FILE SIZE AND THE COST OF PROCESSING MARC RECORDS

John P. KENNEDY: Data Processing Librarian, Georgia Institute of Technology, Atlanta, Georgia

Many systems being developed for utilizing MARC records in acquisitions and cataloging operations depend on the selection of records from a cumulative tape file. Analysis of cost data accumulated during two years' experience in using MARC records for the production of catalog cards at the Georgia Tech Library indicates that the ratio of titles selected to titles read from the cumulative file is the most significant determinant of cost. This implies that the number of passes of the file must be minimized and an effective formula for limiting the growth of the file must be developed in the design of an economical system.

Since 1963 several articles on computerized production of catalog cards have reported cost figures for card production. Fasana reported a cost per card of 9.9 cents at the Air Force Cambridge Research Laboratory (AFCRL) (1). Costs at the Yale Medical Library under the Columbia-Harvard-Yale computerized card production system varied from 8.8 cents to 9.8 cents per card (2). Under the Yale Bibliographic System, costs for card production at the Yale Medical Library have been 13.9 cents per card. When the MARC MATE program is used to introduce MARC records into the Yale Bibliographic System the cost of cards produced from the MARC records is 24.9 cents (3). Costs for computer assisted card production at the Philip Morris Research Library have been estimated at 18 cents per card (4). The cost per card for cards produced from MARC records at the Georgia Institute of Technology Library has been reported as 10 cents (5).

The focus of interest in these cost reports has been on a comparison of the costs of computer produced cards and manually produced cards. There is agreement in these reports that computer production can compete favorably in terms of cost with other methods of production. Less attention has been given to variations in the costs of computer produced cards. Since the systems for which costs have been reported vary in scope and objectives, equipment used, nature of input, rates for labor, and charges for computer time, it is not very useful to compare the costs from system to system. Variations in cost within one system are of greater interest, since it is easier to isolate the factors that result in the altered costs. The report on the Yale bibliographic system shows that the introduction of MARC records into a system that was not designed for processing MARC records may produce substantially higher costs. Fasana reported that when a PDP-1 computer was used rather than the specially built Crossfiler in the AFCRL system, the cost per card was quadrupled. Kilgour discusses briefly the effects of three changes in the Columbia-Harvard-Yale system on the cost of cards produced.

The 10-cent-per-card cost reported for Georgia Tech was the average cost during the preceding three-month period, January through March 1968. During the three years in which catalog cards have been produced on the computer at Georgia Tech, costs have varied widely as procedures, personnel, file sizes and work loads have changed. The greatest variation has occurred in the cost of the manual steps in the system, mainly proofreading and making corrections. The greatly improved accuracy of the MARC II records has resulted in a reduction in the time required for proofreading and making corrections. The costs of supplies and equipment have been small and shown little variation. The cost of computer time has varied from 18 cents per title (just over 2 cents per card) to a high of 47 cents (6 cents per card), excluding the cost of the merge runs to maintain a cumulative file of MARC records. An analysis has been made to determine the factors responsible for this variation in computer costs, and techniques for reducing computer costs have been developed.

MATERIALS AND METHODS

The Price Gilbert Memorial Library at the Georgia Institute of Technology is a centralized scientific, technical and management collection of 612,000 volumes plus 500,000 microtext and other bibliographic units. In 1968/69 almost 20,000 titles representing about 35,000 volumes were cataloged for addition to the collection. The Library makes use of the UNIVAC 1108 and the Burroughs B5500 computing systems of the Institute's Rich Electronic Computing Center for its data processing needs. The work described here was performed on the B5500. The Georgia Tech B5500 configuration includes two central processing units, 32,000 forty-eight-bit words of core storage, 29 million characters of disc storage and 10 magnetic tape drives. Library programs are written in COBOL and are

multi-processed with other programs in the standard work stream. The Library is billed $140 per hour for central processor time and $47 per hour for IO channel time. The system for production of catalog cards from MARC I records which was in operation for over two years has been described previously (6).

Statistics were recorded for all computer runs in the processing of 73 batches of MARC I titles. These statistics include number of records processed, file sizes, processor time, IO channel time, and cost, for each run. The time and cost remained fairly constant for some runs. The cost of runs to produce the sorted catalog cards from edited MARC records ranged from 6 to 9 cents per title and averaged a little over 7 cents. The cost of runs to make changes and additions to the MARC records ranged from 1 to 5 cents per title and averaged 2 cents. The cost was usually about 1 cent per title for each time the correction program was run. It often had to be rerun several times before all records in the batch were correct. The Library's improved MARC II system avoids the cost of correction reruns by permitting independent corrections to any record in a direct access file rather than requiring records to be processed as a batch.

Most of the variation in the cost of computer time occurred in the run in which records were selected from the cumulative MARC file and the selected records were then converted to the B5500 character codes, reformatted and prooflisted. The cost of this run varied from a low of 10 cents per title selected to a high of 36 cents per title; the variation is primarily an effect of the increasing size of the cumulative MARC file and of variation in the number of titles selected in the run. As the MARC file increased in size the cost of selecting a small number of titles increased dramatically. The precise relationship of file size and batch size to cost per title is not apparent, however, because the cost of character conversion, reformatting, and printing the prooflist were combined with the cost of selection in a single run. An additional complication results from the effects of the other jobs being processed by the computer concurrently. For example, one batch which had to be rerun because the output tape was defective cost 23 cents per title the first time and 28 cents per title when rerun with a different job mix.

Although the part of the run cost which can be attributed to passing the MARC file and the part attributable to code conversion, formatting and printing cannot be determined for a single run, this can be calculated from a number of runs with varying file sizes and batch sizes. It is assumed that variations in the time required for processing individual records of varying lengths and variations due to the mix of jobs run concurrently will average out and may be disregarded. Statistics for the selection runs include the number of records read from the cumulative MARC file, the number of records selected and processed, the processor time and IO channel time required for the run, and the cost of the run. Using the method of the least squares, these statistics were used to calculate the

average time and cost for each record read from the cumulative MARC file. Once these constants are calculated it is possible to predict the cost per item or the total cost of a select run with any given file size and batch size.

In order to determine the average cost for processing a selected record and the average cost for reading a record from the cumulative MARC file, it was postulated that

$$C_T = \left(\frac{FS}{BS}\right) C_R + C_P$$

where

- C_T is the total cost per title
- FS (File Size) is the number of records read from the cumulative MARC file
- BS (Batch Size) is the number of records selected in the run
- C_R is the cost of reading a record from the cumulative MARC file
- C_P is the cost for processing a selected record

The method of least squares yields the following equations:

$$\left[\Sigma\left(\frac{FS}{BS}\right)^2\right] C_R + \left[\Sigma\left(\frac{FS}{BS}\right)\right] C_P = \Sigma\left(\frac{FS}{BS}\right) C_T$$

and

$$\left[\Sigma\left(\frac{FS}{BS}\right]\right. C_R + NC_P = C_T$$

Solving these equations for the data from the 73-batch sample gives the following values:

$C_P = \$.073$
$C_R = \$.00068$

Since charges for computer time are determined differently at other installations, the figures for processor time and IO channel time may be more useful to others than the cost figures. Using the same techniques but substituting processor time for cost gives the following values:

Processor time per record read = .00646 seconds

Processor time for selected records = 1.339 seconds

Again, using the same technique but substituting IO channel time for cost gives the following values:

IO channel time per record read = .02048 seconds

IO channel time for selected records = .456 seconds

These values may be substituted in the formula, $C_T = \left(\frac{FS}{BS}\right) C_R + C_P$, to find the cost or time per title for any batch and file size. For example, the per title cost for selecting and processing a batch of 200 records from a MARC file of 40,000 records:

$$C_T = \left(\frac{FS}{BS}\right) C_R + C_P$$

$$C_T = \left(\frac{40000}{200}\right) (\$.00068) + \$.073$$

$$C_T = \$.21$$

It will cost about twenty-one cents per title. The total cost of the run can be predicted as follows:

$$C = (FS - BS)\ (C_R) + (BS)\ (C_P)$$

$$C = (40000 - 200)\ (\$.00066) + (200)\ (\$.073)$$

$$C = \$41.27$$

RESULTS

Table 1 shows the predicted cost per title for various file sizes and batch sizes; it is based on the cost of the select run at Georgia Tech and ignores the cost of maintaining the MARC file. Since the Library of Congress cumulated MARC I records until a reel of tape was filled and provided a cumulative card number listing of the records on the reel, it was not essential to update the cumulative MARC file each week. The MARC II tapes issued from the MARC Distribution Service are not cumulative. Most libraries maintaining a cumulative file of MARC records will find it necessary to update this file each week. Weekly updating of the MARC file requires that all records on the file be not only read but also written on a new tape each week. For most systems this will rapidly become the most expensive machine procedure in the entire system. Combining the selection function and any index production with the file update means that no additional passes of the file will be required, but the cost of writing the file each week must be added to the figures in Table 1. Statistics from the merge runs at Tech show that if the number of old MARC file records read, the number of records read from the weekly update tape, and the number of records written on the new MARC file are totaled, the average cost per IO operation for the merge runs ranged between $.00062 and $.00073 and averaged $.00068 for all merge runs. Since this is the same cost as that obtained for each record read from the cumulative file in the select runs, it seems reasonable to use this figure as the cost for reading or writing a MARC record in calculating the cost of

Table 1. Relationship of File Size and Batch Size to Cost per Title

File Size	BATCH SIZE 50	100	150	200	250	300	400	500	750	1000
10K	$ .209	$ .141	$ .118	$ .107	$ .100	$ .095	$ .090	$ .087	$ .082	$ .080
20K	.345	.209	.164	.141	.127	.118	.107	.100	.091	.087
30K	.481	.277	.209	.175	.155	.141	.124	.114	.100	.093
40K	.617	.345	.254	.209	.182	.164	.141	.127	.109	.100
50K	.753	.413	.300	.243	.209	.186	.158	.141	.118	.107
60K	.889	.481	.345	.277	.236	.209	.175	.155	.127	.114
70K	1.025	.549	.390	.311	.263	.232	.192	.168	.137	.121
80K	1.161	.617	.436	.345	.291	.254	.209	.182	.146	.127
90K	1.297	.685	.481	.379	.318	.277	.226	.194	.155	.134
100K	1.433	.753	.526	.413	.345	.300	.243	.209	.164	.141
110K	1.569	.821	.572	.447	.372	.322	.260	.223	.173	.148
120K	1.705	.889	.617	.481	.399	.345	.277	.236	.182	.155

Table 2. Relationship of File Size and Batch Size to Cost per Title — File Update and Record Selection Functions Combined in Same Program

Old File Size	BATCH SIZE									
	50	*100*	*150*	*200*	*250*	*300*	*400*	*500*	*750*	*1000*
10K	$ .378	$ .225	$ .175	$ .149	$ .134	$ .124	$ .111	$ .104	$ .093	$ .088
20K	.650	.361	.265	.217	.188	.169	.145	.131	.111	.102
30K	.922	.497	.356	.285	.243	.214	.179	.158	.130	.115
40K	1.194	.633	.447	.353	.297	.260	.213	.185	.148	.129
50K	1.466	.769	.537	.421	.352	.305	.247	.212	.166	.143
60K	1.738	.905	.628	.489	.406	.350	.281	.240	.184	.156
70K	2.010	1.041	.719	.557	.461	.396	.315	.267	.202	.170
80K	2.282	1.177	8.09	.625	.515	.441	.349	.294	.220	.183
90K	2.554	1.313	.900	.693	.569	.486	.383	.321	.238	.197
100K	2.826	1.449	.991	.761	.624	.532	.417	.348	.257	.211
110K	3.098	1.585	1.081	.829	.678	.577	.451	.376	.275	.224
120K	3.370	1.721	1.172	.897	.732	.622	.485	.403	.293	.238

combined merge-select runs. Table 2 shows the predicted costs per title for combined merge-select runs with varying file and batch sizes. The costs shown are based on the following equation:

$$C_T = \left(\frac{FS_O + FS_A \; FS_D + FS_N}{BS} \right) C_{IO} + C_P$$

where

C_T is the cost per title

FS_O is the file size for the old MARC file

FS_A is the file size for the add records (1200)

FS_D is the file size for the delete records (1200)

FS_N is the file size for the new MARC file

BS (Batch Size) is the number of records selected in the run

C_{IO} is the cost of reading or, writing a record ($.00068)

C_P is the cost of processing a selected record ($.073)

Calculations for this table are based on several assumptions: it is assumed that the file has reached a state of equilibrium in which the weekly additions and deletions are equal; it is also assumed that delete records have the same average length as other records and therefore take as long to read. While it is unlikely that these assumptions will hold perfectly, the variations are not great enough to destroy the usefulness of the resulting figures as a guide.

DISCUSSION

The figures presented in the two tables have several implications for the design of systems based on the maintenance of a cumulative MARC file and the selection of records from that file. First, they show the importance of assuring that no unnecessary passes of the cumulative MARC file are made. Updating of the MARC file, production of indexes to it and selection of records from it should be accomplished in a single pass of the file. If it is desired to select records from the file more often than once a week, Table 1 provides a means of estimating the cost of the improved response time. If for example, the file size is 100,000 and the weekly volume is 500, twice-a-week runs would increase the cost by 14 cents per title or by $68.00 a week for the select runs.

The figures presented in the two tables also show the critical importance of controlling the growth of the cumulative MARC file, especially for

libraries with a relatively small volume of titles to be processed. Three characteristics of the acquisitions program of the library largely determine the possibilities for controlling the growth of this file. The number of titles acquired by the library determines the batch sizes for records to be selected from the file each week. The acquisition rate is also an important determinant of the growth rate of the cumulative file provided that records which have been selected and used are then purged from the file. If the Library of Congress issues an average of 1200 titles per week and a library uses an average of 1000 titles a week from the file, the net annual growth of the cumulative file will be only slightly over 10,000 records. On the other hand, a smaller library selecting an average of only 100 titles a week would have a net annual growth rate of about 57,000. If unused records were purged after one year, the file size would remain stable at these levels. Table 2 indicates that the cost per title for file maintenance and selection at these two libraries would be about 9 cents and 86 cents respectively.

A second characteristic of the acquisitions program of the library that is important in controlling the growth of the cumulative MARC file is the scope of the subject coverage attempted. If most of the monographs acquired fall within well defined subject classes, the probability of utilizing MARC records in many other subject classes may be low enough that these records need not be added to the cumulative MARC file at all. For a special library that attempts to collect everything published in a few well defined subject areas it may be economical to maintain and utilize a limited MARC file even though the number of records selected is small. On the other hand, a small or medium-sized public library acquiring the same number of titles would probably find a much larger percentage of its records on the MARC file but still not be able to use the MARC tapes economically. Since the public library is likely to collect titles in most subject fields, the probabilities of utilizing records in different classes would not vary as widely and it would not be possible to limit the file to records in a few classes having a high probability of utility. Consequently, the per-item cost of MARC records would likely be too high for consideration. If it is determined that the probabilities of using MARC records vary widely for other characteristics, such as publisher, these characteristics may be used for restricting the records to be added to the cumulative file, thus limiting its size, but subject class seems to be the most promising characteristic for this purpose.

An analysis by subject class of all non-juvenile records in the MARC I file and of those records selected from it for use by the Georgia Tech Library has been used as the basis for restricting the growth of the cumulative file of MARC II records. Overall, 8,953 out of 46,486 records were utilized, 19.3% of the file. The percentage selected varied from more than 50% in some engineering classes to less than 1% in a few classes such as CS (Genealogy) and BW (Practical theology). Elimination of thirty

classes in which fewer than 4% of the records were eventually used would have reduced the file by 7,710 records or 16.6%. Only 184 of these records (2.4%) were eventually selected for use. Records for these thirty subject classes are not being added to the Georgia Tech file of MARC II records.

A third characteristic of the acquisitions program important in controlling the growth of the cumulative MARC file is the speed with which newly published monographs are acquired. If most monographs are acquired soon after publication, the probability of using a MARC record that has not been selected in the first few months after its receipt may be low. Unselected records may therefore be purged after a relatively short time and the file size thereby controlled. Use of the MARC tapes for book selection will help to increase the probability of records being selected during the first few months on the file. A system that uses the weekly MARC tapes for book selection and does not retain on the cumulative MARC file those records not selected for purchase might be quite economical. The frequency with which decisions are later made to acquire titles that were initially passed over, and the added cost for manual input of those records, would have to be considered in deciding on this policy.

An analysis has been made of the interval between the date records were added to the MARC file and the date on which they were selected for use by the Georgia Tech Library. Distributions by time intervals for each Library of Congress subject class were prepared. The distributions varied significantly for reasons that are not yet clear. Generally, it appeared that in those subject classes for which a smaller percentage of the titles available on the MARC file were acquired, they were acquired more rapidly. This seems to be advantageous for keeping the MARC file small. For those classes in which a large percentage of titles are selected, unselected records will be retained on the file for a long period, such as eighteen months. Use of a large percentage will mean that the number of unused records remaining on the file will be relatively small and they will have a high probability of selection over the extended period. For those classes in which a smaller percentage of titles are acquired, the unselected records will be retained on the file for a shorter period, such as six months. Since titles in these fields tend to be acquired more promptly, few potentially useful records will be lost by purging unselected records after a shorter interval.

Over the past year major changes have been made in acquisitions procedures in the Georgia Tech Library. A much larger proportion of monographs are now received on approval plans. The MARC distribution service now provides about twice as many records each week as were provided during the pilot project phase. The effects of these changes on the proportion of titles selected and the time required for acquiring titles in the various subject classes have not yet been determined. Continuous monitoring of the operation of the system for changes in these characteristics

will be required for efficient operation. The improved program for maintenance of the MARC II file and selection of records from it provides for designating subject classes which are not to be added to the file and designating how long unselected records in other classes are to be retained on the file.

This study of variations in the computer costs of card production lends support to the decision to continue using COBOL as the primary language for the MARC II system being implemented on the UNIVAC 1108 rather than using assembly language. The inefficiency of COBOL for character-by-character code conversion and for manipulating variable length data had been a source of some concern. The cost of all processing of selected records, including code conversion, reformatting, prooflisting, making corrections, generating and formatting added entry records, and sorting and printing catalog cards, averaged only about 16 cents per title. A reduction of even 50% through the use of assembly language and increased effort directed to program efficiency would reduce costs by only about 8 cents per title or 1 cent per card. These savings do not seem to justify the increased original programming costs and the likelihood of eventual costly reprogramming. On the other hand, the cost of selecting records from the MARC file varied from 3 cents per title to 29 cents per title. With the added cost of weekly maintenance of the MARC file and with more than twice as many MARC records being received, the costs of processing the cumulative MARC file might easily go much higher. By careful attention to controlling the growth of this file, significant savings in the cost of the system may be achieved.

CONCLUSION

Some librarians have assumed that as the scope of the MARC distribution service expands to include other languages and other types of materials their problems of inputting current records will be solved. This analysis shows that the situation is not so simple. Probably only a few of the largest general research libraries will be able to maintain complete MARC files for their individual use during the next few years, though reductions in computing costs may eventually change this prediction. Even medium-sized libraries such as Georgia Tech will not be able to use economically the foreign language materials when they are included in the MARC program.

Some libraries which do not use a large enough proportion of the MARC records to make it economically practical to maintain a complete MARC file may be able to make economical use of MARC records by carefully controlling the retention of records on the cumulative file. Continuing analysis of the probabilities for selecting records of varying age and subject classes may be utilized in developing a formula for maintaining the file at near optimum size if the system provides for collection of the required statistics.

For libraries which cannot profitably use the MARC tapes, there is another prospect. Cooperative centers that do the processing for large library systems or for several systems will have the volume to justify maintenance of complete files. Certainly, a processing center serving all libraries of the University System of Georgia could economically maintain a more complete MARC file than Georgia Tech alone can justify. The development of cooperative processing programs in Ohio, New England, Oklahoma, (7, 8, 9) and elsewhere indicates that some librarians are coming to this realization.

ACKNOWLEDGMENTS

Mrs. Julie Gwynn wrote most of the computer programs referred to in this paper. Her husband, Professor John Gwynn, gave valuable advice on the statistical techniques employed in analyzing the data. The University of Toronto Library generously provided a copy of its MARC file, which included the date each record was added to the file, for use in analysis of the time lag between availability of the record and selection of it.

REFERENCES

1. Fasana, Paul J.: "Automating Cataloging Functions in Conventional Libraries," *Library Resources and Technical Services,* 7 (Fall 1963), 350-365.
2. Kilgour, Frederick G.: "Costs of Library Catalog Cards Produced by Computer," *Journal of Library Automation,* 1 (June 1968), 121-127.
3. Stone, Sandra F.: *Yale Bibliographic System; Time and Cost Analysis at the Yale Medical Library* (Unpublished document, New Haven: Yale University Library, 1969).
4. Murrill, Donald P.: "Production of Library Catalog Cards and Bulletin Using an IBM 1620 Computer and an IBM 870 Document Writing System," *Journal of Library Automation,* 1 (September 1968), 198-212.
5. Kennedy, John P.: "A Local MARC Project: The Georgia Tech Library." In University of Illinois, Graduate School of Library Science: *Proceedings of the 1968 Clinic on Library Applications of Data Processing* (Urbana: University of Illinois, 1969), pp 199-215.
6. *Ibid.*
7. Kilgour, Frederick G.: "A Regional Network — Ohio College Library Center," *Datamation,* 16 (February, 1970), 87-89.
8. Agenbroad, James E.; et al.: *Systems Design and Pilot Operations of the New England State Universities. NELINET, New England Library Information Network. Progress Report, July 1, 1967 - March 30, 1968* (Cambridge, Mass.: Inforonics, Inc., 1968). ED 026 078.
9. Bierman, Kenneth John; Blue, Betty Jean: "Processing of MARC Tapes for Cooperative Use," *Journal of Library Automation,* 3 (March 1970), 36-64.