THE RECON PILOT PROJECT: A PROGRESS REPORT OCTOBER 1970-MAY 1971

Henriette D. AVRAM and Lenore S. MARUYAMA: MARC Development Office, Library of Congress, Washington, D. C.

Synopsis of three progress reports on the RECON Pilot Project submitted by the Library of Congress to the Council on Library Resources covering the period October 1970-May 1971. Progress is reported in the following areas: RECON production, foreign language editing test, format recognition, microfilming, input devices, and tasks assigned to the RECON Working Task Force.

INTRODUCTION

With the implementation of the MARC Distribution Service in March 1969, the Library of Congress and the library community have had available in machine readable form the catalog records for English language monographs cataloged since 1969. Most libraries, however, also need to convert their older cataloging records, and the Library of Congress attempted to meet these needs by establishing the RECON Pilot Project in August 1969. During the two-year period of the pilot project, various techniques for conversion of retrospective bibliographic records have been tested, and a useful body of catalog records is being converted to machine readable form.

The pilot project is being supported with funds from the Library of Congress, the Council on Library Resources, and the U.S. Office of Education. Earlier articles in the *Journal of Library Automation* have described the progress through September 1970 (1, 2, 3). This article covers the period October 1970 through May 1971.

PROGRESS—OCTOBER 1970 THROUGH MAY 1971

RECON Production

The conversion of 8476 records in the 1969 and 7-series of card numbers that had not been included in the MARC Distribution Service was completed, and these records were sent to 47 subscribers of the MARC Distribution Service. The subscribers were not charged for these records but were asked to send a tape reel to the Library for the duplication process.

At present, the RECON data base consists of 25,206 records in the 7, 1969, and 1968 series of card numbers. Records in the 1968 series that were part of the data base for the MARC Pilot Project are being converted by program from the MARC I format to the MARC II format, proofed, and updated. To date, 7551 out of 7583 MARC I records have been processed. Prior to the implementation of the MARC Distribution Service, records were input for test purposes, and the resulting practice tapes contain data requiring correction or updating to correspond with the present specifications of the MARC II format. Of the 8340 titles on the practice tapes, 3460 have been updated and reside on the RECON master file. These updated machine readable records will be distributed with the RECON titles in the 1968 card series.

Foreign Languages Editing Experiment

A foreign language editing experiment was conducted to test the accuracy of MARC/RECON editors in editing French and German language records. Records used for this test included 1180 of the 5000 RECON research titles. At least 50 percent accuracy was expected since half of the task of editing a MARC record involves being able to read the language of the record. The other half involves identifying the data elements by their location in the record. The three editors used in the experiment had studied French in high school, one having had an additional year in college; none had studied German. Each editor was required to edit approximately 200 records in each language.

Statistics on the number of records edited per hour and the number of errors made, when compared with the same editors' statistics for editing English language records, showed that each editor maintained an approximately equal rate of speed in editing foreign language records as in editing English. The error rate for each editor, however, was more than tripled on foreign records, and each made approximately as many errors in French (the language studied) as in German. Each editor averaged more than 12 errors per batch in French and 12 in German. Since the MARC Editorial Office has established a standard of 2.5 errors per batch (20 records comprising a batch) as being acceptable for trained MARC editors, this error rate would have to be lowered in a production environment.

The majority of errors occurred in the title field, which is a portion of the record that must be read for content in order to be edited correctly. The second largest number of errors occurred in the fixed fields, which are also dependent upon a reading knowledge of the language of the record for accurate coding.

The number of errors made in each batch of records by each editor was tabulated to determine if any improvement was made during the course of the experiment. In no case was improvement noted. Statistics were also kept on the number of times an editor consulted various sources for help: e.g., dictionaries, the editing manual, the LC Official Catalog, the reviser, or a language specialist. Dictionaries were consulted frequently, and the reviser and language specialists rarely.

Typing statistics (number of errors) were also recorded for 181 French and 185 German records. The error rate for typing foreign language material was lower than for typing English. The English language statistics, however, were combined for several typists, and the foreign language statistics were for one typist only. Charts showed that there was no improvement in the number of typing errors made at the end of the test.

The primary conclusion drawn from the results of the experiment is that in order to edit foreign language records with an acceptable degree of accuracy, it would be necessary for the editor to have a good knowledge of the language as well as the editing procedures.

Format Recognition

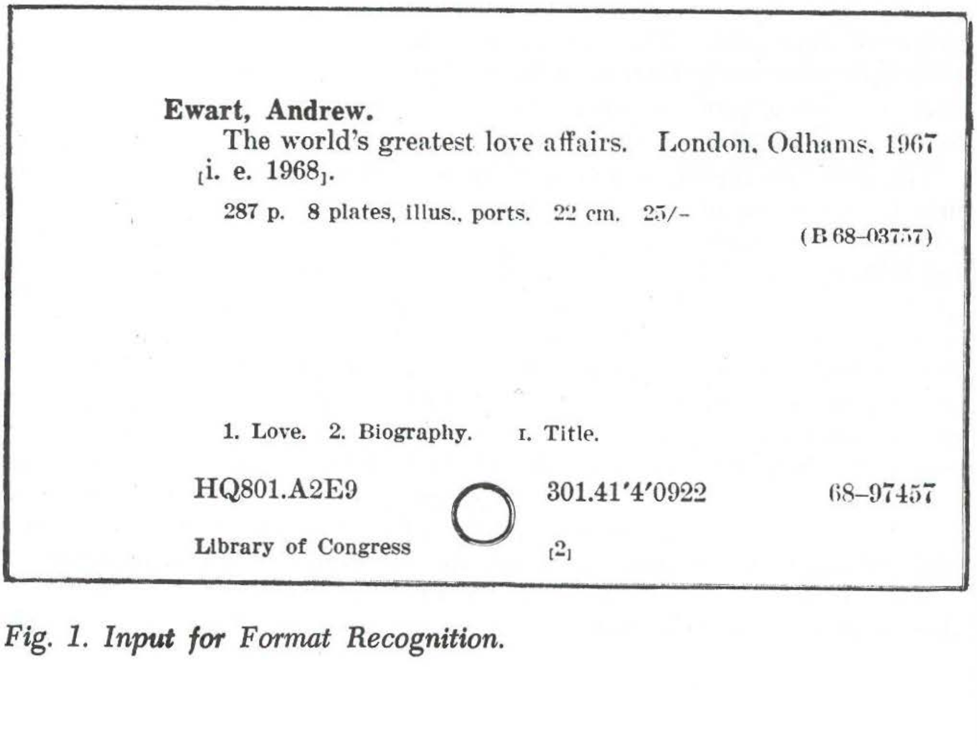
Format recognition is a technique that allows the computer to process unedited bibliographic records by analyzing data strings for certain keywords, significant punctuation, and other clues to determine proper identification of data fields. The Library of Congress has been developing this technique since early 1969 in order to eliminate substantial portions of the manual editing process, which in turn should represent a considerable savings in the cost of creating machine readable records.

The RECON report, which was written prior to the completion of the first format recognition feasibility study, concluded that "partial editing combined with format recognition processing is a promising alternative to full editing." (4) Since that time, the emphasis in the development of the programs has been shifted to no editing prior to format recognition processing. The programs are in the final stages of acceptance testing, and it is expected that 75% of the records can be processed without errors created by the format recognition programs. Preliminary estimates show that it takes approximately half a second of machine time to process one record by format recognition; the manual editing process, on the other hand, takes approximately six minutes per record. The total amount of core storage required is approximately 120K: 80K for the programs and 40K for the keyword lists. Although the keyword lists are maintained as a separate data set on a 2314 disk pack, they are loaded into memory during processing. The format recognition programs have been written in Assembler Language for the Library's IBM 360/40 under DOS.

The logical design of the format recognition process, with detailed flow charts needed for implementation of computer programming, has been published as a working document by the American Library Association so that the technical content would be available to assist librarians in their automation projects (5).

Workflow for format recognition begins with the input of unedited catalog records via the MT/ST following the typing specifications created for format recognition. After being processed by the format recognition programs, these records are proofed by the editors (the first instance in which they see the records), and the necessary corrections or verifications made. Correction procedures for format recognition records are the same as those used for regular MARC records. Figures 1, 2, and 3 are examples of the printed card used for input, the MT/ST hard copy, and the proofsheet of the record created by format recognition.

Initial use of the format recognition programs is for input of approximately 16,000 RECON records in the 1968 card series. Input of current MARC records via format recognition will begin at a later date. RECON records were chosen for large-scale testing because they are not required for an actual production operation such as the MARC Distribution Service. In addition, work has begun on the expansion of format recognition to foreign languages. Analysis is being done on German and French monograph records, and eventually Spanish, for new or expanded keyword lists and some changes to the algorithms.



Ewart, Andrew.

The world's greatest love affairs. London, Odhams, 1967 [i. e. 1968].

287 p. 8 plates, illus., ports. 22 cm. 25/-

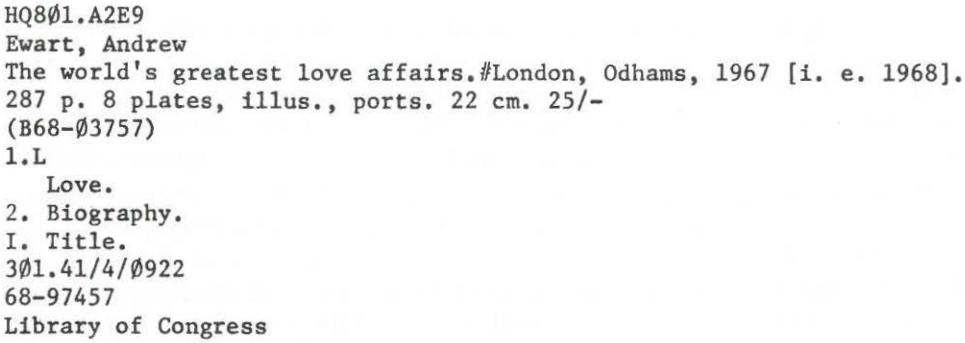
(B 68-03757)

1. Love. 2. Biography. I. Title.

HQ801.A2E9 301.41'4'0922 68-97457

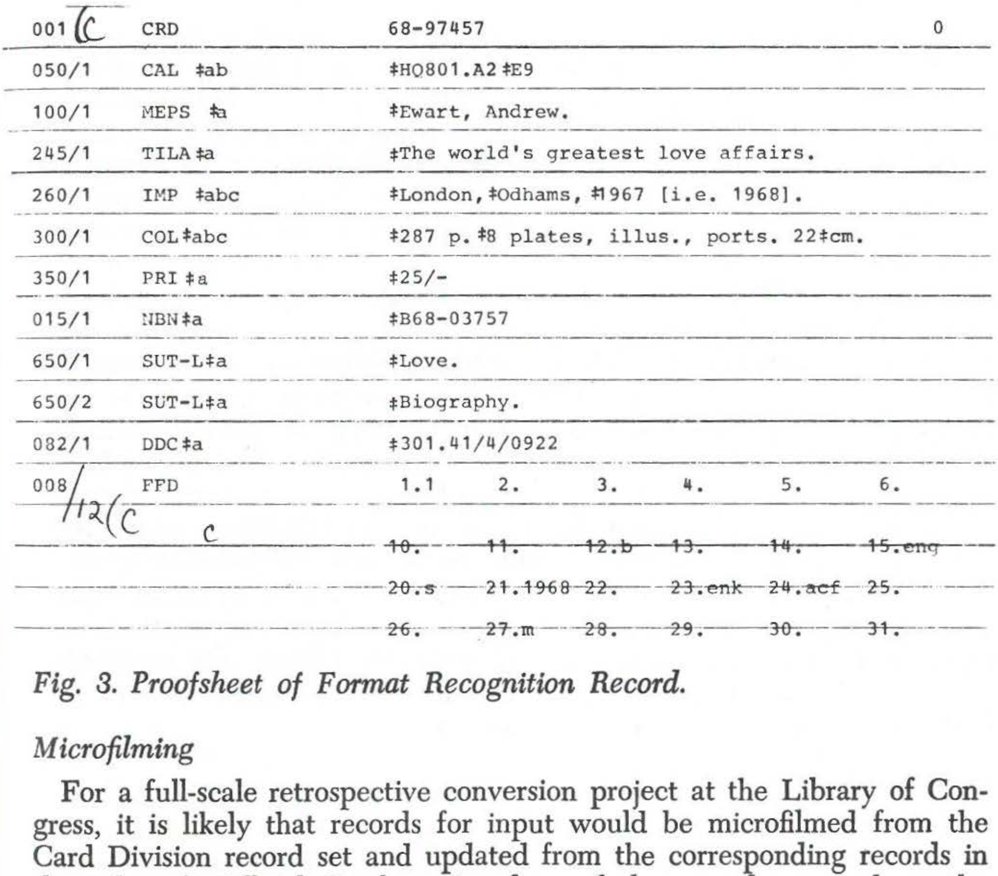
Library of Congress [2]

Fig. 1. Input for Format Recognition.



```
HQ8Ø1.A2E9
Ewart, Andrew
The world's greatest love affairs.#London, Odhams, 1967 [i. e. 1968].
287 p. 8 plates, illus., ports. 22 cm. 25/-
(B68-Ø3757)
1.L
   Love.
2. Biography.
I. Title.
3Ø1.41/4/Ø922
68-97457
Library of Congress
```

Fig. 2. MT/ST Hard Copy.



001 (C	CRD	68-97457						0
050/1	CAL ‡ab	‡HQ801.A2 ‡E9						
100/1	MEPS ‡a	‡Ewart, Andrew.						
245/1	TILA ‡a	‡The world's greatest love affairs.						
260/1	IMP ‡abc	‡London, ‡Odhams, ‡1967 [i.e. 1968].						
300/1	COL ‡abc	‡287 p. ‡8 plates, illus., ports. 22‡cm.						
350/1	PRI ‡a	‡25/-						
015/1	NBN ‡a	‡B68-03757						
650/1	SUT-L‡a	‡Love.						
650/2	SUT-L‡a	‡Biography.						
082/1	DDC ‡a	‡301.41/4/0922						
008/12(C	FFD c		1.1	2.	3.	4.	5.	6.
			10.	11.	12.b	13.	14.	15.eng
			20.s	21.1968	22.	23.enk	24.acf	25.
			26.	27.m	28.	29.	30.	31.

Fig. 3. Proofsheet of Format Recognition Record.

Microfilming

For a full-scale retrospective conversion project at the Library of Congress, it is likely that records for input would be microfilmed from the Card Division record set and updated from the corresponding records in the Library's Official Catalog. A subset of the record set, such as the catalog cards for a given year, would be microfilmed and then the appropriate records, i.e., English language monographs, German monographs, etc., would be selected after filming. Costs were calculated for a base figure of 100,000 records for the year 1965, and four different methods of microfilming have been estimated as follows by the Library's Photoduplication Service: 1) microfilming for a direct-read optical character reader ($2000); 2) microfilming for reader/printer specifications ($2350); 3) microfilming for reader specifications ($400); and 4) microfilming for a Xerox Copyflo printout of a card overlaid on a 8 x 10½ worksheet ($7000).

The differences in cost are primarily attributable to the type of camera used (rotary or planetary) and the kind of feed mechanism (manual or automatic). Other factors need to be considered, such as the fact that film suitable for OCR requirements could not be used on Xerox Copyflo or even for contact printing to positive film. Since a readable copy of the original printed card is necessary for updating and proofing, microfilming for direct-read OCR would not be a viable alternative.

Input Devices

The monitoring of existent input devices was continued with an investigation of Dissly Systems' Scan Data optical character reader. Scan Data has been modified, via software, to read 55 different type fonts which are recognized by a "best compare" technique using six stored fonts to match against the remaining 49. According to the manufacturer, direct-reading is accomplished with approximately 95% level of accuracy. Errors are recorded during a proofing cycle and corrected in the machine readable data base.

The Scan Data equipment does not have a transport for a 3 x 5 document, so that a number of 3 x 5 cards must be attached to an 8 x 14 document for scanning, and therefore these cards would not be returned to the Library by the manufacturer. Under these conditions, cards to be read by Scan Data equipment would have to be obtained from stock rather than from the Card Division record set. Unfortunately, many cards are out of stock; and of those that are in stock many may be cards reprinted several times by photo-offset methods and consequently have a poor image. Therefore the use of this device would be severely hampered.

Fifty good quality cards were submitted to Dissly Systems for an experiment that was run without any modifications to the existing machine and software. Five of the 50 cards were returned to the Library with a matching printout. The results were not encouraging because many lines of text were missed and many characters misread.

RECON Working Task Force

The RECON Working Task Force has compiled work statements for contractual support for two of its research projects. These projects involve investigations on the implications of a national union catalog in machine readable form and the possible utilization of machine readable data bases other than that of the Library of Congress for use in a national bibliographic store. Preliminary tasks related to these projects have been described in earlier progress reports (6, 7).

The first part of the work statement deals with the products that could be derived from the machine readable national union catalog: a bibliographic register, indexes by name, title, and subject, and a register of locations. These indexes would provide multiple access points to the records in the National Union Catalog.

The bibliographic register will contain a full bibliographic record on each title covered. The indexes will contain partial records which are associated with the full records in the register, and a given index file will carry one or more partial records for every record in the register. For each title in the register, the register of locations lists those libraries where copies of the title are held.

The assumption is made that the indexes under consideration will contain the following data elements (the numeric designations and subfield codes are those used in the MARC format fields):

Name Index
- Name (100, 110, 111, 400, 410, 411, 600, 610, 611, 700, 710, 711, 800, 810, 811)
- Short title (245)
- Main entry in abbreviated form
- Date (fixed field Date 1)
- Language (fixed field language code)
- LC card number
- Register number

Title Index
- Short title (130, 240, 241, 245, 440, 630, 730, 740, 840)
- Main entry in abbreviated form
- Date (fixed field Date 1, or may be omitted if in heading)
- Language (fixed field language code, or may be omitted if in heading)
- LC card number
- Register number

Subject Index
- Subject heading (650, 651)
- Main entry (100, 110, or 111)
- Short title (245)
- Date (fixed field Date 1)
- Language (fixed field language code)
- LC card number
- Register number

The abbreviated form of main entry noted above is to be included in the record of the name or title index unless the name itself is carried in the main entry of that record. It is defined as follows: 1) for a personal name, a conference, or a uniform title heading—subfield "$a" is appended in brackets after the title; and 2) for a corporate name—subfield "$a" plus the first "$b" subfield are appended, within a single set of brackets, after the title.

The specific objective of this project is to define and investigate alternative processing schemes associated with an automated National Union Catalog. This study will explore and examine these processing schemes and the following components:

1) Techniques for introducing the necessary input into the automated NUC system. The considerations to be covered include the relationship to MARC input, use of the format recognition programs, and the problems of language in terms of selection of input.
2) Techniques for structuring or organizing the data contained in the register and the various indexes to establish and maintain the relationships among the records contained in these data bases.
3) Techniques and procedures connected with the production of the products listed above. This investigation will also cover any selection and sorting procedures necessary.
4) Analysis of the format, i.e., graphic design and printing, size, style, typographic variation, condensation, etc.
5) Examination of alternative cumulation patterns associated with the products of the system. In this connection, items such as number of characters in an average entry, average number of entries on a page, expected rate of increase of number of entries in catalog, and segmentation of catalog are to be taken into consideration.
6) Feasibility of producing a register through automation techniques. If this can be accomplished, further investigation will be directed toward the feasibility and cost of segmenting the register into three sections: one produced from machine readable records (English and whatever roman alphabet language records are in machine readable form); one produced from roman alphabet language records which are only in printed form; and one produced from non-roman alphabet language records which are only in printed form.

The costs associated with the various techniques and procedures enumerated above as well as with their components will be calculated. From these figures an average total cost per title cataloged is to be determined for each alternative processing scheme. These cost values (one per alternative scheme) are to be compared with those associated with a purely manual processing scheme. Included in this cost analysis will be the associated costs for different forms of hard copy as well as for the use of COM (Computer Output Microfilm).

From any one index and the register of locations, the maximum number of alphabetic and numeric lists (registers of location ordered by register number) will be determined, taking into account ease of usage and technical and economic feasibility. The intent is to have as few lists as possible and still keep the cost within reasonable bounds. Supplements to the indexes should be issued monthly; supplements to the register of locations may be issued monthly or quarterly.

The second project is a continuation of a previous investigation on the possible utilization of machine readable data bases other than that produced by the Library of Congress for use in a national bibliographic store. The results of this project should determine if the use of other data bases is economically and technically feasible. Using three or four data bases selected by the RECON Working Task Force, the study will determine the following:

1) Method and cost of acquiring these other data bases in machine readable form.
2) Analysis of the kinds of programs capable of converting records from a number of these data bases into the MARC format. Different level data bases might require different kinds of programs. If such an effort is deemed feasible, a cost estimate for such a program or array of programs will be calculated.
3) Method and cost of printing the records for examination, corrections, etc.
4) Method and cost of eliminating records already in the MARC data base.
5) Method and cost of comparing these records against the LC Official Catalog and making the necessary changes in the data or content designators.
6) Cost for input of additions and corrections.
7) Method and cost of incorporating the additions and corrections in the machine readable file.
8) Cost of providing means by which these records would not be input again by any future LC retrospective conversion effort.

A result of this project should be a determination as to whether high potential or medium potential files, or both, are suitable for conversion. A determination will be made of the minimum yield or the minimum number of titles needed to justify writing the programs to convert these data bases. A factor to be considered is that the number of unique titles will decrease as more data bases are converted for this pool of records.

It was decided that the research tasks to study the problems in distributing name and subject cross reference control files would be dropped because of limitations of time and funds. An additional task, however, has been added that can be performed within the time limits of the pilot project. During the past year, the Library of Congress Card Division has recorded information about card orders in machine readable form. This information will be analyzed as to the year and language of the most frequent orders because it is assumed that the most popular card orders bear a relationship to the potential use of a data base in machine readable form by libraries in the field. This study involves the following:

1) Analysis of a frequency count of LC card orders for a one-year period and preparation of a distribution curve for card series.

2) Analysis of a sample of frequently ordered cards to determine with fair reliability the proportion of English language titles in this group. The sample will be large enough to give an indication of other language groups that might be significant for any RECON effort.
3) Preparation of distribution curves for English language and non-English titles by card series.
4) Mathematical analysis of the results of 1)-3) above to arrive at a table to show the anticipated utility of converting specified subsets of the LC card set.

OUTLOOK

Research in input devices has not uncovered any equipment that offers a significant technical and cost improvement over the MT/ST currently used in the Library of Congress. On-line correction and verification of MARC/RECON records will, however, speed conversion and will offer relief in the flow of documents and paper work required in a purely batch operation. Since MARC/RECON records will be corrected and verified in one operation rather than by the cyclic process of the present system, cost savings should be realized. The Library of Congress will have this on-line capability through the Multiple Use MARC System. This new system is still in the design phase, and a projected date for implementation has not yet been set.

To date investigations in the use of direct-read optical character readers have demonstrated that there are no devices currently available capable of scanning the LC printed card.

The format recognition programs are operational, and RECON titles in the 1968 card series are being converted without any prior editing of the records. Procedures are being implemented to gather the necessary data to compare costs of the format recognition technique with costs of conversion with human editing.

Production statistics have shown that retrospective records are more costly to convert than current records. This higher cost is attributed to the additional tasks in RECON of selecting the subset for input from the LC record set and comparing the records with the LC Official Catalog for updating. Since cards in the LC record set do not necessarily reflect the latest changes made to the cards in the LC Official Catalog, the Official Catalog comparison is necessary to ensure that RECON records are as up-to-date as the cards in the Official Catalog.

Although the RECON report (8) recommended conversion in reverse chronological order with highest priority given to the last ten years of English language monograph cataloging, the Working Task Force study on the Card Division popular titles may reveal that selective conversion is a more practical approach. The orderliness of chronological conversion by language does mean that records in machine readable form can be ascertained easily. It is interesting, however, to speculate on the use of these records compared with popular titles which may cross many years and languages.

The MARC/RECON titles constitute the data base for the Phase II Card Division Mechanization Project, and close liaison continues to be maintained between both projects. It is recognized that the distribution of cards and MARC records requires the same computer based bibliographic files and has similar hardware and software requirements. Plans are presently underway to transfer the duplication of tapes for MARC subscribers from the Library's IBM 360/40 to the Card Division's Spectra 70 when the Phase II system is operational.

The RECON Pilot Project does not officially end until August 1971. In an attempt to make information available as rapidly as possible, the preparation of the final report will begin this summer, since several aspects of the project are complete enough to be documented. The final report will be published by the Library of Congress, and its availability will be announced in the LC *Information Bulletin* and in professional journals.

ACKNOWLEDGMENTS

The authors wish to thank the staff members associated with the RECON Pilot Project in the MARC Development Office, the MARC Editorial Office, the Technical Processes Research Office, and the Photoduplication Service of the Library of Congress for their contributions to the project and, therefore, to this report. Special thanks are due to Patricia E. Parker of the MARC Development Office for her work on the foreign language editing experiment and for writing that section of this article.

REFERENCES

1. Avram, Henriette D.: "The RECON Pilot Project: A Progress Report," *Journal of Library Automation,* 3 (June 1970), 102-114.
2. Avram, Henriette D.; Guiles, Kay D.; Maruyama, Lenore S.: "The RECON Pilot Project: A Progress Report, November 1969-April 1970," *Journal of Library Automation,* 3 (September 1970), 230-251.
3. Avram, Henriette D.; Maruyama, Lenore S.: "RECON Pilot Project: A Progress Report, April-September 1970," *Journal of Library Automation,* 4 (March 1971), 38-51.
4. RECON Working Task Force: *Conversion of Retrospective Catalog Records to Machine-Readable Form: A Study of the Feasibility of a National Bibliographic Service* (Washington, D.C.: Library of Congress, 1969), 179.
5. U. S. Library of Congress. Information Systems Office. *Format Recognition Process for MARC Records: A Logical Design* (Chicago, American Library Association, 1970).
6. Avram, Guiles, Maruyama, op. cit., 248-249.
7. Avram, Maruyama, op. cit., 49-51.
8. RECON Working Task Force, op. cit., 11.