THE RECON PILOT PROJECT: A PROGRESS REPORT

Henriette D. AVRAM: Project Director, Information Systems Office,
Library of Congress, Washington, D. C.

A synthesis of the progress report submitted by the Library of Congress to the Council on Library Resources under an Officer's Grant to initiate the RECON Pilot Project that gives an overview of the project and the progress made from August-November 1969 in the following areas: training, selection of material to be converted, investigation of input devices, and format recognition.

INTRODUCTION

The RECON Pilot Project is an effort to analyze the problems of large-scale conversion of retrospective catalog records through the actual conversion of approximately 85,000 non-current records. This project has grown directly out of the implementation of the MARC Distribution Service. Libraries considering the use of machine readable records for their current materials have naturally begun to consider conversion of their older records as well. Some libraries have even begun such conversion projects.

Since the Library of Congress is also interested in the feasibility of converting its own retrospective records, it seemed appropriate to explore the possibility of centralized conversion of retrospective cataloging records and their distribution to the entire library community from a central source. A proposal having been submitted by the Library of Congress to the Council on Library Resources, Inc. (CLR), the Council granted funds for a study of this problem. An Advisory Committee was appointed to provide guidance, and direct responsibility for the study and report (1) was assigned to a Working Task Force.

A recommendation of the Working Task Force was the implementation of a pilot project to test the techniques suggested in the report in an operational environment. Since any feasibility report, no matter how detailed, refers to a theoretical model, the recommended techniques should be tested to determine a most efficient method for a large-scale conversion activity. The Advisory Committee concurred with this recommendation. The Library of Congress submitted a proposal for a pilot project (hereinafter referred to as RECON) to CLR, and received an Officer's Grant in August 1969 to initiate RECON while the Council continued its evaluation of the full-scale pilot project.

A progress report was submitted to CLR by the Library covering the period from mid-August to November 1, 1969. So that CLR might have a clear understanding of the work in progress, the report addressed itself to both the areas of RECON supported by the Council and those activities supported by the Library of Congress. In December 1969, CLR awarded the Library the funds requested for the entire pilot project. To make the library community cognizant of RECON as quickly as possible, CLR granted permission to modify the progress report for publication.

OVERVIEW OF THE RECON PILOT PROJECT

The pilot project is concerned with the conversion and distribution of an estimated 85,000 English language titles: 22,000 titles cataloged in 1969 and not included in the MARC Distribution Service, and 63,000 titles from 1968. The creation of this data base partially satisfies the conclusions and specific recommendations of the RECON Working Task Force as stated in the report (2): 1) there should be no conversion of any category (language or form of material) of retrospective records until that category is being currently converted; 2) the initial conversion effort should be limited to English language monograph records issued from 1960 to date and converted into machine readable form in reverse chronological order. (MARC Distribution Service covers current English language monographs cataloged by the Library of Congress). In order to explore the problems encountered in encoding and converting cataloging records for older English language monographs, and monographs in other roman alphabet languages, 5,000 additional titles will be selected and converted.

The Library further intends to investigate, through the design and implementation of a format recognition program, the use of the computer to assist in the editing of cataloging records. This technique should significantly reduce the manpower needs of the present method of conversion and therefore have an impact on any future Library of Congress conversion activity, either of currently cataloged or retrospective titles.

RECON will include experimentation with microfilming and producing hard copy from the LC record set.

The record set in the LC Card Division consists of a master copy of the latest version of every LC printed card, arranged by card series and, within each series, by card number. Although a specific time period can be selected for conversion, the primary disadvantage of the record set for this purpose is the fact that not all changes in cataloging made to the LC Official Catalog are reflected in the record set. After considering all the alternatives, the RECON Working Task Force recommended (3) that the record set be used for selection of titles, but that the titles be compared with the Official Catalog and updated to insure bibliographic accuracy and completeness. Since the record set is in constant use by Card Division personnel, the selected titles for conversion must be reproduced, and the original file reconstituted, as quickly as possible.

The state of the art of direct-read optical character recognition devices suitable for large-scale conversion will be monitored and experimentation will be conducted with a variety of input devices.

RECON is closely related to the LC Card Division Mechanization Project, which is based upon the availability of records in machine readable form. RECON will be closely coordinated with the Card Division project, both in the design of specifications for implementation and in the investigation of a common hardware/software configuration.

The project was organized during August 1969. The first group of records being edited are those cataloged by the Library of Congress in 1969. In June 1970, the editing of the 1968 records will begin. Since these records will have to be compared with the LC Official Catalog to record any changes, present thinking includes the design of a print program (referred to as a two-up print program) to cut printing time by providing a listing with records arranged in card number sequence (the order of input) and in alphabetic sequence by main entry on the same page. The records will be arranged by main entry to reduce the effort of checking them against the Official Catalog and the changed records will be inserted in their proper place in sequence by LC card number.

The process of manual editing may be greatly reduced, or perhaps even eliminated, by October 1970, when the format recognition program is scheduled for completion. After this time, the records will be input with little or no prior tagging and further editing will be performed by the computer. The resulting records will be examined by the MARC editors both for accuracy in transcription and for correctness in the assignment of MARC tags, indicators, and subfield codes.

The duration of the pilot project will be twenty-four calendar months, August 1969-August 1971. It is anticipated that by November 1970 enough data should be available to determine whether a full-scale conversion project should be undertaken. An early evaluation of the project is advantageous in order to explore the funding possibilities of a conversion effort if the results of the pilot are affirmative.

Figure 1 is a calendar indicating the major milestones of RECON as postulated during August 1969.

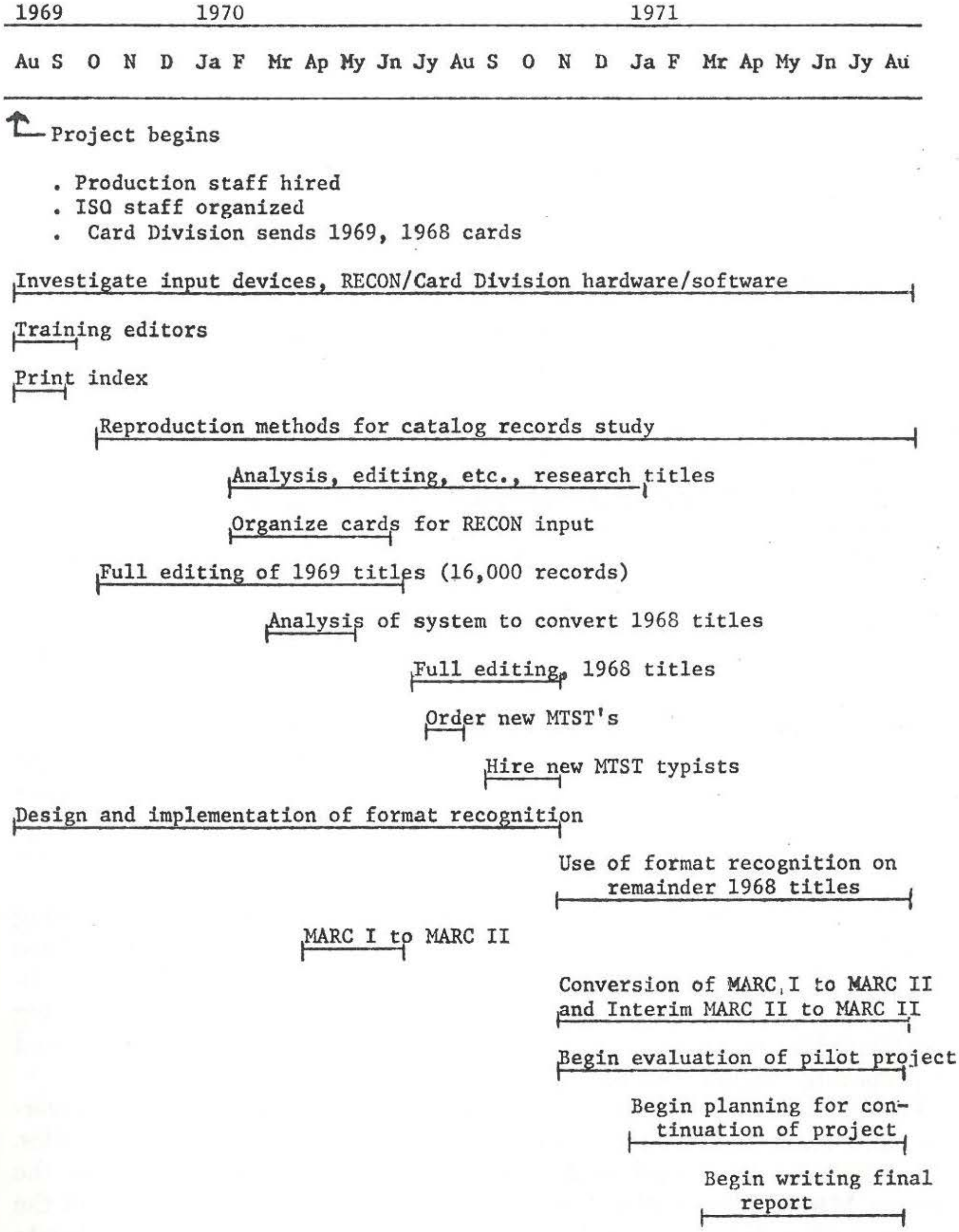

Fig. 1. RECON Calendar.

Essentially the same Advisory Committee and Working Task Force selected for the RECON feasibility study have agreed to serve in their respective capacities for RECON. The implementation of the Library of Congress' MARC Distribution Service and the initiation of RECON are providing the nucleus of a national bibliographic data base. Creation of this data base is not in itself a panacea for libraries but, in fact, amplifies the need to explore some of the larger issues at this time to provide the direction for future cohesive library systems. Certain aspects of the problems were discussed in general terms in the RECON report but time did not permit full analysis.

During the two-year period of RECON, the Working Task Force will consider some of those issues (defined as four tasks listed below) under the grant from CLR. The ability to complete all of the tasks described will be dependent on additional funding, which, it is hoped, may be available early in 1970.

1) Any national data store should have a data base in which all records are consistent. It is possible, and highly probable, that libraries may convert bibliographic records for local use, which may not require the detail of a MARC II record. It is imperative that before levels of completeness of MARC records are defined with respect to content and content designation, the implications of these definitions to future library networks be thoroughly explored.

2) Any consideration of a national bibliographic data store in machine readable form should include the possibility of recording titles and holdings from other libraries. Although the resolution of the problems associated with a machine readable national union catalog are enormous, it is time to begin an exploration of the problems to provide guidance for future design efforts.

3) Several institutions have begun the conversion of their cataloging records into machine readable form. The possibility of utilizing these records in building a national bibliographic data store should be investigated. This will involve evaluating the difficulty and cost of converting and upgrading records converted by others to a MARC format as opposed to preparing original records.

4) The Library of Congress maintains, and is considering the conversion into machine readable form, of its name and subject authority files. Many libraries have expressed interest in receiving these records in the present MARC Distribution Service. Little thought has been given to the storage and maintenance of these large files in each library subscribing to MARC Distribution Service. A library may not have in its collections a bibliographic record requiring either a name or subject cross reference record distributed by the Library of Congress. However, the library will keep the cross reference record because it cannot predict when a title will be added to the collection that does require the cross reference structure. The result will be the eventual storage and maintenance of the

entire LC name and subject reference files in each library. This problem should be explored to determine if there is a possible efficient method of libraries accessing these files from either a centralized source or several regional sources.

PROGRESS—AUGUST 1969 TO NOVEMBER 1969

Organization

The RECON staff is divided into two sections: 1) the Production Section, responsible for the actual editing and keying of the records; and 2) the Research and Development Section, responsible for liaison with the Production Section, determination of the criteria for the selection of the 1968 and 1969 titles, actual selection of the 5,000 research titles, investigation of input devices and photocopying techniques, liaison with the Card Division Mechanization Project, and the design and coding of special computer programs unique to RECON. In addition, staff members of the MARC project team in the Information Systems Office (ISO) are working in areas of format recognition and MARC system programming that will affect RECON.

Training

The MARC experience at the Library of Congress has demonstrated that staff members assigned to the editorial process of preparing catalog records for conversion to machine readable form must be exposed to cataloging fundamentals.

Phase I of the training program for the RECON editors was a two-week cataloging class conducted by the supervisor of the Production Section, a professional librarian with experience in teaching cataloging principles at the Library of Congress. Each day was formally structured into reading, discussion, and practice. The editor-trainees applied the *Anglo-American Cataloging Rules* (4) to practice problems and to actual cataloging of books. Experience in using the LC subject heading list, filing rules, and classification schedules was provided to a lesser extent. In order to insure that the editor-trainees would have a wider range of experience in examining cataloging copy, the mnemonic MARC tags and the more simple indicators and subfield codes were taught and used to identify explicitly cataloging elements on LC proofslips.

Phases II and III of the training, MARC editing and correction procedures, were also taught by professional librarians. The editing class, which lasted two weeks, was divided into lecture sessions and laboratory sessions. Each lecture period was from two to three hours; then, during the laboratory session, the instructions given in the lectures were applied to practice worksheets. The course covered input of variable and fixed fields, assignment of bibliographic codes for language and place of publication, and identification of diacritical marks included in the LC character set. Phase III of the training program, on correction procedures, was a one-week class covering the addition, deletion, and correction of entire records or data elements at the field level.

The training period was followed by an intensive practice period using MARC input worksheets, which were reviewed by the experienced editors.

Selection of Cards

The actual selection of the 1968 and 1969 titles is a joint effort by the Card Division staff and the RECON staff. The procedures for the selection of cards from the Card Division for RECON differ from those described in the original report. Since only cards for 1968 and 1969 titles are being selected, it is more expedient to draw the cards from the Card Division card stock than to microfilm the record set. These cards will include all titles cataloged by the Library of Congress during 1968 and 1969 regardless of language or form of material, which will yield approximately 250,000 cards.

The cards are forwarded to the Production Section from the Card Division, where each record is inspected to determine whether it meets the criteria established for RECON, i.e., all English language monographs with an LC catalog card number representing works cataloged by LC in 1968 and 1969 that are not already in machine readable form.

The determination as to whether or not an item is in English is based upon the text, not the title page. An anthology of literature in Spanish with a title page in English would not be included in RECON; a book with text in English but title page in French would be included. If a book is multilingual (complete text in more than one language), the language of the first title determines inclusion or exclusion for RECON. Atlases are included, but not single maps or set maps. Music or music scores are excluded, but books about music are included. Records representing film strips, moving pictures, serials, and other kinds of materials not regarded as monographs are excluded.

Once the cards eligible for RECON are selected and arranged in LC card number sequence, the cards are compared with the Print Index listing all records already in machine readable form.

Those records not in machine readable form are photocopied onto the input worksheet for editing and keying. To date, 60,000 cards have been selected by Card Division staff and forwarded to the production staff for further processing.

Selection of Research Titles

An integral part of RECON is the conversion of 5,000 titles to machine readable form for research purposes. Ideally, these titles should serve not only the needs of RECON but also be useful for some other purpose in the Library of Congress. These titles would include English language monographs cataloged before 1950, and foreign language material using the roman alphabet, and would be used to test various methods of input and certain aspects of the format recognition program. The older material would represent records cataloged under earlier cataloging rules and would reveal problems in conversion in an area in which little information exists.

Two sources were initially considered for the selection of research titles: 1) titles in the Main Reading Room collection for conversion into machine readable form for the production of book catalogs, and 2) the popular titles (cards ordered most frequently) of the Card Division Mechanization Project. A decision was made to study the titles in both sources with priority given to solution of conversion problems and to determine: 1) if overlap existed in records for both projects that would also serve the needs of RECON; 2) if overlap did not exist, which titles (Main Reading Room collection or Card Division popular titles) best served the needs of RECON; and 3) if the titles in neither project were suitable, the method of selection to be used from the Card Division record set.

The first task was a study of the characteristics of the Main Reading Room collection. The collection consists of approximately 14,000 titles, and printed cards have been collected to compile a complete shelf-list catalog. These cards represent a wide range of material cataloged from 1900 to date. Approximately one-fourth to one-third represent serials. The collection includes material in most of the roman alphabet languages currently processed at the Library, the more common non-roman alphabet languages, such as Russian, Japanese, Hebrew, etc., and a number of "difficult" titles, such as encyclopedias, dictionaries, etc., that would present a variety of cataloging and editing problems.

The second task was a study of the popular titles from the Card Division. The Card Division provided a printout of card numbers for titles with 25 or more orders. There were 4,765 such card numbers listed with their corresponding number of orders. Only 210 of these were for pre-1950 cards, and 97 of the 210 cards were for serial titles. Only 15 out of the 210 cards were for "difficult" titles.

Another list was produced which contained card numbers for titles with ten or more orders. This list (with 39,148 card numbers) did produce more titles that would meet the research needs of RECON. A sampling technique was designed by the Technical Processes Research Office to determine the percentage of overlap of this list with the titles in the Main Reading Room reference collection. The estimated number of matches (15.5%) indicated that not enough overlap existed to consider a selection of titles that would serve the needs of both projects (Main Reading Room collection and Card Division) and RECON. Therefore, the research titles are being selected from records for the reference collection.

ISO is working closely with staff members of the Reference Department on this project. The Reference Department is providing local information (e.g., local call number to locate the item in the reference collection as opposed to the LC call number which locates the item in the general collection) for all titles. As this process is completed, the responsible RECON staff member is selecting the research titles. To date, "local" information has been added to 2,000 records, and 400 RECON titles have been selected from this group of records.

Computer Programs

The only computer program implemented to date is the Print Index Program. This program was required to check the records meeting the manual selection criteria for inclusion in RECON against records in existing machine readable data bases to avoid duplicate input. Print Index lists by card number all records in machine readable form in either the MARC I or MARC II data bases. At a later date, the 1968 titles found on the MARC I data base will be processed by a subset of the format recognition program and converted to the MARC II processing format.

The Print Index Program is made up of two routines. The LC catalog card number routine reads each record, extracts the LC card number and creates a magnetic tape file of numbers (called Print Index Tape). The tape created contains a card number right justified for machine sorting, a card number in the same form (zeros deleted) as the number on the printed card, and a data base code indicating the file in which the record originally resided (e.g., MARC II Data Base, MARC II Practice Tape, MARC I Data Base). A parameter card is used to indicate which format and data base is to be processed.

The IBM Sort is used to arrange the output of the LC catalog card number routine into the following order: all 6x-series numbers, all 6x-series numbers with alphabetic prefixes (by year of cataloging—i.e., 1968 followed by 1969), all 7-series numbers (disregarding the check digit, the second digit in the number).

The LC card number print routine prints the card numbers, which are in numeric sequence as described in the preceding paragraphs, from the Print Index Tape. Each page of the listing contains a heading, a running index, a date, and a page number. The program prints 200 card numbers and data base codes per page. The numbers are in ascending order, top to bottom in four columns of 50 numbers each.

Format Recognition

The experience of the Library in the creation of machine readable cataloging records during the MARC Pilot Project and the MARC Distribution Service has clearly demonstrated that the highest cost factor of conversion is the human editing and proofing. The editing presently consists of assigning tags and codes to the bibliographic record to explicitly identify the content of the record for machine manipulation. The

Library has completed a format recognition feasibility study which concluded that the probability of success of automatically assigning tags and codes by computer is high. Since the format recognition feasibility study was only concerned with cataloging records for current English language monographs, the study must be extended to cover other roman alphabet languages and as part of RECON, records which were created according to different rules and conventions.

Although the progress report submitted to CLR included the definition and status of each of the tasks that make up the format recognition program, these have been omitted to avoid duplication with an article recently published in the *Journal of Library Automation* (5) describing format recognition concepts in some detail and elaborating on the tasks completed and projected at that time.

Investigation of Input Devices

The investigation of input devices and the testing of several selected devices in an operational mode will continue throughout RECON. A study of the use of a mini-computer operating in an on-line mode for input, editing, and formatting of MARC records is in progress at the Library and will supplement the RECON effort and provide additional data.

A preliminary investigation was begun of optical character readers commercially available and in the developmental phases. Only those readers capable of reading numerous characters on many lines (page reader) as opposed to a limited number of characters or lines per document (document reader) were included in the study.

The machines evaluated were considered as possible candidates if they were capable of processing upper- and lower-case alphabetic characters, numerals, standard punctuation and some special symbols. Each manufacturer has specifications for the type of paper required and the font style which can be recognized. Paper handling is a major drawback of optical character readers. Excessive handling of the paper or any type of smear, crease, or crinkle could cause rejection of a character or conversion of a character to some specified symbol indicating an invalid character. Error rates for the devices considered range from one to 35 characters per 10,000 characters and 80% of the errors are caused by paper handling. Typewriters used to prepare the source document must be constantly cleaned and ribbons changed to keep impact keys free of dirt. Frequent jamming appears to be a characteristic of most machines; unjamming these machines can be difficult and is highly dependent upon the skill of the operator.

Ten companies that have various types of optical character recognition equipment commercially available were considered in the first study. Five were immediately rejected because their devices did not meet the criteria as specified above.

The devices remaining had the following characteristics:

Control Data Corporation 915 Page Reader. Accepts 2.5x4 to 12x14-inch paper; OCR-A standard type font; recognizes upper-case alphas, numerals, and standard punctuation; through programming and use of special symbols, lower-case alphas can be coded.

Farrington Model 3030. Accepts 4.5x5.5 to 8.5x13.5-inch paper; OCR-A standard and 12L (Farrington) type fonts; recognizes upper-case alphas, numerals, standard punctuation and special symbols; through programming and use of special symbols, lower-case alphas can be coded.

Scan-Data Models 100/300. Accepts 8.5x11-inch paper; multi-type fonts; recognizes upper- and lower-case alphas, numerals, standard punctuation, and special symbols; has programmable unit for formatting.

Philco-Ford General Purpose Reader. Accepts 5.7x8.5x-11 inch paper; multi-type fonts; recognizes upper-case alphas, numerals, standard punctuation and special symbols; through programming and use of special symbols, lower-case alphas can be coded.

Recognition Equipment Retina. Accepts 3.25x4.88 to 14.14-inch paper; multi-type fonts; recognizes upper- and lower-case alphas, numerals, standard punctuation, and special symbols; has a programmable unit for formatting.

The possibility exists of using any of these five machines for the input of English language material. The keying of an extraneous character is required with the Farrington and Control Data Corporation equipment for lower-case and some special symbols. This is not necessary with Philco-Ford, Scan-Data, and Recognition Equipment machines. Since the number of special symbols vary by machine, each machine must be studied to determine a method of coding the entire library character set as developed by the Library of Congress and this method must be evaluated in terms of the burden placed on the typist.

With the added feature of lower-case recognition, the price of the machine increases substantially. Adequate information has not been obtained from these companies to give an accurate accounting of cost. It should be noted that the rental price for the majority of optical character readers is high, a factor which will have to be taken into consideration at the time of selection of an input device. The most economic route to conversion may be through a service bureau, depending on the volume of records to be converted.

OUTLOOK

It is too early in the life of the project to predict the outcome or to describe any factual conclusions. The Library of Congress is greatly encouraged by the interest expressed in the project and the assistance offered by the members of the Advisory Committee and the Working Task Force. The scope of the assignments and the fact that all members of the Working Task Force have responsible positions in their own institutions are clear evidence of the spirit of cooperation that has been exhibited by the Working Task Force members and their parent organizations. Other members of the library community have been and will continue to be contacted throughout the project for their expertise in certain facets of the many problems under exploration.

Several developing regional networks were requested to describe their plans in the hope that smaller scale efforts would shed some light on the problems involved on a national level. Those organizations contacted have responded, and a continuing liaison will be maintained not only to avoid duplication of effort but, more important, to attain a better understanding of how to approach the requirements of future library systems in terms of what is possible today.

The report submitted to CLR described progress made to November 1, 1969. Since that time, the RECON production staff has selected all the 1969 titles from the card stock to be included in RECON, 5,200 records have been edited, and the first 250 have been forwarded to a service bureau to test its procedures for keying. The staff has begun the selection of the 1968 titles and out of approximately 26,000 records received to date from the Card Division 19,000 are RECON candidates.

The production section continues its training by the proofing of MARC records until the RECON records are processed through the MARC system to provide the required diagnostics for the proofing process.

Procedures were set up for typing records without any editing and in accordance with the requirements for the format recognition program. Sample records selected for testing the procedures were of above-average difficulty in order to include all types of data that might be encountered. The procedures will be continually evaluated until some optimal method is determined.

The format recognition algorithms are being evaluated by having RECON staff simulate a computer and follow through the logic of the algorithms on actual data. Results of the simulation will provide the necessary feedback to adjust the algorithms prior to the coding of the computer programs.

Detailed design work has begun on the expansion of the MARC system to include random access capability and on-line correction. This effort is being coordinated with the Card Division Mechanization Project and is considering the requirements of a large-scale conversion activity.

Although it has a long way to go, RECON is on schedule and for any project concerned with automation, that is an encouraging note. For the moment the future looks bright.

ACKNOWLEDGMENT

The author wishes to thank the RECON staff members of the Library of Congress for their respective reports which were incorporated into the progress report submitted to the Council on Library Resources, Inc., and as such, are significant contributions to this paper.

Without the aid of the Council on Library Resources the RECON Project would not have become a reality. Through three important grants the Council has made a major contribution to the Project: 1) the first was a grant in support of the RECON Feasibility Study and the Working Task Force that resulted in the RECON Report; 2) an Officer's Grant enabling the establishment of the RECON Production Unit to create additional machine readable records not included in the MARC Distribution Service; and 3), most importantly, a grant providing full funding for the two-year Pilot Project.

REFERENCES

1. Library of Congress; RECON Working Task Force: *Conversion of Retrospective Catalog Records to Machine Readable Form.* (Washington: Library of Congress, 1969).
2. Ibid, pp. 10-11.
3. Ibid, pp. 20-38.
5. *Anglo-American Cataloging Rules.* (Chicago: American Library Association, 1967).
4. Avram, Henriette D., et al.: MARC Program Research and Development: A Progress Report," *Journal of Library Automation,* 2 (December 1969), 242-265.