

AN INTEGRATED TECHNICAL PROCESSING SYSTEM FOR A LARGE
RESEARCH LIBRARY

James S. Aagaard
Northwestern University
Evanston, Illinois

Emphasis in the design of the Northwestern University on-line technical processing system was on economical operation and integration of as many functions as possible. The present system accomodates ordering, cataloging, serial holdings, and the production of book marking materials, with provisions for the addition of accounting functions.

For nearly three years the Northwestern University Library has been operating an automated technical processing system. Recently, planning has started for extensive changes to the system, and so it seems an appropriate time to report on our experiences. These observations may be of value in helping others who are designing similar systems.

One point should be emphasized at the very beginning of this report. That is, our system was designed with the needs of a large research library in mind; a library which must deal with a substantial quantity of foreign language materials, antiquarian purchases, and specialized collections and branches. We believe the state of the art in library automation is a long way from producing a sufficiently general system that will serve adequately the needs of a small public library, a large university library, and everything in between or beyond.

As described in a previous report (1) we were concerned with developing an economical system. Northwestern University supports the entire cost of system development as well as operational costs, and it was vital that the computer requirements be kept to a minimum. These considerations led to the concept of using a single file to hold a combined record which would serve all of the needs of technical processing (this is in addition to the circulation file described in the earlier report). Because the mass storage device available to us was an IBM 2321 Data Cell, which can have an access time to a record as long as a second, the use of elaborate indexing schemes did not seem feasible. However, in order to use the system for serial check-in, it was essential that access to each record be possible without the necessity for consulting some kind of a printed index.

These considerations led to the basic design of a direct-access file, wherein each record is assigned a key derived from the author and title of the work. A hashing algorithm is applied to the key and the result is used to locate the record within the file. Appropriate provision must be made to handle overflow conditions, of course.

The record itself is basically a reformatted MARC record with the discard of most of the fixed field information and with the addition of local ordering, payment, and cataloging fields. The original development

of the system coincided with the announcement of the MARC II format, and fortunately it was possible to include many of these improvements in our system. One of the major difficulties in processing bibliographic data on a computer is related to the extreme variability in record size. The standard file management techniques which the computer manufacturer supplies do not provide much help in this respect, particularly when random access and the capability to change the length of an existing record are required. To provide a space large enough for the longest possible record is extremely wasteful of expensive storage space, so the alternative of breaking up a record into fixed-length segments was chosen. A segment size of approximately 500 characters was selected as a reasonable compromise. Too small a size wastes space between blocks, and also in repeating the key for each segment, while too large a size wastes space in partially filled segments.

Record Key

Because of the great amount of attention which the OCLC system has received in the library community, the use of an author-title key to access a bibliographic record has become widely understood. However, unlike the OCLC system, which provides several keys of different construction (2) our key provides the only access to a record, and thus must be more nearly unique than those used by OCLC.

After trying a number of different ideas on a quantity of bibliographic records, a 13-character key finally evolved. This key is composed of four segments; the first one being four characters and the other segments three characters each. The basic idea was that the first segment would be derived from the main entry, and the remaining three segments from the title. It soon became apparent that corporate main entries were a special problem, and the rules for generating the key were modified for such cases. There is, of course, a conflict between the attempt to get a unique key and the desire to keep the rules simple enough so that both human and computer can apply them readily. The resulting compromise is still too complicated to be described completely here, but a simplified statement and some examples will be presented.

For a personal name main entry, the first segment of the key is

obtained by taking the first three letters of the author's last name plus his first initial. For a corporate main entry, the first one or two segments of the key are obtained by using the first letter of each word in the name. One segment is used if there are five or fewer words; two segments if there are more than five words. The remaining segments of the key are obtained by taking the first three letters of each of the first two or three words of the title. In each of these cases only words of four or more letters are used. To aid recognition, the parts of the key derived from the main entry are written in upper case letters; the parts derived from the title are written in lower case letters. The examples in Figure 1 show application of these rules as well as exceptions which are not covered by the simplified version given above.

Even the most elaborate rules will not provide uniqueness when confronted with two periodicals with the same title (we have 17 different ones named Africa, at last report), or with new editions of the same work. Therefore, provision was made for a two-digit suffix to the key which is manually assigned. When such duplicates are encountered, inspection of each record is necessary to determine which is the correct one.

Batch Processing

Although the system is an on-line system, with real time access to records for display, addition, and update, certain of the needs of technical processing were more readily met in a batch mode. Items produced in this way are catalog cards, purchase orders, temporary catalog slips, book pocket labels, punched cards for the circulation control system, claim notices, and work sheets. The work sheets, which are simply a printout of an entire record, were necessitated by the fact that no display terminals with lower case capability were available to us, forcing the use of typewriter terminals. These are much too slow (13 characters per second) for the display of an appreciable number of records. Thus, when a display of a record is not needed immediately, a work sheet is requested.

All of the batch products except the permanent catalog cards are produced nightly and are available to the library at the beginning of the next working day. The catalog cards are presorted for filing, and

it is more efficient if they are produced less often, in larger batches. To request a batch product, a terminal operator enters a one-character request on the terminal, which causes a copy of the complete record, with the request code, to be written in a sequential "transaction file". The batch programs scan this file, searching for records with a particular code, and process the records appropriately. For example, if the operator enters the letter "P" as a request, the purchase order program will recognize this code and produce a purchase order from the record in the transaction file.

Use of MARC

It was clear during the initial development phase of our system that some method should be provided to make use of the MARC records which were beginning to come from the Library of Congress, even though the restriction to English-language imprints meant that MARC would supply only a fraction of our needs. The cost of on-line storage made it impractical to keep any substantial part of the MARC file available for searching from the remote terminals. The method which finally evolved was a combination of a batch search, requested by the operator when a new record is created, plus a monthly automatic review to capture records which have arrived since the original request was made.

The Data Cell was retired nearly two years ago in favor of disk units with a total capacity of about 60 million characters of storage. One such disk will hold nearly 100,000 MARC records plus an associated index. This disk represents the most current file, and is updated weekly as the tapes are received. Approximately every six months the oldest records are removed and transferred to magnetic tape.

The search request is treated in much the same way as other batch requests. The batch processing program extracts (if present) the LC card number and the International Standard Book Number from the provisional record entered by the terminal operator. These quantities, plus the search key, are used to search the MARC index after the disk has been mounted on the computer. When a match is obtained on the ISBN or LC card number, the MARC record is written on a temporary disk file. The teleprocessing program is then made to think that the temporary disk file is actually a

remote terminal, and the MARC record is used to update the provisional on-line record. However, if the match is obtained only on the search key, the record is printed instead of transferred to the processing file, so that a manual verification can be made that the correct record has been found. This procedure occasionally leaves a record in the processing file with more than one title or main entry; the unwanted ones are removed during the cataloging process.

On a monthly basis, those records which are still waiting for MARC data (the book itself either has not been received or is in the cataloging backlog) are searched first against the same current disk file, and then those less than a month old are searched against approximately an equal number of older MARC records on tape. This covers a total period of about three years; further searching against older files does not seem to turn up enough records to justify the time involved. If the work does not fall within the scope of the MARC distribution, cataloging data is entered through a terminal.

Data Integrity

One of the reasons for the complexity of many on-line computer systems is the need to provide for the recovery of the data base in case of abnormal conditions. In our system this need is minimized since there are no links or pointers which could be left set incorrectly. Successful completion of an update operation is acknowledged to a remote terminal only after the record has been rewritten in the file; if the system fails before this acknowledgement is received, it is the operator's responsibility to verify the status of the record at the next opportunity.

With any system there is always the possibility of a disaster wiping out a large part of the file. Precaution against this requires the frequent copying of the data to tape and storage of the tape at a location physically removed from the computer. Complete protection would also require the separate recording of all transactions which occur between copies, but the likelihood of such a catastrophe is small, and most of the changes could probably be reconstructed from paper copies in existence. In any case, the restoration of the entire file would be done only as a last resort after all attempts to restore only the damaged part had failed.

Costs

Although the problem of allocating costs among various systems sharing a common computer is difficult, our best estimate of the computer costs per year attributable to technical processing is as follows:

Batch processing time charges	\$23,000
Teleprocessing charges	40,000
Terminal equipment (9)	17,000
Total	\$80,000

The major cost item, teleprocessing charges, would be significantly lower if there were other University users of on-line systems, so that the Library would not have to bear the entire cost of the additional computer equipment which is needed.

It is also difficult to determine the net effect on personnel requirements, since there have been significant changes in the duties of many people. We estimate that one nonprofessional position has been eliminated, while the volume of work has increased because of an increase in the Library's book budget of about \$100,000 per year.

Approximately 50,000 new volumes per year are processed by the system. Of these about 35,000 are new titles and the remainder are added copies or bound serial volumes. Many items are received on standing order, although the computer system produces about 22,000 purchase orders per year. At any time there are 60,000 to 70,000 monograph records in the on-line file for books in various stages of processing, and about 35,000 serial records, 20,000 of which are considered to be active. The computer produces over 300,000 catalog cards in a year, as well as about 40,000 temporary slips which are placed in the catalog at the time an item is received.

Future Improvements

Northwestern's on-line system has worked very well for most purposes. Perhaps the most significant weakness is with serial check-in. It was originally planned that the system would be used for check-in of individual issues. This was done for a number of months simultaneously with the conversion of existing files to computer records. However, at about the same time as all the serial records were converted, we experienced a temporary shortage of staff. As a result the check-in process fell behind. Once this happened it became very difficult to catch up, and finally the on-line check in of individual issues had to be abandoned until some system improvements could be made. (Serials are temporarily being recorded when they are ready for binding.)

There were two main problems which caused the check-in of an issue to require too much time. The first of these was the difficulty in locating some records. The only access to the file is by the author-title key, which is derived from the cataloging data. It may bear little resemblance to what appears on the cover of a piece to be checked in. The obvious solution to this problem is to have more and better access points, which will probably require indexes.

The second problem involves the slowness of the typewriter terminals. It is usually necessary to display several lines of the record to verify that the correct record has been found, and also to display some existing holdings information to check for missing issues. Then the most recent holdings line must be completely retyped to add the issue in hand. This problem will be alleviated by a change to cathode-ray display terminals.

Another significant improvement which is planned is a change from a single general-purpose file maintenance program to a collection of specialized programs for specific functions such as bibliographic information update, ordering, check-in and accounting. This approach will facilitate using the full capabilities of newer display terminals such as screen formatting and protected fields. It will also enable a much higher level of on-line validity checking of data. The validity checking capabilities become of greater importance as the use of the terminals is extended to applications other than the entering of bibliographic data. Additional flexibility will be possible due to the removal of restrictions imposed by the Data Cell.

As suggested above, the plan for the system includes provision for processing invoices and for fund accounting. These modules were nearly ready for implementation a year ago, but the decision was made to postpone the implementation and concentrate on the basic redesign of the entire system. We expect to continue following our philosophy of developing compact, efficient programs which can provide all of the functions needed by our library without incurring the excessive overhead of extremely sophisticated file management or text editing systems. However, we do plan to keep in mind the possible affiliation with a network, which could provide needed on-line access to a much larger bibliographic data base than a single institution can afford.

Acknowledgements

Overall systems design, and the development of the search key format, were done by Velma Veneziano. The suggestions and encouragement from the staff of the Technical Services Department, library administration, and the director of administrative data processing have contributed greatly to the success of this project.

References

1. J.S. Aagaard: "An Interactive Computer-Based Circulation System: Design and Development" Journal of Library Automation 5 (March 1972) 3-11.
2. F.G. Kilgour, P.L. Long, A.L. Landgraf, and J.A. Wyckoff: "The Shared Cataloging System of the Ohio College Library Center" Journal of Library Automation 5 (September 1972) 157-183.

GAUP aar gre mou	Gauch, Patricia Lee Aaron and the Green Mountain Boys.
TEST aba hom es	Tessier, Thomas Abandoned Homes.
CBEA PPR abs pap	Conference on Biochemical and Ecological Aspects of Plant Parasite Relations, Budapest, 1970. Abstracts of papers presented at the Conference on Biochemical and Ecological Aspects of Plant-Parasite Relations.
aca sup mar ket	Academic super markets.
aac rao sur gra	The AACRAO survey of grading policies in member institutions.

Figure 1. Examples of search keys