to OCLC members.

Members using the OCLC system report a large reduction in cataloging effort. Two libraries using LC classification report that they are cataloging at a rate in excess of ten titles per terminal hour when cataloging already exists in the system. Libraries using Dewey classification are experiencing a somewhat lower rate.

The original cost benefit studies were done on the basis of a calculated rate of six titles per hour for those books for which there were already cataloging data in the system. The net savings will be realized when the file has reached sufficient size to enable the largest libraries to locate records for 65 percent of their cataloging and for the smallest to find 95 percent. To reach this level, members collectively would have to use

existing bibliographic information to catalog 350,000 titles in the course of a year, or an average of approximately 1,460 titles for the total system per working day. It was thought that this rate would be attained by the end of the second year of operation. However, at the end of the first month of on-line operation, over a thousand titles per day were being cataloged.

The new catalog card production programs operating on the Sigma 5 are much more efficient than the programs used in the older off-line system. Earlier in this paper it was reported that cost of the older programs to format catalog cards, but not to print them, was 2.27 cents per card. If costs of the Sigma 5 are calculated at commercial rates, the new programs format cards at 2.21 cents per card. However, if actual costs to OCLC are used and with the total cost being assigned to one shift, the cost of formatting each card becomes 0.86 cents. The total cost of producing catalog cards is, of course, much more than the cost to format them on a computer. Nevertheless, either the 2.21 cents or 0.86 cents rate might serve as a criterion for measuring the efficiency of computerized catalog card production.

The low terminal response-time delay for the operation of seventy terminals is a good gauge of the efficiency of the on-line system. In particular, the file organization is efficient, for it enables retrieval of a single entry swiftly from a file of over 600,000 records. Moreover, no serious degradation in retrieval efficiency is expected to arise as the result of the growth of file size.

The system operates from 7:00 A.M. to 7:00 P.M. on Mondays through Fridays, and at times the interval between system downtimes has exceeded a week. It is rare that the system will be down on successive days, and when a problem does occur, the system can be restored within a minute or two. Moreover, when the system goes down, only two terminals will occasionally lose data, and most of the time, there is no loss of data. Hence, it can be concluded that the hardware and software are highly reliable.

In summary, it can be said that the OCLC on-line shared cataloging system is easy to use, efficient, reliable, and cost beneficial.

ACKNOWLEDGMENTS

The research and development reported in this paper were partially supported by Office of Education Contract No. OEC-0-70-2209 (506), Council on Library Resources Grant No. CLR-489, National Agricultural Library Contract No. 12-03-01-5-70, and an L.S.C.A. Title III Grant from the Ohio State Library Board.

REFERENCES

1. Ellen Wasby Miller and B. J. Hodges, "Shawnee Mission's On-Line Cataloging System," *JOLA* 4:13-26 (March 1971).

2. John P. Kennedy, "A Local MARC Project: The Georgia Tech Library," in *Proceedings of the 1968 Clinic on Library Applications of Data Processing.* (Urbana, Ill.: University of Illinois Graduate School of Library Science, 1969) p. 199–215.
3. New England Board of Higher Education, *New England Library Information Network; Final Report on Council on Library Resources Grant #443.* (Feb. 1970).
4. Charles T. Payne and Robert S. McGee, *The University of Chicago Bibliographic Data Processing System: Documentation and Report Supplement,* (Chicago, Ill.: University of Chicago Library, April 1971).
5. Judith Hopkins, *Manual for OCLC Catalog Card Production* (Feb. 1971).
6. Ohio College Library Center, *Preliminary Description of Catalog Cards Produced from MARC II Data* (Sept. 1969).
7. F. G. Kilgour, "Libraries—Evolving, Computerizing, Personalizing," *American Libraries* 3:141-47 (Feb. 1972).
8. American National Standards Institute, *American National Standard for Bibliographic Information Interchange on Magnetic Tape* (New York: American National Standards Institute, 1971).
9. F. G. Kilgour, P. L. Long, and E. B. Leiderman, "Retrieval of Bibliographic Entries from a Name-Title Catalog by Use of Truncated Search Keys," *Proceedings of the American Society for Information Science* 7:79-82 (1970).
10. F. G. Kilgour, P. L. Long, E. B. Leiderman, and A. L. Landgraf, "Title-Only Entries Retrieved by Use of Truncated Search Keys," *JOLA* 4: 207-210 (Dec. 1971).
11. Philip L. Long, and F. G. Kilgour, "A Truncated Search Key Title Index," *JOLA* 5:17-20 (Mar. 1972).
12. P. L. Long, K. B. L. Rastogi, J. E. Rush, and J. A. Wyckoff, "Large On-Line Files of Bibliographic Data: An Efficient Design and a Mathematical Predictor of Retrieval Behavior." *IFIP Congress '71: Ljubljana –August 1971.* (Amsterdam, North Holland Publishing Co., 1971). Booklet TA-3, 145-149.
13. Martin Snyderman and Bernard Hunt, "The Myriad Virtues of Text Compaction," *Datamation* 16:36-40 (Dec. 1970).
14. W. D. Schieber and G. W. Thomas, "An Algorithm for Compaction of Alphanumeric Data," *JOLA* 4:198-206 (Dec. 1970).
15. A. C. Day, "Full Table Quadratic Searching for Scatter Storage," *Communications of the ACM* 13:481 (Aug. 1970).
16. New England Board of Higher Education, *New England Library Information* . . . , p. 100-101.