RECON PILOT PROJECT: A PROGRESS REPORT, APRIL-SEPTEMBER 1970

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

A synopsis of the third progress report on the RECON Pilot Project submitted by the Library of Congress to the Council on Library Resources. An overview is given of the progress made from April through September 1970 in the following areas: RECON production, format recognition, research titles, microfilming, and investigation of input devices. In addition, the status of the tasks assigned to the RECON Working Task Force are briefly described.

INTRODUCTION

The RECON Pilot Project was established in August 1969 to test various techniques for retrospective conversion in an operational environment and to convert a useful body of records into machine readable form. It is being supported with funds from the Council on Library Resources, the U.S. Office of Education, and the Library of Congress. This article summarizes the third progress report of the pilot project submitted by the Library of Congress to the Council and has addressed itself to all aspects of the project, regardless of the source of funding, in order to present a meaningful document.

Two previous articles in the *Journal of Library Automation* summarized the first and second progress reports, respectively (1), (2). This article describes the activities occurring April through September 1970.

PROGRESS—APRIL THROUGH SEPTEMBER 1970

RECON Production

At the present time, the RECON data base contains approximately 20,000 records. It appears that the original estimates on the number of titles to be input during the RECON Pilot Project were considerably higher than the actual number found to be eligible. This situation occurred because of the following circumstances:

1) The original estimates were derived from the number of English language monographs cataloged during 1968 and 1969. Since the MARC Distribution Service began in March 1969, it was felt that the number of titles eligible for RECON in the 1969 and 7-series of card numbers would be equal to the number cataloged during January-March 1969. In actuality, the titles cataloged during this period were primarily records with 1968 card numbers.

2) The estimate of records with 1968 card numbers was higher because it was thought that many more of these titles had been through the cataloging system than were actually processed prior to the beginning of the MARC Distribution Service. Instead of being included in RECON, these records have been input into the MARC Distribution Service.

In order to obtain 85,000 records for conversion, several alternatives, including the conversion of English language monographs in the 1967 card series, are being studied.

Format Recognition

Format recognition is a technique that will allow the computer to process unedited catalog records by examining data strings for certain keywords, significant punctuation, and other clues to determine the proper content designators. This technique should eliminate substantial portions of the manual editing process and, if successful, should represent a considerable savings in the cost of creating machine readable records.

The logical design for format recognition has been completed, and the manual simulation to test the efficiency of the algorithms was described in an earlier article (3). Completion date for the programs is expected in February 1971.

The programs were designed in several modules so that they could be adapted for different input procedures without disturbing the logic. Once the programs have been implemented, tests may show that certain fields should be pretagged because the error rate is too high or the occurrence of the field is too low to justify the processing time. The complete logical design for format recognition has been published as a separate report by the American Library Association (4).

As part of a manual simulation to test the format recognition algorithms, one hundred fifty records for English language monographs were typed on an MT/ST, a typewriter-to-magnetic tape device. The MT/ST hardcopy output was used as the raw data for the simulation. The results of the test were analyzed for possible changes to the algorithms, keyword lists, or input specifications. Then the records with the content designators assigned by the format recognition algorithms were retyped and processed by the existing MARC system programs. Proofsheets were produced and given to the RECON editors for proofing, a process to verify content designators and bibliographic information.

Each editor proofed all of the format recognition records; their hourly numbers of records proofed were as follows: highest, 9.3; lowest, 5.3; average, 6.8. The average number of current MARC records edited and proofed in an hour is 4.8.

When format recognition is implemented, present workflow—editing, typing, computer processing, proofing—will be replaced by a new one—typing, format recognition, proofing. In comparing production rates in the two systems, time needed to proof format recognition records must be compared against time needed to edit and proof in the current system.

Several factors should be considered when evaluating this portion of the simulation experiment. Although all the records chosen for the test were of English language monographs, they were generally more difficult than those encountered in a normal day's work for both editors and typists. In addition, numerous errors were made by the human simulators, such as omission of subfield codes, delimiters, or fixed field codes.

Format recognition does appear to have reduced the amount of time spent in the combined editing and proofing process, but the success of the program depends heavily on the following factors: 1) extensive training for the input typists with greater emphasis placed on their role in this project; and 2) extensive training for the editors to alert them to kinds of errors the format recognition programs might make.

Proofing time for the test was greater than anticipated. With fewer errors from the typing input and the elimination of human errors from the simulation, it is possible that the proofing rate will be higher under actual work conditions. Editors might reach an average of 9.3 records proofed, or double the number presently done in a combined editing/proofing process.

Two programs are being written to support the format recognition project. Format Recognition Test Data Generation (FORTGEN) will provide test data for format recognition by stripping MARC records of delimiters, indicators, and subfield codes, and reformatting the data to be identical with the product from the initial input program. Thus, a large quantity of high quality test data can be provided without additional keystroking.

The Keyword List Maintenance Program (KLMP) maintains approximately sixty keyword lists used by the format recognition program in processing bibliographic data. These lists are maintained as a separate data set on a 2314 disk pack. The actual lists themselves, along with associated control data, are referred to as "keyword list structures." The general function of KLMP is to read the entire set of keyword list structures from the file on disk, modify them as specified by parameter cards to KLMP, and write a new file on disk. The individual actions performed by KLMP are as follows: 1) create a list; 2) remove a list; 3) add a keyword; 4) delete a keyword; 5) augment a table (translation tables to generate codes such as Geographic Area Code, Language, Place of Publication); and 6) list structures (printout of all or selected portions of a list).

Since the keyword lists will be dynamic in nature, this program provides the flexibility required to change or update them without recataloging the entire format recognition program. New lists will be added as format recognition is extended to other languages, and keywords will be added to or deleted from existing lists as experience is gained in the use of format recognition.

Research Titles

Since the production operations of the RECON Pilot Project have been limited to English language monographs in the 1968, 1969, or 7-series of card numbers, it was recognized that many problems concerning retrospective records would not be revealed in the conversion of relatively current titles. For this reason, a project to identify and analyze 5,000 research titles was included as part of the pilot project. These research titles would consist of records for older English language monographs and foreign language monographs in roman alphabets and would be studied for problems in the following areas: 1) earlier cataloging rules which caused certain elements to be omitted from the record or transcribed in a different style; 2) different printed card formats which placed elements in different locations; 3) difficulty in working with foreign languages when converting records to machine readable form; 4) problems arising from shared cataloging records; and 5) problems arising when expanding the format recognition algorithms to cover these kinds of records.

The selection of these records was described in an earlier article (5). The initial analysis of the research titles has been completed, and a few of the problems encountered are listed as follows:

1) Ellipses at the beginning of a title field (. . . *Dictionnaire-manuel-illustré des écrivains et des littératures)* were used frequently on older cataloging records. Since they are no longer prescribed by the present cataloging rules unless they appear on the title page at the beginning of a title, it was recommended that such ellipses be deleted from the machine record because they would affect the format recognition algorithms.

2) Card numbers without digits representing the year (F-3144) were assigned during 1901. Generally, these numbers appear with an alphabetic prefix representing the language of the publication or the classification number. It has been recommended that such numbers be revised to read "f01-3144" for the machine record.

3) Records cataloged under the 1908 *A.L.A. Catalog Rules* included in the series statement such information as the editor of the series or the location of the series statement *(Half-title:* Everyman's library, ed. by Ernest Rhys. Reference). It has been recommended that such information be deleted from the machine record.

4) An asterisk preceding personal name added entries (I. *Spence,

Lewis, 1874- joint author.) indicated that the name had appeared in a fuller form at an earlier date; if this name were used as the main entry, there would have been a corresponding full name note at the bottom of the catalog card. It has been decided that this asterisk will be deleted from the machine record.

5) The national bibliographies from which shared cataloging copy is derived use punctuation conventions which differ from the AA Rules. For example, the West German bibliography uses parentheses to indicate that the data are not on the title page, brackets to indicate the data are not in the publication, and angled brackets to indicate that the data are enclosed in parentheses on the title page (<22.-27. Mai 1967>. Köln ([-Ehrenfeld] Bundesinstitut für Ostwissenschaftliche und Internationale Studien) 1967). Such conventions would affect the expansion of the format recognition algorithms to foreign languages. This is an area in which the Standard Bibliographic Description would be of great value.

6) In the MARC II format, each place of publication is a separate subfield so that when each place is connected by hyphens (Milano-Roma-Napoli . . . ,), there would be a problem in inputting the data and having the data printed out in the same fashion. It has been recommended that each place of publication be separated with a comma instead of a hyphen (and the ellipsis deleted from the imprint statement).

7) Conjunctions have been used between places of publication on records cataloged according to the 1908 rules and on some shared cataloging copy (London, Glasgow and Bombay) (Neuwied a. Rh. u. Berlin). In the machine record, each place is a separate subfield, and the presence of a conjunction means that one subfield contains non-essential data. It has been recommended that conjunctions be omitted from the machine record and that places of publication be separated by commas.

8) The *A.L.A. Cataloging Rules for Author and Title Entries* states that with certain well-known persons, dates of birth and death can be omitted when the heading is followed by a subject subdivision (1. Shakespeare, William—Language—Glossaries, etc.). Since the rules provide a list of such persons, it has been recommended that when such names are used as subject headings, they should include dates of birth and death in the machine record.

9) A collation statement like the following (25 p., 27-204 p. of illus., 205-232 p., 233-236 p. of illus., 237-247 p. 28 cm.) would cause the format recognition algorithms some difficulty in identifying the proper subfields. This is another area in which the adoption of a Standard Bibliographic Description would aid format recognition programs.

10) Both East and West German bibliographies give information about illustrations in the title paragraph rather than in the collation (Title paragraph: [Mit] 147 Abbildungen und 71 Tabellen. Collation: xii, 418 p. 26 cm.). The cataloging policy at the Library has been revised so that

on current cataloging records information about illustrations is also repeated in the collation. It has been recommended that for retrospective records the data should be input as it appears on the catalog card. In this example, the machine record would not contain illustration information in the collation.

11) The method of transcribing non-LC subject headings has been changed in recent years, and the MARC II format reflects this change. In previous years, the following conventions were used: subscript brackets enclosed headings or portions of headings that were not the same as the LC form; subscript parentheses enclosed portions of headings that were the LC form but not the contributing library's; if two headings had the same number, the LC form was listed first; if both forms of the heading were the same, there would be only one number, and the heading itself would not have the subscript brackets or parentheses. It has been recommended that either the non-LC forms be deleted from the machine record or the transcription of such subject headings be revised to follow the current practice.

12) NLM subject hearings have different capitalization conventions from those used by LC, and the geographic subject subdivisions are often in a form different from that which the Library of Congress uses ([DNLM: 1. Public Health Administration—U.S.S.R. W6 P3]). In analyzing these research titles in terms of possible problems with format recognition, it was discovered that NLM subject headings would be incorrectly identified for the above reasons. Format recognition depends heavily on capitalization and keyword lists; in this example, the heading "Public Health Administration" would be identified as a corporate name because of the capitalization.

Examinination of the research titles showed the similarity of the cataloging of the older records (pre-1949) and the current foreign language records based on shared cataloging copy. Certain stylistic conventions, such as the use of ellipses or the transcription of imprint statements, were similar for both kinds of material. It would be necessary to have a thorough knowledge of the *ALA Catalog Rules* (published in 1908) in order to interpret the data on the older printed cards correctly during a conversion project.

The experience of the editors in the RECON Production Unit has been that retrospective records, even those cataloged during the last two years, require a considerable amount of interpretation in order to assign the correct content designators in the fixed fields. For pre-1949 records, the problem becomes more acute when one attempts to apply the procedures and techniques for current material to older records. It is very likely that a higher level of personnel would be required to process these records because in many instances the changes would be similar to recataloging the entire record.

The expansion of format recognition to foreign languages would be extremely difficult without a greater degree of consistency in shared cataloging copy. Each national bibliography, from which the cataloging copy is derived, has its own rules and style of cataloging, so that although the language of the works may be the same, e.g., German, the entries from the West German, East German, Austrian, and Swiss bibliographies may differ in terms of punctuation or style of cataloging. These problems have been compounded by printer's errors on the printed cards as the result of conventions that differ from the AA Rules. The adoption of the Standard Bibliographic Description (6) would be a tremendous aid in interpreting cataloging data by both humans and format recognition programs.

Microfilming Techniques

The Library's Photoduplication Service is supporting the RECON Pilot Project by providing the cost estimates for the various alternatives of microfilming techniques and providing technical guidance as required. Several discussions with them confirmed that the method of filming a portion of the record set containing the subset of records to be converted first and selecting the appropriate records afterward would be more advantageous than selection prior to microfilming (7).

It was considered unrealistic to attempt to project microfilming costs for the entire RECON effort. Because of the paper handling problems involved in the management of input worksheets, the microfilming rate should be in reasonable proportion to the actual conversion rate. There is no point in providing a huge supply of input worksheets which will not be used in actual conversion for a long time. The data may become "dated," and there may be storage and handling problems. In addition, cost estimates provided by the Photoduplication Service can only be expected to prevail over the next twelve months. Beyond that period, any quotation given is likely to be higher because of the general trend of rising costs.

Any projection of costs should be based on a manageable portion of the whole. Just what this portion should consist of has yet to be determined. Assuming a *modus operandi* as described above, there is needed a determination of the "rate floor," which is defined as the minimum number of records that must be microfilmed to achieve the maximum cost benefits resulting from a relatively high volume job. Once the rate floor is determined, it should probably be translated into year equivalents, i.e., if the rate floor is 100,000 and the catalog card production is 50,000, then two years' worth of cards would be microfilmed. Estimates would be obtained for the following alternatives: microfilming for OCR device specifications; microfilming for reader-printer specifications; microfilming for reader specifications; and microfilming for Xerox Copyflo printouts of the LC printed cards onto RECON worksheets.

Certain ground rules were assumed for the actual microfilming process. The selected drawers of the record would be "frozen" for a day or two prior to being filmed, i.e., the file would be complete and no one would remove cards from the file while filming was in process. The filming would take place during the day. Assuming that 100,000 cards for the year 1965 would be used as a base figure and that approximately 5,000 cards per day can be filmed with a planetary camera, it would take twenty working days to film the collection of cards for one year in the record set (rate floor as defined above). All cost estimates will include quality control; i.e., quotations would indicate degree of inspection of film for technical quality and degree of preparation of the file before filming.

Input Devices

During 1969 the Library of Congress conducted an investigation to determine the feasibility and desirability of using a mini-computer for MARC/RECON input functions (original input and corrections). This study was performed with contractual support and consisted of three basic tasks: 1) analysis of present operations to determine functional requirements, to measure workloads, and to identify problem areas; 2) survey and analysis of mini-computers that are potentially capable of meeting the requirements of the present operations; 3) evaluation of available hardware and software capabilities relative to MARC data preparation requirements and determination of economic feasibility based on present and projected workloads.

The intent of this study was to provide a basis for future planning and procurement activities by the Library of Congress relative to improvement of the MARC/RECON man-machine interface. The survey of hardware was not intended to be all-inclusive. There were time and funding limitations, and in addition it was recognized that the mini-computer field was a rapidly expanding one; therefore, it was not possible at any cut-off point to have surveyed the totality. Six firms were included in the survey, and the machines considered were the Burroughs TC-500, the Digital Equipment Corporation PDP-8/I, the Honeywell PDP-516, the IBM 1800, the Interdata Model 4, and the XDS Sigma 3. Of these, the DEC PDP-8/I and the Honeywell PDP-516 were determined to have the highest potential for meeting MARC/RECON requirements.

Additional analysis revealed that software availability for mini-computers is minimal. Manufacturers covered in this investigation supplied an assembler as well as testing and editing routines. Some provided a FORTRAN, ALGOL, or BASIC compiler and an operating system with foreground/background processing. Systems that support FORTRAN and the operating system are quite substantial, generally requiring 16,000 words of core, memory protect, disc, etc. The cost of this kind of system is generally a minimum of $10,000.

Few low-cost peripheral devices are available for use with mini-computers. High-speed tape readers, punches, and punched card readers are the most inexpensive input/output devices available. The addition of a magnetic tape unit to most systems significantly increases the overall cost.

The conclusion reached as a result of this investigation was that there is no gain, either technically or economically (considering the hardware configuration of the Library of Congress), to using a mini-computer in performing present MARC/RECON functions.

Another input device investigated during this reporting period was the Keymatic Data System Model 1093, which was selected for a two-month test and evaluation period because it appeared to have the following advantages for the recording of bibliographic data: 1) this device has 256 unique codes; 2) data is recorded directly on computer compatible magnetic tape; 3) through manufacturer supplied software, the user may assign to certain keys, called expandables, the value of whole strings of characters; thus a single key would equate to a MARC tag; 4) correction procedures are built into the device, i.e., the ability to delete a character, word, sentence, or entire record; and 5) the single character display screen obviates the necessity for hard copy. It is often claimed that hard-copy output is scanned by the typist unintentionally to the detriment of typing rates.

The machine tested was specifically set for the Library's requirements. Four separate keyboards contained 184 keys, of which 103 had upper- and lower-case capability, and the remaining 81 had only a single case. The 256 possible codes were divided into the following categories: 1) 94 were used as expandables and assigned to those MARC tags and data strings (correction and modification symbols) that appear most frequently; 2) 10 were used as machine function codes; 3) 150 were assigned unique values in the MARC character set; and 4) 2 were left unused.

The keys on the four keyboards were assigned values such that the most frequently used keys were located in a strong stroke area. The main character keyboard was designed to be closely compatible to the device currently in use at the Library to lessen the training requirements for the typist. Therefore, the typist had only to learn the expandable keys and some lesser used special characters. The program supplied by the manufacturer was modified for code conversion and output format acceptable to the MARC system and to conform to the Library's computer system assignments.

The two typists selected to participate in the test were both experienced MARC production typists. Both typists were given individual instruction on the machine and spent three weeks practicing; at the same time, their performance was being analyzed and discussed with them. During the official evaluation period, the typists spent two weeks working full time on the machine. When the typists began their practice period, their speeds were relatively slow, 6-7 records per hour. As time progressed, their speed increased, leveling off to approximately 11-12 records per hour by the end of the test period.

Each typist reported problem areas during the official evaluation. One problem was the hesitation which resulted when the typist had to determine whether to use an expandable key or actually type the data, character by character. If she chose the former, the expandable key had to be found. The number and different combination of tags caused some confusion. The opinion of both typists concerning the keyboard arrangement was that they would rather type the tags character by character than search for the expandable key. More experience on this device might eliminate this problem.

The absence of hard copy was felt to cause another problem. When a typist intuitively feels that she has made an error in current MARC/RECON typing operations, she uses the hard copy to verify that a mistake has actually been made prior to taking corrective action. The lack of hard copy did not allow for this verification, and the typists reported that this detracted from their efficiency.

The following table lists the results of the official evaluation period. The average production rate of these two typists on the MT/ST is also listed. The figures for MT/ST production have been calculated for a particular three-week period.

	Typist A	*Typist B*	*Total*	*MT/ST*
New records	505	540	1045	1995
Correction records	323	278	601	
Verified records	58	537	595	
Average records/hour—new	10.1	14.0	12.1	14.6
Average records/hour—corrected	21.3	27.7	24.5	
Keystrokes				
Total	238,435	259,630	498,065	
Expandables Used	12,280	14,646	26,926	

The Keymatic model used for the test rents for $768.25 per month (July 1970 pricelist). It is a fully equipped model with several options not required for the MARC system. Without these options, a less expensive model could be used. Keymatic does have a 24-month lease plan in which the basic machine could be rented for $368.00 per month. This is an increase of $258.00 per month per machine over the current method of input.

Costs per record were computed for the Keymatic device and for the MT/ST based on the average record statistics of both typists. Although the same records were not actually typed on the MT/ST, extensive experience with production and error rates on that device made it valid to use average production rates for purposes of comparison.

For purposes of computing the cost per record, the hourly cost per machine was calculated by dividing the cost per machine by 160 working hours. The 24-month leasing price of $368.00 per month was used for the Keymatic, resulting in a machine cost per hour of $2.30. The MT/ST rental cost is $110.00 per month, resulting in an hourly cost of $.69. (The cost of the MT/ST listed in a previous article (8) as being $100.00 was in error.) On the basis of 12.1 records per hour on each device, the cost per record for the Keymatic is $.19 and $.06 for the MT/ST.

In the context of the Library of Congress MARC/RECON Project, the addition of a Digi-Data to translate MT/ST output to computer compatible tape adds an incremental cost to each input device. For the purposes of this report, it was assumed that the project required five input devices. On this basis, the prorated Digi-Data cost per hour is $.33, which makes the total machine cost per hour for the MT/ST as $1.02. Thus, the cost per record for the MT/ST becomes $.08.

The results of the test indicated that the Keymatic used in the Library of Congress environment did not substantially increase production rates or decrease error rates. Thus, no savings in cost were demonstrated. The complex data to be typed and the construction and quality of the worksheets at the Library of Congress impose severe constraints on all machines. (The manuscript card reproduced on the MARC/RECON worksheet results in a source document that is difficult to work with for the following reasons: 1) loss of legibility during the copying process; 2) position of tags in relation to content; and 3) combination of typed and handwritten data as recorded by the catalogers.) In order to make a fair comparison between the Keymatic and the MT/ST, the manuscript card was used for the test rather than the printed card. If, on evaluation, the Keymatic proved to be more efficient than the MT/ST using the manuscript card, it would be even more effective if the printed card were used, since the latter is a far more legible source document.

Keymatic does have a new machine, Model K-103, which has an 80-character visual display option which might correct one of the objections raised by the typists, i.e., lack of hard copy; however, this model requires the use of a converter as does the MT/ST. This device is less expensive than the machine used in the test and may be evaluated during the RECON Project at a later date.

An investigation of Model 370 CompuScan was continued following the initial findings reported in a previous article (9). Twenty-five letterpress Library of Congress printed cards representing English language titles and containing no diacritical marks in the content were sent to the firm for input. This allowed the machine to be evaluated and problems noted within an "ideal" test environment. Depending on these results, further testing could be performed.

Since existing CompuScan software was used to conduct the Library of Congress test, the entire LC card could not be read but only that portion that contained fonts already built into the existing configuration. The printed cards were blocked out, except for the area covering the body of the entry, i.e., title through imprint, prior to microfilming for subsequent scanning.

Operator intervention was required on approximately 1%-25% of the characters on each card. In addition to the problems offered by variant and touching characters, fine lines in certain characters caused a misreading by the machine. This was particularly true with the letter "e" being interpreted as the letter "c." CompuScan felt this problem might be resolved by increasing the size of the comparison matrix of the hardware. In some instances, a period was generated in the middle of a word due to the coarseness of the card stock that was microfilmed.

Initial discussions have begun on the possibility of testing a retyped version of the printed card. The only rationale behind this test would be to investigate if typing for a scanner that could read upper-and lower-case and special characters made any significant difference in speed and/or error rate compared to costs and production rates of typing for a scanner which could read only upper-case characters. The latter was described in an earlier article on RECON (10).

RECON Working Task Force

The Working Task Force continued the discussion on the implications of a national union catalog in machine-readable form. From the postulated reporting system for a future NUC described in a earlier article (11), several items were isolated for further consideration. These included: 1) grouping of records in a register (by language, alphabet, etc.) to allow for a segmented approach to computer-produced book catalogs (a register is defined as a printed document containing the full bibliographic descriptions of works sequenced by unique identification numbers. As each record is added to the register, it is added at the end and assigned the next sequential identification number); 2) the need for additional indexes to the register by LC card number and classification number (the class number was not included in the list of data elements required for the machine-readable NUC); 3) the requirement to include the author statement in the title index versus using the main entry in all cases; and 4) clarification of subject index to mean only topical or geographic subjects.

The following tasks were outlined for further consideration: 1) Format of the printed NUC (graphic design and printing, size, style, typographic variation, etc.); 2) Physical size of the volume depending on pattern of distribution (monthly, bimonthly, etc.); 3) Input (relationship to MARC input, use of format recognition, problems of languages in terms of selection for input); 4) Output (cost of production for register and indexes, cost of sorting, costs of selection, etc.); 5) Cumulation patterns in terms of cost and utility (number of characters in an average entry, number of items on a page, rate of increase, etc.); 6) The use of COM (Computer Output Microfilm) as an alternative to photocomposition for printed output.

Work on Task 3, the investigation of the possible use of existing data bases in machine readable form for a national bibliographic service, has been continued. Phase 1 of this task consisted of a survey of existing machine readable data bases. Selection of data bases for analysis was based on the following criteria: 1) The data base had to include monograph records. 2) Any data base known to have predominantly LC MARC records was excluded. 3) The data base had to be potentially available to RECON (security organizations or commercial vendors might not be willing to give their files to a RECON effort). 4) Data bases of less than 15,000 records were excluded.

A data analysis worksheet was prepared to reduce the documentation to a standardized form for each system studied in the survey. It was initially anticipated that once documentation was received from the various institutions, additional contact would be made via telephone or on-site visits. This proved to be unnecessary, as the submitted documentation was generally sufficient. Since many of the formats submitted were complicated, errors could have been made in interpretation; however, this possibility was not considered important enough to affect the findings of this task. If necessary, additional information can be requested from the library systems at a later date. The analysis of the submitted documentation was difficult for the following reasons: 1) The amount of documentation ranged from extremely detailed to very sparse; 2) Neither the technical nor the bibliographic terminology was consistent for all organizations; 3) In some instances, the format descriptions were more detailed with respect to control and housekeeping data fields than bibliographic data fields.

The formats were ranked according to three broad categories: low potential, medium potential, and high potential. To arrive at a ranking, the data fields of each format were compared to the MARC II format. Comparison was made on the following basis: 1) present in both formats; 2) not present in local format and not capable of generation by format recognition algorithms; or 3) not present in local format but capable of generation by format recognition.

The result of this analysis distributed the twenty-two institutions into the following ranked order: 1) Low potential—3; 2) Medium potential—8; 3) High potential—11.

The figure for the number of low potential data bases is in addition to the eight out of the eleven originally rejected due to a small data base or very limited content in the record. It is significant to note that although no attempt was made at an all-inclusive survey of machine readable data bases, the total number of records in machine readable form reported by the respondents amounted to approximately 3.7 million of all types. Of this figure, about 2.5 million represented monograph records.

The Phase 1 study included procedures required to transform a record into a certified RECON record, thus outlining the areas requiring cost analysis to compare the economics of using existing files for a national bibliographic store, as opposed to original input. (Certification in this context means comparing the record of the local institution to the record in the LC Official Catalog and, if required, making the record consistent with the LC cataloging as well as upgrading it to the bibliographic completeness of the LC record. Input in this sense includes the editing of the record as well as the keying.) The results of the study, prior to any further analysis, seems to indicate that the next phases of Task 3 will concentrate on a very large data base with a high degree of compatibility with MARC II (high potential) and another data base with a format differing from MARC II both in level of explicit identification and in bibliographic completeness (medium potential). The first data base tests the most favorable situation; the latter a much less favorable situation.

The carry-on phases of Task 3 will include: 1) a determination of a cut-off point at which a particular data base would not be included in future studies (although the composition and the format of the records in the data base might fit the selection criteria, the number of records in the file might be insufficient to warrant the costs of the hardware/software for the conversion effort); 2) investigation of the hardware and software effort involved; and 3) determination of the costs of comparing the records with the LC Official Catalog and the resultant updating costs to bring the records up to the level of the records in the LC machine readable MARC/RECON data base.

ACKNOWLEDGMENTS

The authors wish to thank the staff members associated with the RECON Pilot Project in the MARC Development Office, the MARC Editorial Office, and the Technical Processes Research Office in the Library of Congress for their contributions to this report. The LC Photoduplication Service provided valuable assistance in certain phases of this project.

Work on the RECON Pilot Project has continued to be supported by the Council on Library Resources and the U.S. Office of Education.

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. Ibid., p. 235
4. U.S. Library of Congress. Information Systems Office. *Format Recognition Process for MARC Records: A Logical Design.* Chicago: ALA, 1970.
5. Avram, Henriette D.; Guiles, Kay D.; Maruyama, Lenore S. Op. cit., p. 236.
6. Ibid.
7. Ibid., p. 237.
8. Ibid., p. 246.
9. Ibid., pp. 244-245.
10. Ibid., pp. 245-248.
11. Ibid., p. 248.