DOCUMENT RESUME

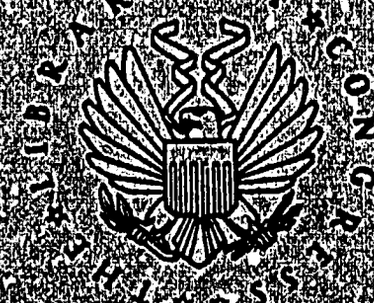
ED 082 793 LI 004 515

AUTHOR Avram, Henriette D.
TITLE RECON Pilot Project. Final Report.
INSTITUTION Library of Congress, Washington, D.C.
SPONS AGENCY Council on Library Resources, Inc., Washington, D.C.; Office of Education (DHEW), Washington, D.C.
PUB DATE 72
NOTE 54p.; (0 references)
AVAILABLE FROM Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 (Stock No. 3000-00061; $1.50)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Bibliographic Citations; *Cataloging; Computer Programs; Costs; Data Bases; *Information Processing; Input Output Devices; *Library Automation; Library Technical Processes; Microfilm; Pilot Projects
IDENTIFIERS *Library of Congress; Machine Readable Cataloging; MARC; RECON

ABSTRACT

One of the specific recommendations in the Retrospective Conversion (RECON) feasibility report (ED 032 895) was that a pilot project be established to test various conversion techniques, ideally covering the highest priority material (English-language monograph records from 1960-68). A two-year pilot project was initiated in August 1969. This report is oriented toward the work of the project as a whole. The pilot project conducted at the Library of Congress covered five major areas: (1) testing techniques postulated in the RECON report in an operational environment, (2) development of procedures and computer programs to implement format recognition, (3) analysis of techniques for the conversion of older English-language materials in foreign languages using the roman alphabet, (4) monitoring the state-of-the-art of input devices that would facilitate conversion of a large data base, and (5) a study of microfilming techniques and their associated costs. The accomplishments of the pilot project are discussed in detail in this document. (Author/SJ)



RECON Pilot Project

RECON Pilot Project

Final Report on a Project Sponsored by the Library of Congress, the Council on Library Resources, Inc., and the U.S. Department of Health, Education, and Welfare, Office of Education

Prepared by Henriette D. Avram
Project Director
MARC Development Office

Library of Congress Washington 1972

Library of Congress Cataloging in Publication Data

RECON Pilot Project.
RECON Pilot Project.

1. MARC project. I. Avram, Henr'ette D.
Z699.4.M2R4 029.7 72-7314
ISBN 0-8444-0034-3

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402 - Price $1.50
Stock Number 3000-00061

Foreword

Since March 1969 the Library of Congress has been converting its bibliographic records for currently cataloged English-language monographs into machine-readable form for dissemination to the library community through the MARC Distribution Service. During fiscal 1972 this program was expanded to include motion pictures and filmstrips. Monograph records in French will be added in fiscal 1973, provided the necessary funding is available, and plans for future expansion include adding records in German, Spanish, and Portuguese. Thus, the prospects for centralized conversion of catalog records for current materials are encouraging.

There has also been widespread interest in centralized conversion of retrospective records. The Library of Congress, whose concerns in this respect include both its own requirements and those of the library community, proposed to the Council on Library Resources that a study be conducted to determine the problems associated with centralized conversion of retrospective catalog records and distribution of these records from a central source. Funds to support such a study were granted to the Library, and direct responsibility was assigned to the RECON (*Re*trospective *Con*version) Working Task Force. The task force's major conclusions and recommendations were presented in a report entitled *Conversion of Retrospective Catalog Records in Machine-Readable Form; a Study of the Feasibility of a National Bibliographic Service.* One recommendation was that a pilot project be undertaken to test empirically the techniques suggested in the feasibility study and, at the same time, to convert a useful body of data. Proposals were submitted to the Council on Library Resources and the U.S. Office of Education, and these organizations agreed to provide support for both the pilot project and the continuation of the activities of the RECON Working Task Force.

Most of the people who have served on the advisory committee and task force for the RECON feasibility study agreed to participate in the RECON Pilot Project and are to be commended for continuing their contributions to a project of national scope.

This report describes the pilot project conducted by the Library of Congress staff. A subsequent publication will present the results of the studies conducted by the RECON Working Task Force. In light of the problems encountered during the pilot project, the prospects for a large-scale retrospective conversion activity do not appear encouraging at present. Nevertheless, the results of the project have far-reaching implications for the conversion of current material and for future activities, in both manual and machine systems, of the library community. The profession is urged to study this report and to comment on the findings so that future planning and implementation will continue to be responsive to the most criticial requirements of libraries and their users.

John G. Lorenz
Deputy Librarian of Congress
Officer-in-Charge, RECON *Pilot Project*

Acknowledgments

Projects are conceived, proposed, initiated, supported, conducted, evaluated, and reported on by individuals. As director of the RECON Pilot Project and chairman of the RECON Working Task Force, I would like to express my appreciation not only to the organizations but also to the people that have contributed to the successful completion of the project.

For funding, I am indebted to the Council on Library Resources, Inc., and to the U.S. Office of Education. Fred C. Cole, President of the Council, made possible the prompt initiation of the project, with an initial officer's grant.

Members of the RECON Advisory Committee, with John G. Lorenz, the Deputy Librarian of Congress, as chairman, provided valuable counsel during the course of the project. Thanks are due also to the directors of those organizations that generously contributed the time of the members of the Working Task Force and to the members themselves for their involvement with the pilot project as well as with the research studies.

The support and interest of the Librarian and Deputy Librarian of Congress and of William J. Welsh, Director of the Processing Department, are gratefully acknowledged. Within the Processing Department, the MARC Development Office, the MARC Editorial Office, the Technical Processes Research Office, and the Card Division all played significant roles in both the operational and the research aspects of the project. Among the many staff members who contributed to the project, the following persons merit special mention: Lucia J. Rather, T. Arlene Whitmer, Lenore S. Maruyama, Patricia E. Parker, Kay D. Guiles, and Ivey S. Andrews of the MARC Development Office; Barbara J. Roland and Margaret Patterson of the MARC Editorial Office; and John C. Rather and Susan C. Biebel of the Technical Processes Research Office. The extensive technical experience of Charles LaHood and Robert Sullivan of the LC Photoduplication Service was drawn upon in the investigation of microfilming techniques and costs.

Several contractors performed valuable services for the project. Ken Benson of Input Services, Inc., was contractor for part of the actual RECON production and the experiments in different methods of conversion. Coyle and Stewart, Computer Application Consultants, performed the logical analysis and the programming for format recognition. Josephine S. Pulsifer of Becker and Hayes, Inc., assisted in the preparation of sections of this report.

Finally, I wish to thank the Publications Office of the Library of Congress for its assistance in the publication of the final report and the editors of the *Journal of Library Automation* for publishing several progress reports on the project.

Henriette D. Avram

Table of Contents

CHAPTER 1

Introduction

Availability of machine-readable catalog records from a central source has long been considered a necessary condition for effective application of computer techniques in libraries. A significant step in this direction was taken in November 1966, when the Library of Congress began distributing MARC records for English-language monographs as part of the MARC Pilot Project. The success of the pilot project led to the implementation of the MARC Distribution Service in March 1969, and since that time over 60 subscribers have received approximately 200,000 MARC records representing the current English-language monograph cataloging at the Library of Congress.

When the MARC Distribution Service expands its coverage to catalog records for foreign-language monographs and for other forms of materials, libraries will be able to obtain machine records for a large number of their current titles. Obtaining machine-readable data for retrospective cataloging, however, remains a very serious problem. Recognizing the need for more research in this area, the Council on Library Resources provided funds to the Library of Congress, and in November 1968, a working task force of librarians and systems analysts representing various types of libraries began a study of the feasibility of converting retrospective catalog records, which became known as RECON (*Re*trospective *Con*version). The final report of the RECON Working Task Force was published by the Library of Congress in June 1969.[1]

The RECON feasibility report addressed itself to the following areas: 1) the state-of-the-art of hardware and software applicable to large-scale conversion, storage, and retrieval of retrospective bibliographic information; 2) the organizational and administrative aspects of a conversion project, including identification of the most suitable existing file for conversion, determination of which segments of that file should have the highest priority for conversion, and development of an effective methodology to accomplish the tasks associated with the conversion process; 3) costs of hardware, software, and manpower as well as timing and funding for such a project; and 4) identification of areas that require intensive additional study. The report also included analysis of 1) user needs for retrospective cataloging data; 2) means of maintaining standardization of the format for machine-readable records to allow libraries to exchange information in this form; and 3) systems design and software required to create, maintain, and disseminate information from a large data base.

In its original feasibility study, the RECON Working Task Force reached the following conclusions:

1) The MARC Distribution Service should be expanded to cover all languages and all forms of material as rapidly as resources and technology allow. Retrospective conversion of any category of material should not take place until that category is being converted on a current basis.

2) An early goal of library automation efforts should be the conversion of some portion of retrospective records to machine form.

3) Standardization of bibliographic content and machine format is necessary for a national bibliographic data base; the standard for converting retrospective records should be the same as those for current records.

4) Highest priority for retrospective conversion should be given to records most likely to be useful to the largest number of libraries; subsequent priorities should also be determined by the same criterion.

5) Because decentralized conversion would be more costly and unlikely to meet the requirements for standardization, large-scale conversion should be undertaken as a centralized project under the direction of the Library of Congress.

One of the specific recommendations in the RECON feasibility report was that a pilot project be established to test various conversion techniques, ideally covering the highest priority material (English-language monograph records from 1960–68). In August 1969, a two-year pilot project was initiated with funds provided by the Council on Library Resources, the U.S. Office of Education, and the Library of Congress. The grants from the Council and the Office of Education also included funds for the RECON Advisory Committee and support for several research projects to be carried out by the RECON Working Task Force.

The advisory committee, whose role was that of a sounding board for the Library of Congress and the working task force, met twice during the pilot project. The committee members expressed their opinions on the work in progress, recommended changes in the emphasis or direction of the project, and reported on activities in their sphere of interest that had implications for RECON.

The present report is oriented toward the work of the project as a whole rather than toward individually funded activities. The pilot project conducted at the Library by LC staff members covered five major areas:

1) *Testing of techniques postulated in the* RECON *report in an operational environment by converting English-language monographs cataloged in 1968 and 1969 but not included in the* MARC *Distribution Service.*

This phase of the project partially satisfied the recommendation in the RECON feasibility report to the effect that the initial conversion effort be limited to English-language monograph records issued from 1960 to 1968. The work performed during this phase included the training of RECON editors and typists, selection of records for conversion from Card Division card stock, modification of records already in machine-readable form (MARC I and MARC II practice records) for inclusion in the RECON data base, comparison of records from card stock and from the machine-readable data files against the LC Official Catalog and updating of the records when necessary, inputting into the MARC system records that had been manually edited and records that had received no editing preparation but were keyed for processing by the format recognition programs, and analysis of production costs by function to determine cost per record. Production was handled by a new unit in the MARC Editorial Office, the RECON Production Unit.

2) *Development of procedures and computer programs to implement format recognition.*

The format recognition technique was described in an appendix to the RECON feasibility report.[2] Format recognition is a machine process that assigns content designators and fixed field codes to the bibliographic record by analyzing punctuation, keywords, data content, etc. Content designators are the tags, indicators, and subfield codes that identify data explicitly for machine manipulation. Fixed fields contain such elements as codes to indicate language, country of publication, type of publication, etc. The feasibility report, which was written before the first format recognition feasibility study was completed, concluded that "partial editing combined with format recognition processing is a promising alternative to full editing."[3] Shortly after publication of this report, emphasis was shifted to an approach using format recognition processing without previous editing. The preliminary results were promising and indicated that the conversion of catalog records could be expedited by reducing the amount of human intervention required. The pilot project concentrated on the research to develop these techniques, to implement procedures and programs for English-language records, and to expand format recognition to include records in other languages.

3) *Analysis of techniques for the conversion of*

older English-language materials and titles in foreign languages using the roman alphabet.

The RECON feasibility report had noted the additional complexity of converting foreign-language materials to machine-readable form.[1] Since the production effort was limited to the conversion of recent English-language monographs, a separate phase of the project was instituted to isolate and analyze problems associated with the conversion of records in other languages and cataloged according to other conventions or cataloging rules. The work performed in this phase included selection of a valid sample of titles that would also provide data for other LC projects, as well as the editing and typing of a sample of French and German monograph records for test purposes.

4) *Monitoring of the state-of-the-art of input devices that would facilitate conversion of a large data base.*

The RECON report considered several types of input devices in an analysis of the unit cost per record for various technical alternatives.[5] The present study included a determination of whether significant advances in equipment that would accommodate bibliographic data had been made since publication of the report. In this phase of the project, surveys were conducted of keyboard devices, two of which were tested in a production environment, direct-read optical character readers (OCR), two of which were tested on the vendors' sites, and cathode ray tube (CRT) terminals. The use of a mini-computer on-line for MARC input functions was also investigated.

5) *A study of microfilming techniques and their associated costs.*

The RECON report evaluated several files as candidates for a retrospective conversion effort.[6] The Library of Congress Card Division record set, used in conjunction with the LC Official Catalog, was selected as the best file to meet the criteria established by the working task force. Because the record set is a "high use" file which cannot be withdrawn in whole or in part for any substantial period of time, microfilming was suggested as the least disruptive method of securing records for conversion. This phase of the project postulated four alternative procedures, established microfilming requirements to test the specifications for each procedure, and prepared cost estimates for each alternative.

The accomplishments of the pilot project are discussed in detail in the sections that follow.

Notes

[1] RECON Working Task Force *Conversion of Retrospective Catalog Records to Machine-Readable Form; a Study of the Feasibility of a National Bibliographic Service* (Washington, Library of Congress, 1969). 230 p.
[2] Ibid., p. 169–179.
[3] Ibid., p. 179.
[4] Ibid., p. 79, 82.
[5] Ibid., p. 49–55.
[6] Ibid., p. 20–38.

CHAPTER 2

Summary and Conclusions

The major findings of the RECON Pilot Project are as follows:

1) Format recognition applied to unedited records has proved to be a practical computer technique, and the need for human editing of records before they are input has been eliminated. The costs of keying and proofing for format recognition remain essentially the same as those for the processing of fully edited records, but it appears that format recognition will permit a reduction of about 12 percent in the manpower cost of creating MARC/RECON records. An additional cost reduction will result from the fact that the machine time for format recognition processing is less than that for the format edit and content edit processing required for fully edited records.

2) The preferred device for original input of MARC/RECON records is the IBM Magnetic Tape Selectric Typewriter (MTST). No other device met the Library's keying requirements (easy accommodation of variable record lengths and the expanded character set) with a concurrent reduction in cost, either directly or through an increase in production capacity. Evaluation of CRT devices for on-line correction procedures led to the selection of the Irascope Model LTE as having the most desirable characteristics for the MARC/RECON operation. It was determined that mini-computers offer no gain either technically or economically for input of fully edited records in the LC environment. A reassessment of this finding may be in order, however, in the context of format recognition processing. Since the success of this technique depends on accurate typing, greater flexibility in correcting simple typing errors before processing would promote greater accuracy in machine editing. No direct-read OCR device was found that could perform adequately in converting LC cards to machine-readable form.

3) The most efficient means of producing source documents from the LC Card Division record set is to film all cards in a given series against a worksheet form to produce hard copy via Xerox Copyflo printout. The desired subset of records is then selected for conversion to machine-readable form.

4) Processing of older catalog records and those in foreign languages involves significantly more complex problems, and hence greater conversion costs, than those encountered in the processing of current English-language titles.

5) Many practical difficulties are associated with the conversion of retrospective catalog records on a large scale. The production rates of the pilot project were significantly lower than was anticipated in the RECON feasibility study. Although some of the problems were attributable to the experimental character of the project, there is abundant evidence that recruiting, training, and supervision of the staff in such an endeavor are formidable tasks.

6) The lowest RECON unit cost that can be anticipated is $3.06 for an unedited record processed by format recognition. Even if this rate were to remain constant over a long period, it would cost more than $900,000 to convert the estimated 300,000 English-language records issued in the 1960-67 period.

CHAPTER 3

RECON Production

Background

The RECON Working Task Force evaluated several strategies for conversion of retrospective catalog records. From the standpoint of completeness, accuracy, and quality, the Library of Congress Official Catalog was considered the most suitable file for use in retrospective conversion. Various problems, however, are encountered in using the master records from the Official Catalog as input for a conversion project. The Official Catalog contains over 12 million cards, including main and added entries, name authority records, series treatment cards, and other types of control records. Searching this file for all or part of the four million discrete catalog records produced by the Library of Congress since 1898 would be costly and time consuming. In addition, many of the master records are handwritten or have handwritten changes or additions and are thus very difficult to use in a conversion process.

For these reasons, it was decided that the actual catalog records should be obtained from the Card Division record set, which is a master file of all printed cards produced by the Library since 1898. The file is arranged by year (the first two digits in the LC card number) and then by the sequential numbers that follow. Since the record set is used heavily, the RECON Working Task Force recommended microfilming of the cards as the best means of providing source documents for retrospective conversion with minimal disruption of Card Division operations.

For the pilot project production efforts, it was considered more expedient to obtain the necessary records (in the 1968, 1969, and 7 series of card numbers) from card stock rather than by microfilming the record set. These cards were compared with the corresponding main entries in the LC Official Catalog for any changes or additions not reflected on the cards from stock. Although printed cards sold by the Library are not always as up to date as the records in the Official Catalog, such a limitation was considered undesirable for machine-readable records. The Library itself would be unwilling to accept machine records less accurate than those in the Official Catalog, and a national bibliographic store would also need records of the highest quality possible.

The original estimates of records to be converted, based on LC catalog statistics for 1968 and the first three months of 1969, were:

1969 and 7 series	22,000
1968 series	47,000
1968 and 1969 machine-readable records	16,000
TOTAL	85,000

The machine-readable records consisted of those converted during the MARC Pilot Project (MARC I) and those converted before the MARC Distribution Service was begun (MARC II practice records).

As the selection of records eligible for conversion progressed, it became obvious that the number of records to be converted had been overestimated. Many titles reported in the cataloging statistics for 1968 apparently were not completely processed until 1969 or later because of backlogs, and the cataloging output for the first three months of 1969 consisted primarily of titles with 1968 card numbers. Because of this [illegible] g, many more of these records in the 1968 [illegible] 1969 card series were

received for input as current records for the MARC Distribution Service. The number of records actually converted was

1969 and 7 series	8,541
1968 series	33,904
1968 and 1969 machine-readable records	15,518
TOTAL	57,963

Additional records to make up the deficit may be obtained through other sources.

Production

Records with 1969 and 7 series card numbers were manually edited at the Library and keyed by a service bureau. These records were distributed to 47 MARC subscribers early in 1971. There was no charge for the records or their duplication; instead, subscribers were requested to send tapes on which the records could be duplicated. Approximately 9,500 records in the 1968 card series were also manually edited at the Library and keyed by a service bureau. Another 6,000 were both edited and input at the Library. The remainder have been keyed for subsequent processing by the format recognition program.

Records already in machine-readable form but requiring modifications to make the content designators identical to those in the data base for the MARC Distribution Service have been converted by special programs. MARC I records in the 1968 card series have been converted to MARC II, proofread, compared with the Official Catalog, and updated. Two MARC practice tapes were processed by programs tailored to the modifications required for each tape. The modifications were necessary primarily because of changes made to the format subsequent to the time that these records were input. These records were also compared with the Official Catalog and updated.

Staff

Experience at the Library of Congress has demonstrated that staff members assigned the task of preparing catalog records for conversion to machine-readable form must be familiar with cataloging fundamentals. In assigning content designators or proofing, knowledge of the cataloging rules is necessary to make the correct decisions for machine identification of cataloging information. Because of the large number of new staff members involved in RECON production, it was decided that formal instruction would be more efficient than on-the-job training.

Classes were conducted in elements of cataloging, MARC editing procedures, and correction procedures. Additional sessions were held on LC subject headings and classification, LC filing rules, Dewey decimal classification, workflow through the MARC Editorial Office, and the MARC character set. Three series of classes were held during the period of the pilot project. Formal instruction lasted from 12½ to 19 days, depending on the size of the class and the aptitude of the pupils. After the initial training period was completed, the editors' work was reviewed for at least six months, and if their work was satisfactory, they were promoted to independent editor status.

Instruction was provided by staff members from the MARC Development Office and the MARC Editorial Office. Personnel from other divisions in the Processing Department were also invited to give briefings in their areas of specialization.

The staff of editors varied in size during the course of the pilot project, and the rate of turnover was high. Since all of the positions were temporary, it was sometimes difficult to find qualified individuals for the jobs. Eleven editor positions were originally established, but this number was reduced to nine with the creation of two verifier positions early in 1970. That number was reduced to eight by the end of the project.

Verifiers review records, with both the proofsheets and the input worksheets in hand, after the editors have completed the initial proofreading. Verifiers are required to have been independent editors for at least six months and to have met specific standards in the quality and quantity of their editing and proofreading. They spend a minimum of six months as trainees before becoming independent verifiers. Since promotion to the position of verifier is based on satisfactory performance in MARC editorial functions, no special verifier training program is needed. The two verifier positions were filled in January and May of 1970.

RECON typists were assigned to the Keyboarding Unit of the MARC Editorial Office,

and initial training involved typing of current MARC records. At the end of a six-month training period, those typists who met the quality and quantity standards were promoted to the position of independent typist. The RECON typing staff ranged in size from one to three persons during the pilot project.

Supervision of RECON editors and verifiers was the responsibility of the head of the RECON Production Unit. In October 1970, an additional supervisor was added to the staff, which also included a clerical assistant for Xeroxing. To maintain an even workload, close liaison was established among the different units within the MARC Editorial Office and with the research staff in the MARC Development Office.

As a result of two Government-wide salary increases during the course of the pilot project, funds from the Council on Library Resources grant for RECON production were expended by June 30, 1971, and the Library assumed the costs of completing the conversion of records in the 1968 card series. The RECON Production Unit of the MARC Editorial Office was dissolved, and some of its staff members were absorbed into current MARC operations, although they continued to work on conversion of the 1968 records.

Card Selection

The Card Division supplied the RECON Production Unit with printed cards representing each LC card number in the 1968, 1969, and 7 series. Cards were drawn from stock, beginning with the cards in the 1969 and 7 series. Gaps in the sequence of card numbers were searched by the Card Division staff in the record set. If the gap represented the number of a printed card that was not in stock, the card from the record set was reproduced. Form cards were inserted to indicate cards missing from the record set or cards that had not yet been printed.

Cards sent to the RECON Production Unit were then subjected to additional selection procedures to identify those records that were within the criteria established for the pilot project, i.e., English-language monographs. The determination of whether an item was in English was based on the text rather than the title page. An anthology of literature in Spanish with a title page in English, for example, was not included in RECON; a book with text in English but a title page in French was. Determination of the language of the text depended on the presence of specific information on the printed card. For a multilingual book (complete text in more than one language), the language of the first title determined its eligibility for RECON.

Atlases, which are classified below G3000, were included but not single maps or sets of maps, which are classified as G3000 or above. Music and music scores were excluded, but books about music were included. Other categories excluded were records for motion pictures, filmstrips, and other kinds of materials that were not considered books. Records representing serials were also excluded. Those labeled "MARC" in the lower right-hand corner of the printed card were excluded since they were already in the data base of the MARC Distribution Service.

The cards selected were kept in LC card number sequence and were then checked against a print index of card numbers for records in machine-readable form. This procedure was necessary because catalog records converted into machine-readable form before the beginning of the MARC Distribution Service in March 1969 did not have the special MARC notation on the printed cards. Since March 1969, the word "MARC" has been printed in the lower right-hand corner of the card for titles which are also available in machine-readable form. This notation ensures that revisions or changes on these cards will be forwarded to the MARC Editorial Office to update the MARC data base.

Each number listed in the print index was accompanied by a source code indicating the machine-readable data base in which the record resided. Five codes were used to designate the MARC I data base, first MARC II practice tape, second MARC II practice tape, MARC II data base, or MARC II residual data base.[1]

If the RECON editor found a match on the print index, the appropriate source code was added to the printed card, and the card was placed in a separate file. The remaining cards eligible for RECON input were reproduced on input worksheets. Cards not selected for RECON production were saved for possible future use.

Form cards representing cards not yet available from the Card Division were filed sepa-

rately. The Card Division supplied the missing cards as they became available, and the record selection process was then applied to these cards and eligible records reproduced on worksheets.

Contractor Input

RECON records from the entire 1969 and 7 series and a portion of the 1968 series were input by a service bureau. Because the input worksheets were to leave the Library, stringent controls were necessary. The location, in and out of the Library, of each record had to be known so that worksheets could be reconstituted in the event of any loss. At two-week intervals, the contractor picked up new edited worksheets and corrected proofsheets and returned the worksheets and the corrected proofsheets, from the previous cycle, together with a magnetic tape.

The contractor used IBM Selectric typewriters equipped with an optical character reader (OCR) typing mechanism. The hard-copy sheets prepared on this equipment were run through a Farrington Optical Scanner. The output from the scanner, in the form of a magnetic tape, was processed by the contractor's programs to produce a tape in the MARC pre-edit format.[2] This tape was delivered to the Library for processing through the rest of the MARC system.

In April 1970, a comparison was made of error rates in RECON records typed by the contractor and current records typed in the Library. In analyzing the results, it was found that the contractor's errors were generally more serious, e.g., omission of a field, omission of a record, or an incorrect tag. The general conclusion reached was that the overall accuracy of the two groups was about the same but that the contractor was handicapped by not being able to answer typists' questions or to give special instructions during keying. Since many of the contractor's errors occurred in the input of diacritical marks and special characters, the editors subsequently identified these characters by their hex code equivalents for ease of input.

Problems of RECON vs. Current Records

Because the RECON Pilot Project used printed cards as source documents, the editing process was subject to certain complications which are not associated with the processing of current records, for which the source document is a manuscript card.[3] It was expected that editing would be easier with a worksheet produced from a printed card rather than a manuscript card because the latter includes hand written data, instructions to the printer, etc. Experience, however, showed that Xeroxed printed cards were often difficult to read because of the confusion of such characters as e, o, c, a, and punctuation marks. If these were not clarified by an editor, legibility became a problem for the typist.

Inaccuracies on printed cards may be due to errors in either cataloging or printing. Since assignment of content designators can be made without ascertaining the correctness of the data in the field, errors may be overlooked during the editing process. Problems that arose in this connection were resolved by referral to the principal subject or descriptive cataloger. An analysis of records with cataloging/printing errors showed that 144 of approximately 20,000 records, (0.72 percent) contained 151 errors that required cataloging decisions. It is likely that the actual occurrence of such errors is somewhat higher, since some errors remain unidentified.

Differences in cataloging rules and procedures are critical problems in the conversion of older records and foreign-language records originating from shared cataloging copy. An analysis of these problems is presented in Chapter 6.

Since the book is not examined in the retrospective conversion process, difficulties arise in assigning certain fixed field codes from information on the printed card alone. In converting current catalog records to machine-readable form, many of these codes are assigned by descriptive or subject catalogers who have the book in hand. A RECON editor may encounter problems, for example, in ascertaining the proper language codes for a multilingual publication because of ambiguities in the title paragraph or in the notes. He may also have difficulty determining whether a particular title is a conference publication or a biography. It was concluded that editors and verifiers must devote greater attention to these problems than is required in the editing of current MARC records.

Catalog Comparison

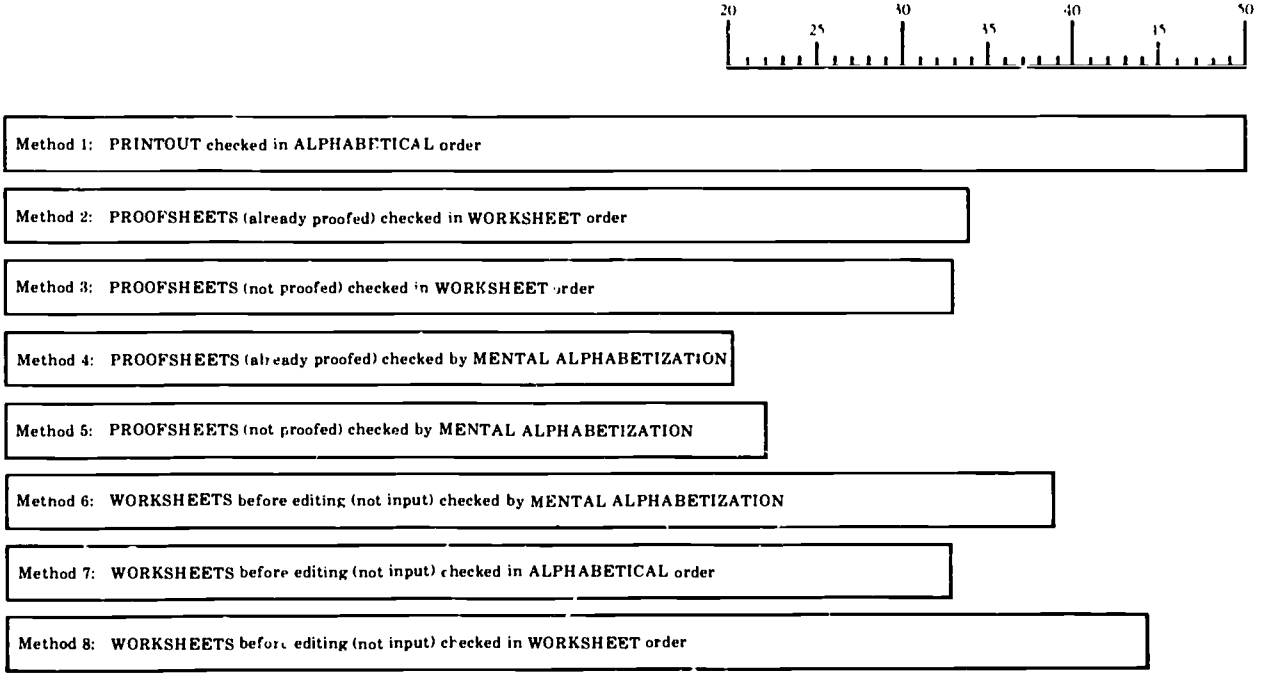
During the RECON Pilot Project, all records were compared against the Official Catalog. It had originally been thought that additional staff members would be hired for this task, but it became apparent that a shortage of qualified staff and the relatively short timespan of the project made such hiring impractical. Catalog comparison was instead assigned to the RECON editors, who already knew how to write in corrections and required only minimal additional training for the work.

Two RECON editors participated in an experiment to test eight possible methods of catalog comparison. The alternatives considered involved the following activities: 1) printouts of verified records arranged and checked in alphabetical order; 2) proofsheets (already proofed) arranged and checked in card number order; 3) proofsheets (not proofed) arranged and checked in card number order; 4) proofsheets (already proofed) arranged by card number but checked by mental alphabetization; 5) proofsheets (not proofed) arranged by card number but checked by mental alphabetization; 6) worksheets (before editing) arranged by card number but checked by mental alphabetization; 7) worksheets (before editing) arranged and checked in alphabetical order; or 8) worksheets (before editing) arranged and checked in card number order.

A group of 200 records was used for each of the proposed methods. For alternatives 2-8, the records were separated into batches of 20. The editors searched the Official Catalog, made the necessary corrections, and recorded the time spent as well as the number of changes made. Figure 3-1 shows the average number of records checked per hour using each of the proposed methods. Table 3-1 gives the estimated cost per record for five of the methods, based on the prevailing salary rates and other costs at the time the test was made.

The editors participating in the experiment found that the task of arranging worksheets in alphabetical order by main entry was time consuming and tedious. They also discovered that checking the Official Catalog with the records arranged in order by card number was not as difficult as anticipated because the entries tended to fall into a rough alphabetical order. Even mental alphabetization (in this context, searching the catalog alphabetically

Figure 3-1. Hourly rates for eight methods of cataloging comparison[1]



[1] *Taken from* Catalog Comparison: An Evaluation *(an internal document prepared for the MARC Development Office).*

Table 3-1. Adjusted cost figures for catalog comparison, by method

Method	Average number of records per hour	Unadjusted cost [1]	Additional costs [2]			Total adjusted cost
			Annual and sick leave	Supervision	Fringe benefits	
4	20	$.220	$.037	$.104	$.027	$.388
3 and 7	33	.132	.024	.063	.016	.235
8	44	.100	.017	.047	.012	.176
1	50	.087	.016	.041	.011	.155

[1] *Taken from* Catalog Comparison Evaluation *(an internal document).*
[2] *Based on assumptions used in the original* RECON *report.*

by main entry although the records in the batch were in order by card number) did not substantially increase searching time. They did find catalog comparison easier when using a worksheet, which consisted of a copy of the printed card, because it was easier to spot revisions. The printing format of the proofsheet, the possibility of the typists' omitting fields, and the fact that many of the diacritical marks and special characters are represented by different characters on the print train used to produce the proofsheets made the proofsheets unlike the printed card in appearance.

Although the results of the test indicated that method no. 1 was slightly faster than the other methods, it would require substantial modifications to the present MARC system if actually implemented. Additional sorting and printing would be necessary to produce hard copy if catalog comparison were performed after all records eligible for RECON had been processed and verified. Since many records would have to be corrected after comparison to reflect the changes found in the Official Catalog main entry, additional updating cycles would also be required. If catalog comparison were performed whenever a batch (approximately 4,000) of verified records were available, extensive system changes or the creation of multiple data bases would be necessary in addition to the sorting, printing, and updating cycles. Since these additional maintenance and processing routines would require more computer time, the total cost of method no. 1 would be higher than that depicted in Table 3-1. In addition, the editors found the printouts more difficult to use than the worksheets when doing catalog comparison. The decision was therefore made to implement method no. 8, under which unedited worksheets in order by card number were also checked in the same order.

Original plans for catalog comparison also included the writing of a new print program to produce a printout with records in two columns. One column would contain records in numeric order by LC card number, and the other column would consist of records by main entry. By cutting the printout in half vertically, the alphabetical sort could be used for catalog comparison, and the numerical sort could be used for proofing. Before coding for the two-up print program was begun, however, the catalog comparison experiment showed that searching the Official Catalog with a printout in alphabetical order did not substantially increase production. Since the editors preferred using the worksheets rather than the printouts, it was decided that the new print program would not be necessary.

Procedures for catalog comparison were worked out for the RECON Production Unit. During the comparison process, "MARC" was written on the main entry cards in the Official Catalog to ensure that corrections or revisions to the card are forwarded to the MARC Editorial Office. If the corresponding record was not found in the Official Catalog, a special cataloging certification code was added to the worksheet by the editor. An additional fixed field was included in the LC internal processing format to carry the catalog certification information. This field is not part of the MARC communications format for books. All worksheets were input regardless of whether the records had been certified in the Official Catalog; in the future, records that have not been certified in the Official Catalog can be obtained from the RECON master data base and checked again.

During the RECON Pilot Project, 7,528 records in the 1969 and 7 series and 34,628 records

Table 3-2. Data elements affected by changes in RECON *records*

Data element	1969 (403 records) Number	Percent	1968 (1989 records) Number	Percent
Total	409	100.0	2,312	100.0
Main entry	81	19.8	359	15.5
Body of entry	15	3.7	108	4.7
Collation	8	2.0	89	3.9
Series statement	5	1.2	67	2.9
Notes	31	7.6	281	12.2
Subject heading	17	4.2	186	8.0
Added entry	54	13.2	423	18.3
Classification number [1]	165	40.3	656	28.4
Dewey number	15	3.7	73	3.2
Added copy	9	2.2	31	1.3
Dash entry	2	.4	3	.1
NBN or SBN number	7	1.7	19	.8
Price	—	—	9	.4
Catalog card number	—	—	8	.3

[1] *The high incidence of changes in classification number primarily reflects the addition of the word "LAW" to many cards.*

in the 1968 series were compared against the Official Catalog, with the following results:

1) Three hundred and thirty-five 1969 main entry records (4.6 percent) and 1,671 1968 main entry records (4.8 percent) were not found in the catalog.

2) As a result of catalog comparison, changes were made in 339 1969 records (4.7 percent) and 2,149 1968 records (6.5 percent). An average of 1.1 changes per record were required.

3) Because almost as *many* records were not found (2,006) as were changed (2,488), a 105-card sample of missing records was studied. Analysis of this sample indicated that 45.7 percent of the records originally not found would have changes on them. Cards that were initially missing were located in the recheck for a variety of reasons. A card out notice had been replaced by the card itself in some cases. In other cases, an added entry pointing to the new main entry had been filed after the first catalog comparison.

4) If figures on the number of changes required are adjusted to take into account records originally not found, the percentages noted in paragraph 2 increase to 6.5 percent for 1969 and 8.4 percent for 1968.

5) The uniform filing title is usually not printed on LC cards; 6.7 percent of the 1969 records and 5.6 percent of the 1968 records required the addition of a uniform filing title. The degree of overlap between records with Official Catalog changes and records with added uniform filing titles was not determined. Table 3-2 shows the data elements in the catalog records that were affected by catalog comparison.

It should be noted that catalog comparison was performed on relatively current records during the pilot project. Additional difficulties would occur in comparisons involving records in foreign languages, handwritten records such as the master record in the Official Catalog, or older catalog records for which the printed card format and cataloging rules differed from present practices.

Notes

[1] In the MARC system, the residual data base contains records in the process of correction and verification. Once the records are declared free of errors, they are transferred to the master data base.

[2] The MARC input system consists of four major programs: pre-edit, format edit, content edit, and update. Tapes received from the contractor in a pre-edit output format could be input directly into the format edit program.

[3] The manuscript card is used at the Library of Congress to record cataloging information and as copy for the printing of catalog cards by the Government Printing Office.

CHAPTER 4

Format Recognition

Background

The preparation of bibliographic data in machine-readable form involves labeling each data element so that it can be identified by the computer. For this purpose, the MARC format employs tags, indicators, and subfield codes (or content designators). In the current MARC system, these content designators are supplied by the MARC editors before the data are typed on the MTST. The MTST tape cassette is converted to computer-compatible tape, which is then run through a series of computer programs to produce a proofsheet. In the proofing process, the editor compares the proofsheet against the original worksheet, checking for errors in editing or keying. Corrections are retyped and processed by the MARC system programs. A new proofsheet is produced by the computer and checked for errors. Records that are error free or "verified" are then removed from the work file and stored in a master file.

Since manual assignment of content designators and fixed field information by the editor is a detailed and somewhat tedious process, it seemed advantageous to develop a method whereby the computer would assign the content designators for bibliographic data by examining data strings for certain keywords, significant punctuation, and other clues. This technique, referred to as format recognition, was not entirely new at the Library of Congress. The need for such a computer program had been recognized during the planning stage of the MARC Distribution Service, but the pressure to implement the distribution service prevented more than minimal development of format recognition processing. The viability of such a technique has since been proved at other institutions, principally the Institute of Library Research at the University of California, Berkeley, and the Bodleian Library, Oxford.

The Library began its work on format recognition with a feasibility study conducted during the winter of 1968 69. At that time a certain amount of editing for MARC records was being performed by catalogers, and the study tested the possibility of using format recognition to assign content designators not already supplied by the catalogers. The study was divided into two parts, the first part analyzing those fields for which the cataloger supplied the tags and indicators but not the subfield codes and the second part analyzing those fields for which no tagging information was supplied by the catalogers. Detailed flowcharts of the algorithms were prepared, and a statistical analysis of the recurrence of types of fields was performed to provide the basis for an estimate of the effectiveness of such a technique. The results indicated that if a format recognition technique were used for partially pretagged data, roughly 85 out of 100 records would be processed without error. The results of this initial study were encouraging enough for the Library to proceed with the development of a format recognition project.

Logical Analysis

Implementation of the project was divided into several tasks. In the first task, the algorithms from the initial feasibility study were examined again to determine how successful they would be if there were no human editing. It was assumed that the typists would type directly from a printed catalog card or a manuscript card. The computer program

would take the raw data and supply the necessary content designators. The accuracy of setting fixed fields completely by computer was also studied. The results showed that, with accurate typing, records could be processed correctly approximately 70 percent of the time; that is, 70 out of 100 records would be correct, and the other 30 records would have errors in one or more fields. On the basis of these results, the decision was made to implement format recognition using unedited catalog records.

The second task covered several areas, including the development of input specifications for the typist. In general, these specifications provide for typing of the record from an untagged card on an input device with a typewriter keyboard. The information on the card is transcribed from left to right and from top to bottom. The data are input as fields, which can be detected by the program because each field ends with a carriage return and each field continuation is indicated by a carriage return. tab. Each field corresponds to a logical portion of the card; thus, the call number is input as a separate field, as are the main entry, collation, each note, each added entry, etc. The title paragraph is input as a single field with the title, edition, and imprint separated by delimiters.

Sixty keyword lists for English-language materials were compiled. The lists contain over 2,500 keywords covering such items as U.S. cities, foreign cities, geographical areas, words frequently appearing in corporate names or meetings, and honorary titles used with personal names.

Three possible processing sequences were considered:

1) Processing of independent fields before dependent fields, e.g., title field before title added entry field.

2) Processing of fields in order of their occurrence; use of a "look-ahead" technique to analyze as yet unencountered independent fields when necessary.

3) Processing of fields in order of their occurrence; use of a "look-back-and-fix-up" technique to modify previously encountered dependent fields when necessary.

The third approach was selected for the logical structure of the program.

The final product of task 2 was the documentation of logical specifications for the entire program,[1] a simplified description of which follows.

Gross identification of fields depends on the location of the collation, which is the only field that is easily identified and always present. Each of the first five fields is searched for the presence of "p." or "v." preceded by an arabic or roman numeral. If there is no hit, the field is searched for the presence of "cm." or "mm." Once the collation is found and identified, the fields preceding it can be identified. The call number is recognized as a character string beginning with one to three uppercase letters followed by one to four numbers. The remaining unidentified fields preceding the collation are identified as the main entry, uniform title, and title paragraph, depending on the number of such fields present. For example, if there are two unidentified fields preceding the collation, the first is tagged as the main entry and the second as the title paragraph. If there is only one unidentified field, it is tagged as the title paragraph.

The fields following the collation are examined separately. The characters at the beginning of each field are analyzed, and gross tag identifiers are assigned. For example, fields beginning with an arabic numeral (number-period-space) are identified as subject entries. Those beginning with roman numerals are tagged as added entries. If the first character is a quotation mark, the field is identified as a note. When the field begins with an open bracket, additional analysis is performed on the characters following the bracket. If the bracket is followed by an arabic numeral, the field is tagged as a subject entry. If the bracket is followed by a alphanumeric combination of characters, it may be an additional LC classification number. Using clues such as these, each field is assigned a tag or a partial tag, and a preliminary record directory is built.

The bulk of the program is devoted to the reexamination of each field to provide final and complete tags, indicators, and subfield codes. Each field is divided into groups and subgroups. A group is roughly defined as a data string ending with a significant period. A subgroup is a data string within the first group.

Examples: (Gr. = Group; Sg. = Subgroup)

1) Chicago. Art Institute.
Gr. 1 Gr. 2

2) Smith, John William, 1898–1965, comp.
Sg. 1 Sg. 2 Sg. 3 Sg. 4
Gr. 1

The first group and its subgroups are analyzed to provide the full tag and the first indicator. In the first example, "Chicago" is matched against the keyword lists, and when a match is found identifying it as a city, the field is identified as a corporate name entered under place. The second group is checked against the keyword list to see if it is a form subheading. Since there is no match, it is identified as a "b" subfield (administrative subunit) by default.

The second example consists of one group made up of multiple subgroups. The subgroups are scanned for a date range, whose presence and location are strong indications that the field is a personal name. The subgroup following the date subgroup is checked against the keyword list and is identified as a personal name relator. The second subgroup is checked against the honorary titles keyword list. Since there is no hit, the second subgroup is identified by default as a forename. Since the first subgroup contains only one word, it is considered a single surname. Thus, the field is finally identified as a personal name/single surname with date and relator subfields.

After each field is fully identified, the data are scanned for information needed for the fixed fields. For example, if the Dewey number field contains the notation "[Fic]" the intellectual level indicator is set to "j" (juvenile) and the fiction indicator to "x." If the words "autobiographical," "diary," "diaries," or "memoirs" appear in the title field, the biography indicator is set to "a." The date of publication field is derived from the date subfield in the imprint. The city transcribed in the place of publication subfield in the imprint is matched against the keyword lists to derive the country of publication code.

Simulation

Before the format recognition programs were coded, a manual simulation was conducted to test and, if necessary, improve the algorithms, keyword lists, and input specifications for the typists. Records for 150 English language monographs, generally consisting of "difficult" records, were selected for the simulation. The source data, in this case the manuscript cards, were input by the typists on the MTST according to the specifications designed for format recognition. The simulators added content designators to the MTST hard copy, following the format recognition logical design. The simulation records were processed by the current MARC system to produce printed proofsheets.

The simulation records were then checked by RECON editors, who were requested to keep track of the time required for proofreading format recognition records and to submit an informal report on their reactions to this work. Format recognition did appear to decrease the amount of time spent in the combined editing and proofing process, but it was demonstrated that success of the program would require extensive training of the input typists since accurate typing of records would be essential for the algorithms to work, as well as training of the editors to alert them to the kinds of errors the format recognition program might make.

Program Structure

The final tasks included the production of detailed flowcharts at the coding level and the actual coding and program testing. Coding began in July 1970, and final testing was completed in May 1971.

The format recognition program was written in Basic Assembler Language for the IBM 360/40 operating under DOS.[2] Input to format recognition consists of a tape created by a preedit program. On the input tape, each data field (call number, main entry, etc.) exists as a separate logical record followed by a field terminator code. The last data field in the catalog record is followed by a record terminator code. Logical records on the input tape are blocked into a physical record, which is about 2,500 bytes in length. Output from format recognition is a machine record in the LC MARC processing format.[3]

The format recognition program is modular in design to facilitate addition of new algoriths or changes in existing algorithms. The program consists of a mainline routine and five

principal routines. The mainline routine, which controls the overall processing, consists of opening of input, output, and keyword list files, calling for the five principal routines as required, and closing of the output file when an end-of-record condition is sensed in the input file. The principal routines are as follows:

1) Step 1 (set up record routine) builds the framework of the MARC processing format, initializes fixed fields, flags, and indicators, and builds those fields of the record which are not dependent on the content of the input data.

2) Step 2 (input and identify record routine) reads the logical records on the input tape created by the pre-edit program until it reaches the record terminator code signifying the end of the catalog record. Data that have been input as a cluster of MARC variable fields (e.g., the title paragraph, containing the title statement, edition, or imprint statement, and the collation paragraph, consisting of the collation, series, or price) are separated into individual fields. Preliminary identification of each data field is made, and a preliminary record directory entry consisting of the first two digits of the MARC tag, a sequence number for the tag, the starting character position of the field in the input buffer area, and the field length is built for each field identified.

3) Step 3 (process input field routine) sorts the record directory by tag and sequence number.[4] Each variable field is processed to: a) build the remainder of the field's entry in the record directory; b) derive the third digit of the tag; c) set variable field indicators which can be generated from analysis of the field being processed; d) delimit and assign subfield codes; and e) set any fixed field information which can be derived from the content of the variable field being processed.

4) Step 4 (complete variable field processing routine) completes the processing of the variable fields. The "look-back-and-fix-up" technique takes place in this step. For example, the geographic area code and the language codes are assigned on the basis of analysis performed in the preceding steps. In addition, all final text cleanup is performed, and the record built by the processing routines is moved and assembled in the output area.

5) Step 5 (output record routine) sorts the completed record directory again by tag and sequence number. Final adjustments are made in assembling the parts of the record, including final entries in the record's communications area. The record is then written on the output tape.

The mainline routine calls the principal routines in order (1-5 above) and repeats the process for as many catalog records as are on the input tape. Each routine exits to the mainline, which calls for the next routine until an end-of-file condition is found and the processing completed.

The program consists of six levels of routines, and the routines at each level can be called for execution by the routines at the next highest level in the hierarchy. The level concept is shown as follows:

Level	*Called by*
1. Mainline routine	——
2. Steps 1–5	Mainline routine
3. Substeps	Steps 1–5
4. Level 0 subroutines	Substeps
5. Level 1 subroutines	Level 0 subroutines
6. Level 2 subroutines	Level 1 subroutines

The communications buffer containing pertinent data that may be required by any of the routines (mainline, steps, substeps, and levels 0-2 subroutines) is the vehicle of communications across all routines. This buffer is assembled as an entity, and all routines in the format recognition program are linked with it as each routine is assembled. This allows all routines to be written as reentrant routines.[5] The current format recognition program is an off-line process; however, the Multiple Use MARC System (MUMS) may provide on-line processing and multiple terminal access to the format recognition program.[6] Existence of the communications buffer and the reentrant routines will facilitate the modifications required to incorporate format recognition in an on-line process.

The keyword lists used by the format recognition program are maintained as a separate data set on a 2314 disk pack but are stored in memory when the format recognition program is running. The total amount of core storage required for the format recognition program under DOS is approximately 120K,

80K for the program and 10K for the keyword lists.

Peripheral Programs

Two peripheral programs were written to support the format recognition project. Format recognition test data generation (FORTGEN) is an assembly language program which provides test data for format recognition by stripping MARC records of tags, delimiters, indicators, and subfield codes, and reformats the data to be identical with output from the pre-edit program. FORTGEN can process any given number of MARC records at any desired point in the MARC data base. Thus, a large quantity of high quality test data was provided without additional keying.

The keyword list maintenance program (KLMP) is an assembly language program which creates and maintains the 60 keyword lists used by the format recognition program in processing bibliographic records. The lists, associated tables, and control data are referred to collectively as "keyword list structures." The principal functions of KLMP are to read the entire set of keyword list structure from the file on disk, modify them as specified by parameter cards, and write a new file on disk. The individual functions performed by KLMP include the following: 1) create a list; 2) remove a list; 3) add a keyword; 4) delete a keyword; 5) augment a table by adding new codes to the translation tables to generate codes such as the geographic area code, language code, country of publication code; and 6) print an entire list or selected portions.

This program provides the flexibility required to change or update the keyword lists which are expected to be dynamic in nature. New lists will be added as format recognition is extended to include other languages, and keywords will be added to or deleted from existing lists as experience is gained in the use of format recognition. If the keyword lists were built into the format recognition program itself, it would be necessary to recatalog the program each time a keyword was changed.

Format Recognition Production

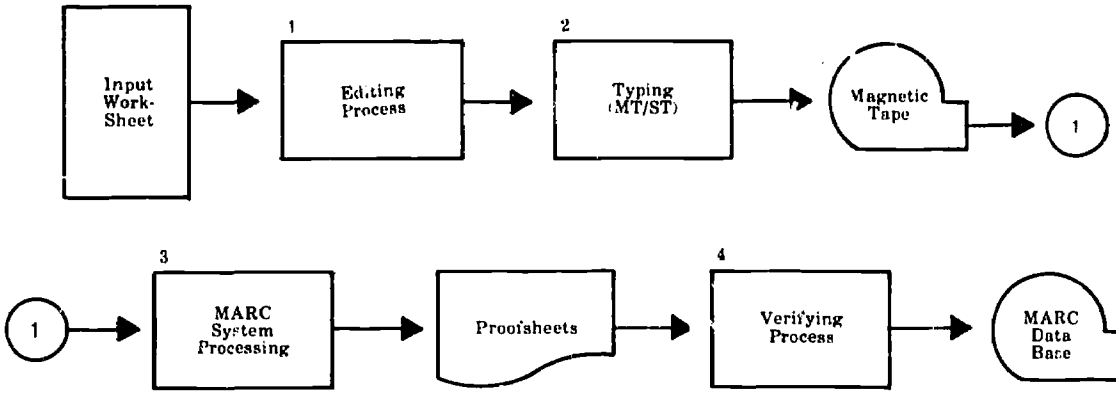
Approximately 17,000 RECON records in the 1968 card series have been processed by the format recognition program since actual production began in May 1971. RECON records rather than current MARC records were used to test format recognition because RECON records were not sent out at regular intervals by the MARC Distribution Service. Current MARC records have been processed by format recognition since January 1972.

The workflow for the manual editing process involves editing the records, keying them on the MTST, processing these records on the computer (including converting the MTST tape cassette to computer-compatible tape), proofing, and verifying. Format recognition eliminates the editing process, as shown in Figure 4-1.

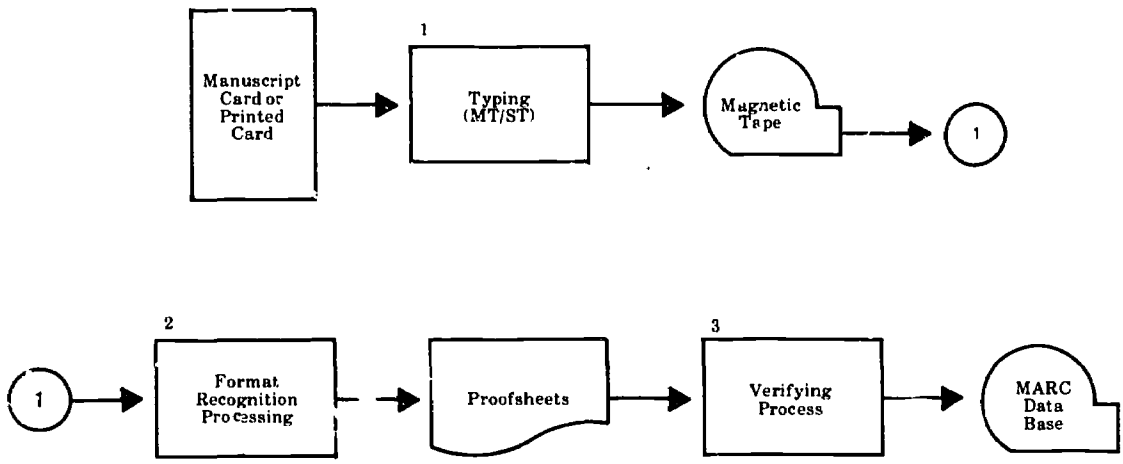
The time estimated in the RECON report for format recognition processing using unedited records as input was four seconds per record.[7] The actual machine time for production runs is ½ second per record, plus approximately ¼ second per record for the pre-edit program, or a total of ¾ second per record. This processing time compares favorably with that for the current MARC system (pre-edit, format edit, and content edit programs) which is approximately three seconds per record.[8]

Although a decrease in machine processing time of approximately 2¼ seconds per record, when projected over thousands of records, represents a significant gain, the principal hope for format recognition lies in relieving the human burden of editing and increasing overall production rates. Production and error rates for RECON editors were tallied for both conversion procedures: 1) editing records before input and proofing and 2) proofing format recognition records. Table 4-1 shows the results of this analysis.

The format recognition production rate of 8.4 records per hour (proofing only) represents a significant increase over the 4.6 records per hour for the combined editing and proofing process. Because proofing format recognition records is more difficult, the rate is less than that for proofing edited records (about 10 per hour). With format recognition records, the editors must be aware of the errors made by the program, which can be quite different from errors made by human editors, as well as keying mistakes. These rates were calculated over a relatively short period, and it is possible that the editors' production would rise as they gained more experience; however,



MANUAL EDITING WORKFLOW



FORMAT RECOGNITION WORKFLOW

Figure 4–1. Workflows for two methods of inputting MARC records

Table 4-1. Editing/proofing production and error rates for edited and unedited records

Type of record	Production rate (per hour)	Error rate [1] (per batch of 20 records)	Comments
Edited records	4.6	3.0	Production rate figure is RECON Unit average from February 1970–April 1971. Editing error rate figure is RECON Unit average from a 10 percent sampling of 6000 records edited from June–August 1970. Proofing error rate figure is RECON Unit average from June 1970–April 1971.
Editing only	9.2	3.0	
Proofing only	9.2	3.0	
Format recognition (Combines editing and proofing)	8.4	2.5	Production rate and error rate are RECON Unit averages from May 10–May 30, 1971.

[1] *The values in this column represent the average number of actual substantive errors per batch made by the editors and/or proofers. These errors include misassignment of tags, indicators, and subfield codes or failure to correct errors in the data content itself, e.g. misspelling of author names. In the case of format recognition, the error rate includes errors made by the format recognition program and not corrected by the proofers.*

it is also recognized that with a repetitive task like proofing, it is unlikely that production rates would continue to rise once a certain plateau has been reached.

Format Recognition Typing

Tests were also conducted to compare production rates and keying errors of input typists at the Library for edited records and for unedited format recognition worksheets. Table 4-2 shows the results of these tests. Of the 1,398 errors, 394 (28.3 percent) caused the format recognition program to misidentify data, i.e., to assign incorrect content designators. Typing speed, however, was slightly higher in keying format recognition records since there were no content designators to be typed.

Since there is a need for typing accuracy in format recognition and since it is possible that future requirements would necessitate the use of a contractor to support the LC staff, a typing test was also conducted by the contractor that had input RECON records. The contractor was asked to type 1,000 records for input to format recognition, with special emphasis on accuracy in typing. A goal of one error in 20 records was established. The first 900 records were considered practice records, and the final evaluation was based on the last 100 records. Since the contractor's proofsheets did not distinguish between uppercase and lowercase characters (i.e., a shift character was used to indicate an uppercase condition), the LC proofsheets produced after the records had been processed by format recognition were used to tally errors. Only errors made in typing were tallied; errors made by the format recognition program were ignored.

To attain an accuracy level of one error in 20 records, two proofreading cycles were required by the contractor before the records were submitted for input to format recognition. An accuracy level of one error per 7.7 records was attained with one proofreading cycle. Since experience has shown that errors are also made when correction records are

Table 4-2. Keying production and error rates for edited and unedited records

Type of Record	Production rate (per hour)	Number of records checked	Error rate Number records with errors	Percent records with errors
Edited records	12.9	300	178	59.3
Format recognition records	13.6	2,879	1,029	35.8

typed, the actual error rate in an operational situation would probably be slightly higher.

The typing was done by an expert typist who had a high school education and 18 months experience typing bibliographic records. Training time was minimal (one day). The typist performed the first proofreading. The second proofreading was done by an employee who had only three months' experience with bibliographic records but who had completed two years of college and one year with a professional secretarial concern. Although a longer training period was required (one week), she showed considerably greater ability in detecting errors. This limited experience does not offer conclusive evidence but does seem to indicate that ability beyond the range of an average typist is needed to detect the kinds of errors that occur in the typing of catalog records. It should be noted that the LC input typists do not proofread their records before computer processing because typing errors cannot be corrected very efficiently on the MTST after the record has been completely typed.

Assuming a production figure of 100,000 records, the contractor estimated that records could be delivered with an accuracy level of one error per 20 records at a cost of $0.85 per record. With an accuracy level of one error in 10 records, the cost would be $0.75 per record.

The format recognition typing test was also used to determine whether printed cards could be produced on a photocomposition device from unproofed format recognition records (assuming one typing error per 20 records). Since the kinds of errors created by the format recognition program would not generally affect a print program, it was thought that a large number of records could be converted to machine-readable form in this manner. Although the format recognition records would not be proofed, it was expected that the printed cards produced as a byproduct of this process would be checked against the original copy for omission of fields, data elements, etc. Although the results were satisfactory, this project was not implemented because of the many problems involved in maintaining these records in a separate data base and updating them for bibliographic content or MARC content designators. It did not appear advisable to convert a large number of catalog records to machine-readable form without providing full retrieval capability for these records at the Library of Congress as well as the potential for printing.

More experience is required to stabilize format recognition production rates for both proofing and typing. Since production rates have increased with no change in the number of personnel and machine processing time has actually decreased, it can be stated that conversion via format recognition is more economical than conversion using records completely edited prior to input. Monitoring of production rates and costs has continued beyond the termination date of the RECON Pilot Project.

Expansion of Format Recognition to Foreign Languages

The format recognition algorithms were formulated to process records for English-language monographs. Because of the commitment to investigate foreign-language material under the RECON Pilot Project and the planned expansion of the MARC Distribution Service to include records in other roman-alphabet languages, the Library was also interested in analyzing the requirements to expand format recognition to include the processing of records in foreign languages.

Although the computer programs have not yet been modified to handle foreign languages, analysis of the algorithms and the necessary modifications to the program specifications have been completed for French and German titles. German keyword lists have been created and converted to machine-readable form. This phase of format recognition will continue as an on-going effort of the MARC Development Office.

Results

As work progressed on format recognition, it became evident that the success of this project depended heavily on standard cataloging practices in recording data and in using punctuation. Format recognition was originally designed to accept cataloging data based on the *Anglo-American Cataloging Rules*, but modifications were necessary to accommodate the catalog records created by the Shared Cataloging Program at the Library, which uses entries from various national bibliographies and adapts them for LC printed cards.

Development of the International Standard Bibliographic Description (ISBD) has broad implications for format recognition and the creation of machine-readable data bases. As a result of the International Meeting of Cataloging Experts sponsored by the International Federation of Library Associations and held in Copenhagen in August 1969, a working party was appointed to prepare a draft proposal for an International Standard Bibliographic Description. The objective was to formulate specifications for bibliographic description, including a standard order of data elements, a minimum set of mandatory data elements, and standard punctuation. Use of the ISBD by national bibliographies and cataloging agencies would aid in the interpretation of cataloging data by humans and by format recognition programs. If all cataloging agencies were to prepare their entries according to the ISBD, format recognition algorithms could then be more easily expanded to encompass foreign-language catalog records produced by the Shared Cataloging Program as well as those originating at the Library.

Notes

[1] United States. Library of Congress. Information Systems Office. *Format Recognition Process for MARC Records; a Logical Design* (Chicago, Information Science and Automation Division, American Library Association, 1970). 1 v. (various pagings).

[2] A converted version has been written to operate under OS.

[3] Henriette Avram, and others. "MARC Program Research and Development: A Progress Report." *Journal of Library Automation*, 2:242-249 (Dec. 1969).

[4] A sequence number or site number is used to distinguish variable fields that have identical tags.

[5] "When processing a variety of input messages, one program may be 'WAITing'—for a file action, for example—and at this time another transaction wishes to use the program. This can cause problems if the program is written in such a way that it modifies itself while being executed, or stores logic information for later use in a location other than the unique message-reference block. Programs to be used by multiple transactions in this way must be carefully written so that no logic error can be caused by this. In particular, they must not modify themselves in such a way that, when control is taken away from them, another transaction can interfere with the modification. Programs which can be entered by multiple transactions without interference are referred to as reentrant programs. If a program is not reentrant, it may be necessary to have more than one copy of it in core at certain times in a multi-thread environment." From James Martin's *Design of Real-Time Computer Systems* (Englewood Cliffs, N.J., Prentice-Hall, 1967), p. 148.

[6] The Multiple Use MARC System (MUMS) is being designed by the MARC Development Office as a data utility. This system will be capable of processing machine-readable records regardless of the source of the record, the content of the record, and the master file on which the record will reside. It will also include all processing required to store, maintain, and retrieve records in both on-line and off-line modes. This project is still in the developmental stages.

[7] RECON Working Task Force. *Conversion of Retrospective Catalog Records to Machine-Readable Form; a Study of the Feasibility of a National Bibliographic Service* (Washington, Library of Congress, 1969), p. 64.

[8] Ibid., p. 63.

CHAPTER 5

RECON Costs

The RECON feasibility study projected costs per record for 20 possible technical alternatives for large-scale retrospective conversion.[1] The four most likely alternatives were: 1) direct-read optical character readers, format recognition processing; 2) no editing, keying using a magnetic tape inscriber, format recognition processing; 3) partial editing, keying using a magnetic tape inscriber, format recognition processing; and 4) full editing, keying using a magnetic tape inscriber.[2] The costs were calculated for the combined man/machine effort, which included staff salaries and overhead apportioned by function, i.e., project direction, editing, keying, proofing, catalog comparison, and quality control, as well as selection of cards from the record set or microfilming of the cards and production of hard copy; the input device; and computer processing of the records. Derivation of these cost figures is described in the feasibility study.[3]

In the pilot project, only the second and fourth alternatives were tested. The first alternative was impractical because no existing device affords the requisite OCR capability. The third alternative was unnecessary because format recognition processing of unedited records proved to be entirely satisfactory.

Computation of the actual cost per RECON record was complicated by two factors:

1) Since the RECON Production Unit was used as a test facility for new devices and techniques, normal production was often interrupted and, therefore, production rates were low.

2) Some RECON records were keyed by a contractor.

The variations in production rates during the lifetime of the project were such that a reliable cost per record could not be obtained by cost analysis based on total manpower costs and total input to the master data base. The use of contractual services resulted in unbalanced workloads, with peak periods of editing, preparation of source documents for the contractor, proofreading after records from the contractor had been processed through the MARC system, etc. This fluctuation in the workflow tended to create a bias in the cost calculations when the analysis included production figures for records processed entirely by LC staff as well as those edited and proofed by LC staff but keyed by the contractor.

It was decided, therefore, to use the average cost of a current MARC record as a basis for calculating a simulated RECON cost. A great deal of experience has been gained at the Library in the conversion of current catalog records to machine-readable form, and cost figures have been maintained since the beginning of the MARC Distribution Service. These costs have been stable for more than a year and thus may be considered valid. RECON production and MARC production are functionally identical with the exception of selection of records from the record set and catalog comparison. The costs of record selection and catalog comparison in the simulation of RECON costs were based on actual RECON Pilot Project experience.

Table 5-1 shows simulated manpower costs for the technical alternatives used in the pilot project. The simulated RECON cost for any given alternative is about 15 percent higher than the comparable MARC cost because of the need for record selection and catalog comparison. It should also be noted that the latter costs would tend to increase as the conditions of

Table 5–1. Simulated costs of converti . RECON records by two different methods [1]

Function	Full editing		Format recognition	
	Percent	Average	Percent	Average
Total	100.0	$3.46	100.0	$3.06
Record selection	6.1	.21	6.9	.21
Catalog comparison	5.5	.19	6.2	.19
Editing and revising	11.6	.40	—	—
Typing	9.2	.32	10.5	.32
Proofing	16.8	.58	18.9	.58
Verifying	17.0	.59	19.3	.59
Other duties	19.4	.67	21.9	.67
Leave and holidays	14.4	.50	16.3	.50

[1] *Derived from* MARC *conversion costs, July 1970–June 1971.*

retrospective conversion changed, that is, when smaller subsets of the total data base were selected or older records were processed.

Machine costs have been omitted from the table because they do not lend themselves to accurate proration per record. For example, the total cost of the input device per record is affected by the number of devices sharing the converter and the number of characters keyed. The RECON feasibility study prorated the converter over 20 keying devices; in the present study, 10 keying devices were used as a basis for calculation (see Chapter 7). The feasibility study also assumed 325 characters per record for unedited records and 500 characters per record for fully edited records.[4] Production experience showed an average of 398 characters per record for fully edited records.

On the basis of the above data, the cost per input device for fully edited records was determined to be $.07 per record, a figure which compares favorably with the prediction of $.063 in the feasibility study. Experience in the project was insufficient to permit an accurate evaluation of the projected cost for the input device of $.041 per record for typing unedited records, but an increase in the number of records keyed was noted in spite of the requirement for greater typing accuracy in format recognition.

Cost estimates for the hardware and software configuration required for a national bibliographic service remain valid since nothing has been found to contradict the assumptions made.[5] Present hardware costs, compared to those given in the feasibility study, could influence total costs, but there is nothing to be gained by updating one estimate with another at this time.

Costs obtained during the RECON Pilot Project cannot be compared on a one-to-one basis with projected figures in the feasibility study for several reasons:

1) The projected costs were based on the experience of the MARC Pilot Project and the MARC Distribution Service in the earliest days of its implementation.

2) Government salaries have been upgraded several times since the projected costs were calculated.

3) RECON production experience dictated a modification of the techniques postulated in the RECON feasibility study.

The study assumed procedures that involved keying, input processing, sorting of records, production of proofsheets, comparison of proofsheets with records arranged by main entry to records in the Official Catalog, and keying of corrections to records requiring changes. During actual production, it was found that the process of catalog comparison and updating of the record was facilitated by using the input worksheet (consisting of a copy of the printed card) rather than the computer-produced proofsheet. This procedure eliminated the necessity to sort the records by main entry before producing the proofsheets. Use of the source document for catalog comparison also allowed changes to be made to the record before the first keying, thus eliminating additional keying and proofing.

The feasibility study also assumed the existence of quality control procedures, consisting of an inspection of 50 percent of the total number of records after the first proofing. A first sampling of 10 percent of all converted records would result in acceptance of 55 percent of the batches. This 10 percent sample plus total inspection of 45 percent of the remaining 90 percent would provide 50.5 percent inspection.[1] Since production rates had not reached the proportions assumed in the feasibility study, the population was not large enough to have confidence in the proposed sampling technique. The majority of RECON records were checked by a verifier in addition to the initial proofing, and although similar inspection procedures were tested toward the end of the pilot project, these were initiated as part of overall quality control for the MARC Editorial Office and were not reflected in RECON costs.

1) The RECON feasibility report predicted that format recognition would add an incremental cost to the total cost of conversion; however, format recognition resulted in the savings of 2¼ seconds of machine time per record.

5) The RECON feasibility study suggested selecting the records for conversion (e.g., all English-language records from 1960 to date) from the Card Division record set, microfilming the records, and reconstituting the record set.

In Chapter 8 of this report, it is recommended that the part of the record set containing the subset of records chosen for conversion be microfilmed (e.g., all records, English-language and others, from 1960 to date) and the records for English-language monographs be selected from the microfilm copy. Since records for the RECON Pilot Project were limited to English-language records cataloged in 1968 or 1969, they were selected from cards in stock rather than from the record set. Since stock cards did not have to be replaced, microfilming was not necessary during the pilot project.

Notes

[1] RECON Working Task Force. *Conversion of Retrospective Catalog Records to Machine-Readable Form; a Study of the Feasibility of a National Bibliographic Service* (Washington, Library of Congress, 1969), p. 224–226.
[2] *Ibid.*, p. 98–99.
[3] *Ibid.*, p. 39–96.
[4] *Ibid.*, p. 59.
[5] *Ibid.*, p. 68–73; 183–223.
[6] *Ibid.*, p. 83–85.

CHAPTER 6

Research Titles Study

Background

Since the production operations of the RECON Pilot Project were limited to English-language monograph records with 1968, 1969, or 7 series card numbers, it was recognized that many problems in converting retrospective records would not be revealed except by a research effort. For this reason, a project was undertaken to identify and analyze 5,000 research titles, consisting of records for older English-language monographs and foreign-language monographs in roman alphabets. These records would be studied for problems concerning: 1) earlier cataloging rules which caused certain elements to be omitted from the record or transcribed in a different style; 2) printed card formats which placed elements in different locations; 3) elements in languages unfamiliar to the editor, such as foreign place names; 4) cataloging originating in different countries under the Shared Cataloging Program; and 5) expansion of format recognition to cover these kinds of records.

Two sources of research records were initially considered: 1) the project to compile a book catalog of the Main Reading Room reference collection; and 2) the popular titles of the Card Division. Both sources were studied for the degree of overlap of titles and suitability for RECON purposes.

The characteristics of the Main Reading Room reference collection were studied first. To compile the book catalog of this collection, printed cards were obtained from Card Division stock for conversion to MARC. The cards represent a wide range of material cataloged from 1899 to the present. Approximately one-fourth to one-third of the estimated 14,000 titles are serials. Most of the roman-alphabet languages currently processed at the Library are included, as well as the more common non-roman-alphabet languages such as Russian, Japanese, and Hebrew. The catalog contains a number of "difficult" titles, such as encyclopedias or dictionaries, which present a variety of cataloging and conversion problems.

The popular titles from the Card Division were studied next. As part of Phase I of the Card Division Mechanization Project, a record was kept of the number of orders received for each LC printed card. A printout which contained 39,148 card numbers for titles with 10 or more orders was produced. A sampling technique was developed to determine the percentage of overlap between this list and the titles in the Main Reading Room reference collection.

The estimated number of matches indicated that there was not enough overlap between the Main Reading Room catalog and the Card Division popular titles to consider a selection of titles that would serve the needs of these projects as well as those of RECON. It was decided that records from the Main Reading Room catalog would be more suitable as the first source from which to choose RECON research titles.

Selection

To obtain 5,000 records for the research titles study, approximately 1,800 cards were selected from the Main Reading Room catalog. The remaining 3,200 records, consisting of current foreign-language cataloging, were selected from printed cards drawn from Card Division stock for RECON production efforts. Emphasis was placed on foreign-language records in French and German, since titles in these two languages constitute a large propor-

tion of the Library's foreign-language cataloging. Other roman-alphabet languages were also represented.

An analysis of the problems that would be encountered in converting the research titles showed some similarity between the cataloging of older records (pre-1949) and of current foreign-language records based on shared cataloging copy. Certain stylistic conventions, e.g., the use of ellipses or the transcription of imprint statements, were similar for both kinds of material. It was felt that a thorough knowledge of the 1908 ALA *Cataloging Rules* would be necessary in order to interpret correctly the data on the older printed cards during a conversion project.

Editors in the RECON Production Unit have found that assignment of content designators for retrospective records, even those cataloged during 1968 or 1969, require a considerable amount of interpretation. For pre-1949 records, the problem becomes more acute because the editors must attempt to interpolate the procedures and techniques for current material to older records. It is likely that higher level personnel would be required to process these records since in some instances the changes would be similar to recataloging the entire record.

Different cataloging rules and printing conventions created even more serious problems for the expansion of format recognition to cover older catalog records and records based on shared cataloging copy. Each national bibliography, from which shared cataloging copy is derived, has its own rules and style of cataloging. For works in German, for example, punctuation, style of cataloging, and printing conventions may vary among entries from West German, East German, Austrian, and Swiss bibliographies, all of which may also differ from LC practices. The analysis also provided the basis for expansion of format recognition to include foreign languages (see Chapter 4).

Foreign Language Test

Decisions were subsequently reached on how to handle problems encountered in the analysis of the research titles (see Appendix I). These

Table 6–1. Production and error rates in the foreign-language editing test

Editor	Language tested	Total records edited	Total editing time	Av. no. edited per hr.	Number without errors	Total errors	Av. no. errors per rec.	Av. no. errors /batch
Editor no. 1	French	181	3.5 days	6.5	97 or 53%	114	.62	12.6
Editor no. 2	French	198	2.5 days	10.4	90 or 45%	175	.88	17.5
Editor no. 3	French	199	3.25 days	7.7	104 or 52%	142	.71	14.2
Editor no. 1	German	185	4 days	5.8	75 or 40%	185	1.0	20.6
Editor no. 2	German	199	2.5 days	9.9	111 or 56%	130	.65	13.0
Editor no. 3	German	200	4 days	6.2	104 or 52%	146	.73	14.6
			Editing statistics—English-language records					
Editor no. 1 [1]	English	383		Estim. 10 per hr.		56		3.0
Editor no. 2 [2]	English	119		11.3 per hr.	104 or 87%	17		2.9
Editor no. 3 [3]	English	139		7.3 per hr.	130 or 94%	11		1.6

[1] *Editor no. 1—No firm figures available. (Daily editing statistics also included proofing and catalog comparison).*
[2] Editor no. 2—Figure taken from daily statistics averaged over a six month period.
[3] Editor no. 3—Figure taken from daily statistics averaged over a six month period.

decisions were incorporated into editing instructions for the MARC Editorial Office as well as for a foreign-language editing experiment. The purpose of this experiment was to determine if experienced editors without much knowledge of foreign languages could maintain acceptable levels of accuracy and rates of production when editing foreign-language records.

Three trained editors edited a total of 1,180 French- and German-monograph records, or approximately 196 records per editor in each language. Two of the editors had taken some college level courses in a foreign language.

The test results revealed an error rate of approximately 50 percent, i.e., about 50 percent of the records contained editing errors. Production rates for the three editors and a comparison with their performance on English-language records are given in Table 6–1. The number of error-free records varied from 45 to 53 percent for French and 40 to 56 percent for German, as compared with 87 to 94 percent for English.

The editors made an average of 12.6 to 17.5 errors per batch of 20 records for French and 13.0 to 20.6 for German, compared with 1.6 to 3.0 errors for English. Since a standard of 2.5 errors per batch has been established by the MARC Editorial Office for trained editors, considerable improvement must be made before foreign-language records are converted on a production basis.

The high error rate for editing of foreign-language records was anticipated since it was known that approximately half of the editing effort is directly dependent upon an ability to read the words that make up the record. The remaining activities involve the identification of various data elements by their location on the printed card and are unrelated to language proficiency.

The majority of the errors took the form of wrong subfield codes, erroneous placement of delimiters, and incorrect fixed field codes (see Table 6–2). Approximately one-third of the errors in subfield codes and delimiters appeared in the title field, where knowledge of the language is essential in order to identify the data elements correctly. The majority of the tagging errors could have been avoided by consulting the name authority records in the Official Catalog.

Although all of the editors had had some training in French and none in German, their editing speed for French was only slightly higher than that for German. Since the editors began the experiment by editing French records and had thus gained additional experience before working with German, it was decided to determine the effects of such experience on the results. The number of errors made by each editor in each batch of records was tallied to see if any improvement had taken place during the entire course of the test. No appreciable improvement was noted for any of the editors. It is doubtful that much improvement would be shown unless extensive training in

Table 6–2. Location and number of errors in foreign-language editing test

Editor	Language	Tag	Indicator	Subfield code	Sequence number	Delimiter	Language	Fixed field	GAC	Other diacritics, punct., etc.
Editor no. 1	French	14	19	16 *1	3	14 *2	3	18	3	24
Editor no. 2	French	7	5	60 *23	2	11 *2	15	17	9	49
Editor no. 3	French	20	8	43 *27	7	13 *6	6	21	8	16
Editor no. 1	German	10	16	37 *12	11	25 *8	19	45	1	41
Editor no. 2	German	11	11	38 *22	0	14 *9	8	23	6	19
Editor no. 3	German	16	16	24 *7	1	18 *8	11	34	5	21

*Error associated with title field.

the editing of foreign-language records were conducted.

The test demonstrated that editors who are not fully knowledgeable in a foreign language cannot accurately edit records in that language without assistance. The editor's work must be carefully revised by someone with a reading knowledge of the language as well as an understanding of editing procedures. If the editor's work requires complete revision, actual editing time is of course drastically increased. During the test, it took the reviser almost twice as much time to correct the test records as it had taken the editors to edit them. Having language specialists edit such critical portions of the record as the title field and fixed fields would involve teaching them the editing procedures, and a staff of regular editors would still have to be maintained to edit the remaining portions of the foreign-language records. It was concluded that the more desirable alternative would be to have editors who are proficient in the language of the records to be edited. Even with the advent of format recognition processing for foreign-language records, the editors would still have to determine if the elements had been correctly identified by the computer.

CHAPTER 7

Input Devices

Since July 1969, the Library of Congress has used the IBM Magnetic Tape Selectric Typewriter (MTST) as the input device for MARC data. The current MARC system is an off-line system. Experience at the Library has indicated that original input of bibliographic data does not call for an on-line system but that correction and verification procedures would be greatly enhanced by on-line capability. Keeping such requirements in mind and seeking a best method for conversion of a large retrospective file, a state-of-the-art review of input devices was therefore conducted to accomplish the following: 1) compare new devices with the MTST and evaluate their relative efficiency for use in the LC environment; 2) determine if the development of direct-read optical character readers had progressed to the point that such equipment could be used to scan LC printed cards; 3) select a terminal device that would meet LC requirements for on-line correction and verification procedures; and 4) compare the use of a mini-computer with the present method of input (off-line to System 360) to determine if there were any technical or cost advantages to be gained.

Since the input of data is still the slowest component of a computer system, and because there is a growing demand for larger character sets, a great deal of emphasis has been given by hardware manufacturers in the past few years to the development of more efficient and sophisticated devices. Naturally, any study on a subject as dynamic as input devices is out of date and incomplete at any point in time. Although the investigation continued throughout the life of the RECON Pilot Project, it is recognized that an ongoing study is necessary and that devices may exist that are not described in this report because 100 percent coverage was not possible.

The investigation included an in-depth literature search, inquiries to various manufacturers, attendance at meetings, and testing of selected equipment, in some cases in an operational mode by the Keyboarding Unit of the MARC Editorial Office.

Keyboard Devices

In order for a device to compete with the MTST in the context of RECON evaluation, it had to meet the Library's keying requirements (easy accommodation of variable record lengths and the expanded character set) and either cost less than the MTST or increase production substantially to offset any increase in price. The equipment monitored, which included both off-line and on-line devices, has been categorized for this report as follows:

1) Key to magnetic tape
2) Key to computer-compatible magnetic tape
3) Key to disk
4) Key to cassette

Key-to-magnetic-tape systems consist of a number of input devices under centralized electronic control that acts as a routing and recording device. The control component may have the sophistication of a mini-computer with the facility to perform many functions such as editing, formatting, etc. In either case, one characteristic of this system is its ability to handle a large number of input devices simultaneously. The devices categorized as key-to-computer-compatible-magnetic-tape systems may either stand alone or share a centralized control device, called a "pooler," which records the information from a number of input de-

vices onto one magnetic tape. A characteristic of the pooler is that it handles fewer input devices simultaneously than the key-to-magnetic-tape system. The key-to-disk system operates in the same way as the key-to-magnetic-tape system. Devices in the key-to-cassette category require a converter to go from cassette to computer-compatible magnetic tape.

Table 7-1, compiled in May 1970, summarizes the characteristics of the devices monitored. Although prices were considered in the actual analysis, they have been omitted from the table to avoid the confusion that might be caused by out-of-date information.

The majority of devices available today do not satisfy the requirements for input of bibliographic data, the principal limitation being in the number of available characters. Among those evaluated, the Keymatic Magnetic Tape Unit appeared to offer enough potential advantage, despite higher cost, to warrant further exploration.

Keymatic Data System Model 1093

The primary attraction of the Keymatic is its ability to encode 256 unique characters without the use of an escape code. The layout of the keyboard is designed according to the user's specifications. The MARC character set, consisting of 175 graphic characters, could be assigned keys in clusters. One cluster might include special characters and diacritical marks, for example, and another cluster might contain uppercase and lowercase alphabetic characters. Common "words" such as MARC tags could be assigned to single keys (called expandables) and translated to their proper value by software, thus reducing the amount of stroking required.

In addition to the flexibility provided by the 256 characters on the keyboard, the machine offers the following advantages: 1) data are recorded directly on computer-compatible magnetic tape; 2) correction procedures are built into the device, i.e., the ability to delete a character, word, sentence, or entire record; and 3) 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 speed.

The keyboard of the machine tested was designed specifically for the Library's requirements. Four separate keyboards contained 184 keys of which 103 had uppercase and lowercase capability and 81 had only a single case. Although 287 codes could be represented, only 256 were used, with some keys representing the same codes. The codes were divided into the following categories: 1) 94 were used as expandables and assigned to those MARC tags and correction and modification commands that are used most frequently; 2) 10 were used as machine function codes; 3) 150 were assigned unique values in the MARC character set; and 4) two were left unused.

The keys on the four keyboards were assigned values so that the most frequently used keys were located in a strong stroke area. To keep additional training of the typists to a minimum, the main character keyboard was designed to correspond closely to that of the MTST. Practice was required only for the expandable keys and some of the less frequently used special characters. The keyboard layouts are shown in figures 7-1, 7-2, 7-3, and 7-4. The program supplied by the manufacturer was modified for code conversion and output format acceptable to the MARC system.

The two typists selected to participate in the test were both experienced MARC production typists. Each typist was given individual instruction on the machine and spent approximately seven days over a three-week period practicing. During the actual test period, the typists spent two weeks working full time on the machine. Their production rates increased from 6-7 records per hour at the beginning of the practice period to 11-12 records per hour at the end of the test period.

Each typist reported on problems that arose during the evaluation. One complication was the hesitation when the typist had to decide 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 large number of tags and their different combinations 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 the device might eliminate this problem.

The absence of hard copy, although consid-

Table 7-1. Characteristics of keyboard devices, May 1970

Manufacturer	Machine Type	Model	Keyboard	Display	Record Length in Characters	Remarks
Cybercom	K/C	MARK I	KP	None	80	Converter—Additional cost
Data Action	K/C	150	KP	Projection	720	Converter—Additional cost
IBM	K/C	50	KP	Backlight	720	Converter—Additional cost
IBM	K/C	MTST V	T	Printed	Infinite	Converter—Additional cost
		IV	T	Printed	Infinite	
Sycor	K/C	301	T	CRT	216	Converter—Additional cost
Tycore	K/C	8500	KP	Light-Emitting Diodes	240	Converter—Additional cost
Viatron	K/C	21	T/KP	CRT	Infinite	Many options affecting price
Burroughs	K/M	N-7000	KP	Projection	160	
Honeywell	K/M	Keytape	T/KP	Backlight	80–400	Pooler for 2 stations—Additional cost
Keymatic	K/M	1091	T	Backlight	Infinite	Price is for basic 88 keys. 256 unique keys available as option.
MAI	K/M	100-02	KP	Projection	100 or 200	Pooler for up to 8 station—Additional cost
Mohawk	K/M	6400	KP	Backlight	80	Pooler for 3 stations—Additional cost
Motorola	K/M	KB800	KP	None	200	Pooler for 7 stations—Additional cost
Potter	K/M	KDR	KP	BCD (Bit)	160	Pooler for 3 stations—Additional cost
Sangamo	K/M	DS9100	KP	Backlight	120	Pooler for 10 stations—Additional cost
Vanguard	K/M	Datascribe	KP	None	200	
Computer Consoles	K/T	Info System	T	CRT	960	One to 12 stations
Computer Entry System	K/T	6000	KP	None	496	Two to 6 stations
Mohawk	K/T	9000	KP	Backlight	80	Four to 16 stations
Computer Machinery	K/T	Key Processing	KP	Backlight	250	Eight to 32 stations
General Computer Systems	K/T	2100	T	Printed	200	Seven to 39 stations
Inforex	K/T	Key Entry	KP	CRT	128	Four to 8 stations
Penta Asscciates	K/T	Key Logic	KP	Backlight	200	Eight to 64 stations
Systems Eng.	K/T	Keytran	KP	None	300	Nine to 48 stations
Logic Corp.	K/D	LC-720	KP	CRT	350	Four to 16 stations

LEGEND
K/T = Key-to-magnetic-tape system
K/D = Key-to-disk system
K/C = Key-to-cassette system
K/M = Key-to-computer-compatible-magnetic-tape system
KP = Keypunch
T/KP = Typewriter or keypunch
T = Typewriter
Blacklight = a matrix consisting of all individual characters that can be keyed. Each character, as keyed, is displayed one at a time in its particular position in the matrix.
Projection and light emitting diodes = A one-character position dot matrix. Each character, as keyed, is displayed one at a time in the same position.
BCD = Lights displaying the bit position (on, off) of individual characters. Each character, as keyed, is displayed one at a time.

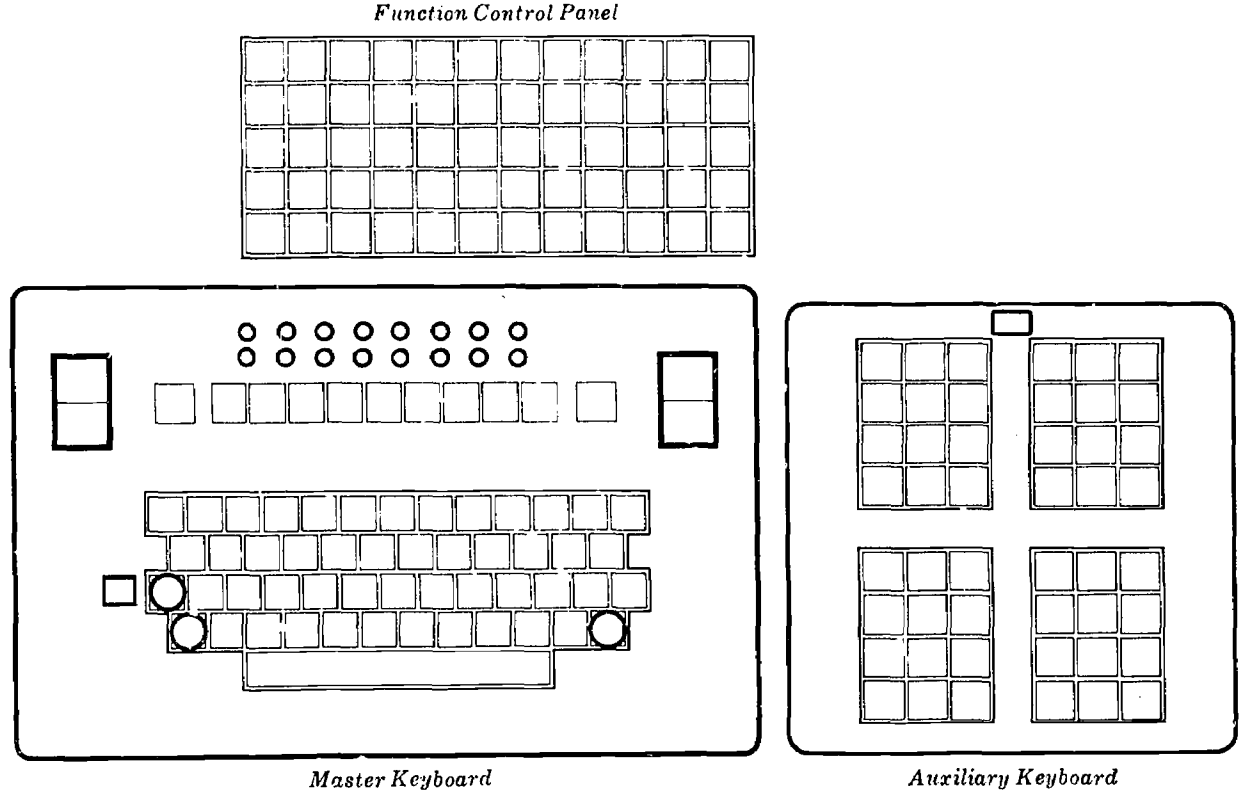
ered beneficial to typing speed, proved to be a handicap for this test. Under current input procedures when a typist thinks that she has made a typing error, she checks the hard copy to verify that a mistake has actually been made before taking corrective action. The absence of hard copy precluded such verification, and the typists reported that this detracted from their efficiency.

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 so that a less expensive model could be used. Keymatic does have a 24-month lease plan under which the basic machine could be rented for $368.00 per month. This would be an increase of $258.00 per month per machine over the cost of the current method of input.

Average production rates were computed for the same two typists for the Keymatic and the MTST. Although the same records were not actually typed on the MTST, experience with production and error rates on that device has been extensive so that it was considered valid to use existing MTST figures for the comparison.

In computing the cost per record, the hourly

Figure 7–1. Keymatic keyboard layouts.



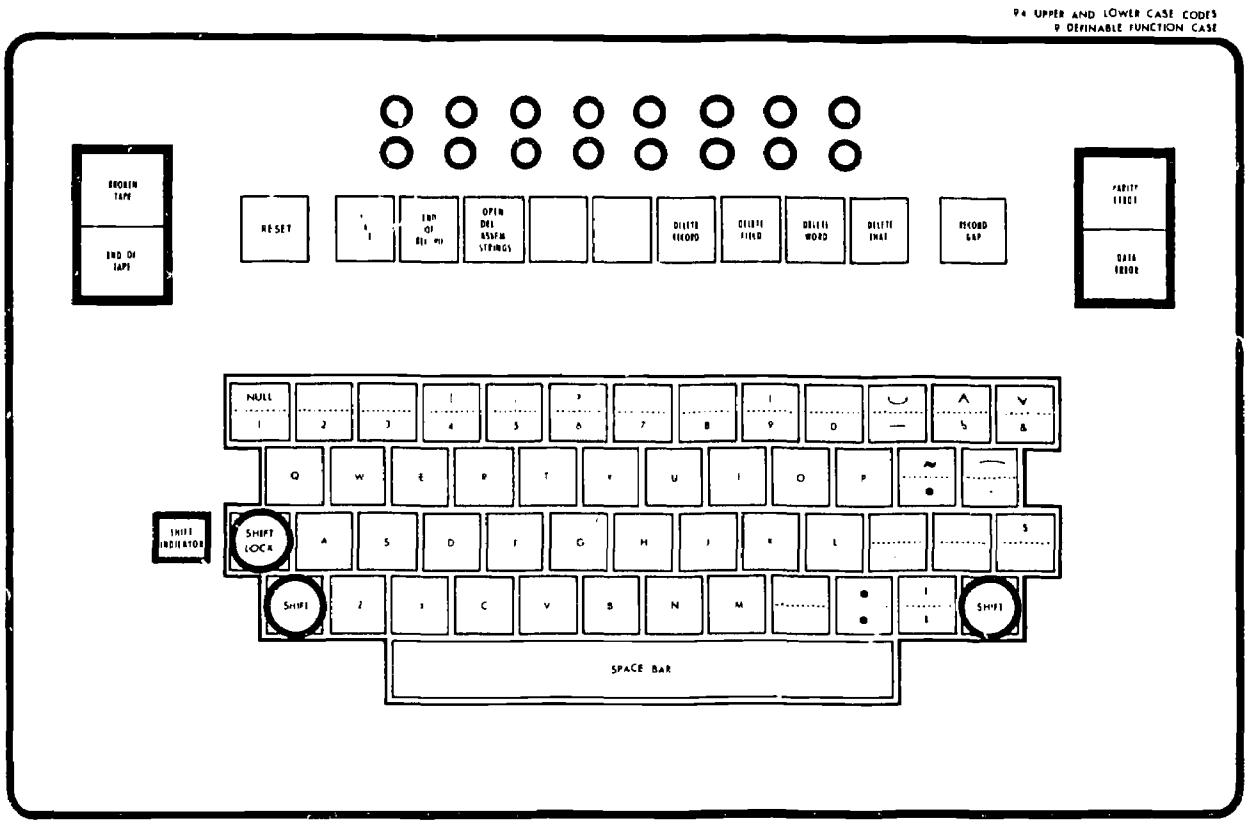


Figure 7–2. Keymatic master keyboard.

cost per machine was calculated by dividing the cost per machine by 132 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 $2.79. The MTST rental cost is $110.00 per month, resulting in an hourly cost of $.83. A converter is required to translate MTST output to computer-compatible tape, adding an incremental cost to each input device. The monthly rental cost of the converter is $260.00. For this report, the total number of MTST devices producing input for the converter (the Library has 10 MTST devices, including the six used for MARC/RECON) was used as a base figure. Addition of the prorated converter cost of $.20 per hour to the MTST cost of $.83 resulted in a total hourly cost of $1.03 for the MTST. On the basis of 12.1 records per hour typed on the Keymatic and 14.6 records per hour on the MTST, the cost per record is $.23 for the Keymatic and $.07 for the MTST.

Since the test indicated that the Keymatic, used in the LC environment, did not increase production rates, no savings in cost were demonstrated. The complexity of the data to be typed and the construction and quality of the worksheet used at the Library impose severe constraints on all machines. In order to make a fair comparison between the Keymatic and the MTST, the manuscript card was used for the test rather than the printed card. Reproduction of the manuscript card on the MARC/RECON worksheet results in a source document that is difficult to work with, owing to the loss of legibility during the copying process, the position of tags in relation to content, and the combination of typed and handwritten data inserted by the catalogers.

Keymatic does have a machine, the Model K-103, which has an 80-character visual display option which might correct one of the objections raised by the typists, i.e., the absence of hard copy. However, the other prob-

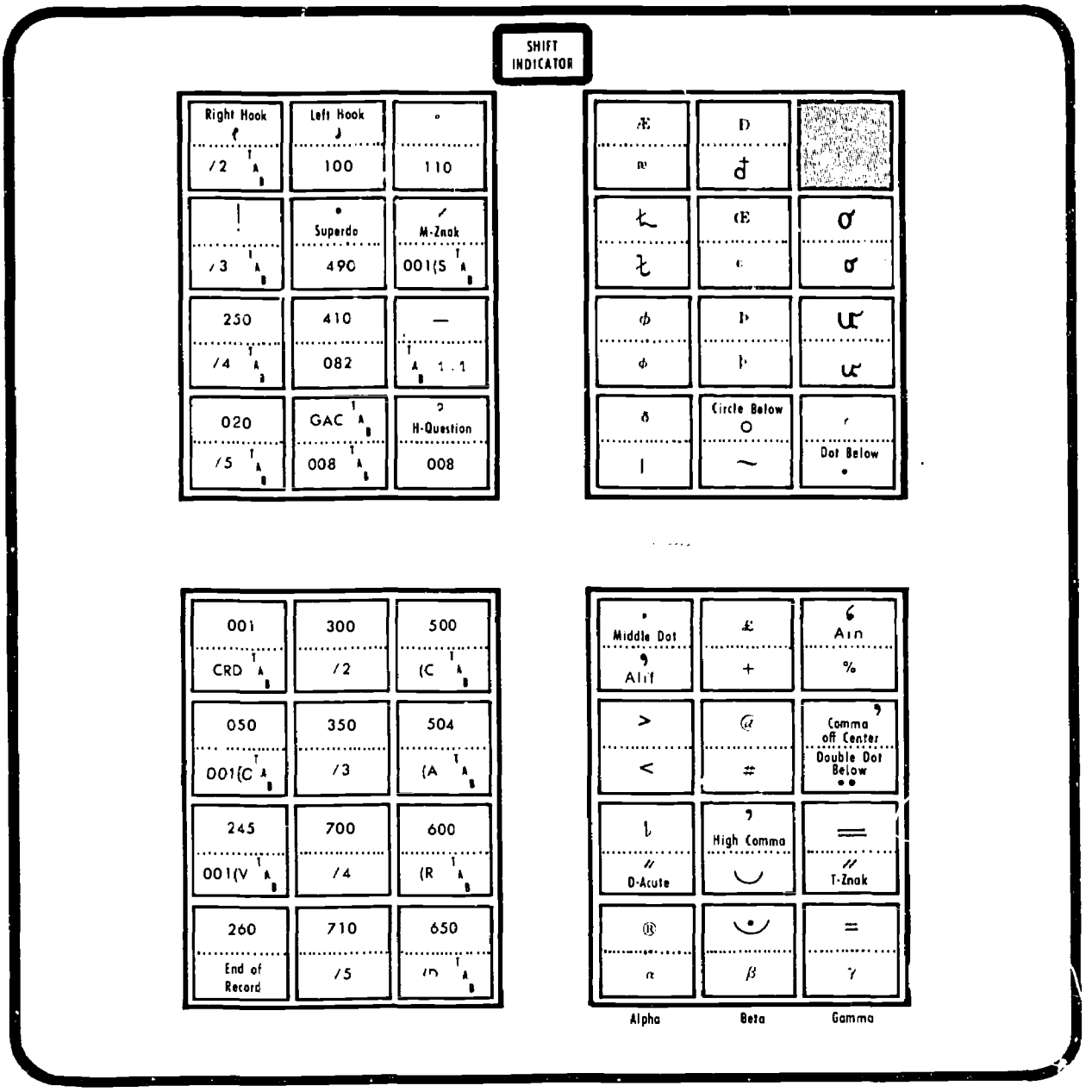


Figure 7-8. Keymatic auxiliary keyboard.

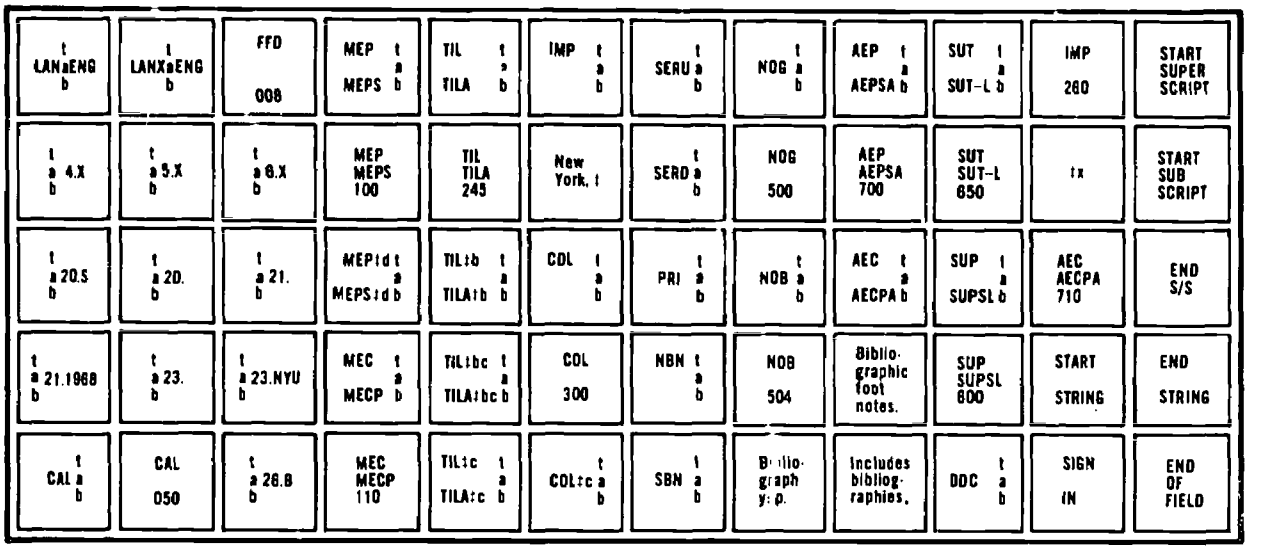


Figure 7-4. Keymatic function control panel.

lems described above still remain, and in addition, the K-103 requires the use of a converter as does the MTST.

Comparison of MTST and OCR Methods

Keying of RECON records in the 1969 and 7 series was performed by a contractor using an IBM Selectric typewriter equipped with an OCR font. The resulting hard copy was fed through a Farrington optical character reader. The contractor monitored and reported the production rates for his equipment, and these were compared with corresponding data for the MTST's used in the Library.

The cost of the typewriter with the OCR font ($500.00) was amortized over 40 months fo a monthly cost of $12.50. If the typewriter were used 132 hours a month, the hourly cost would be $.10. The contractor reported a typing rate of 12 records per hour or $.01 per record for the typewriter. The service bureau rental cost for the contractor's Farrington scanner, which can read 10,000 lines per hour (or 556 records averaging 18 lines in length) is $50.00 per hour (or $.09 per record). The contractor's total equipment cost per record is $.01 for the typewriter with the OCR font, plus $.09 for the scanner, or a total of $.10. This cost is quite close to that for the MTST equipment of $.07 per record.

Even if typing with OCR turned out to be less expensive than with the MTST, this method would probably not be satisfactory. The Farrington scanner is capable of reading only uppercase. Since bibliographic data contain fewer uppercase than lowercase characters, a shift character was typed in front of each uppercase character. The resulting hard copy was difficult to read and contained many typing errors. In order to attain a degree of accuracy comparable to that of the Library's typists, the contractor found it necessary to proof and correct all records before returning them to the Library for further processing. Records typed on the MTST at the Library are not proofed before being processed by the computer. Proofing of uncorrected OCR records might decrease the editor's production to a point that the higher manpower costs would more than offset any savings on equipment. OCR equipment with uppercase and lowercase capability is now becoming available, but it must be assumed that the rental on such equipment will be higher.

Direct-Read Optical Character Readers

An obvious advantage for the use of a direct-read optical character reader is elimination of the need for manual keying of the original input. With format recognition a proven technique, the use of such a device has even greater possibilities. Bibliographic data could be read,

edited, formatted according to MARC format specifications, and output as a proofsheet with a minimum of human intervention.

There are two types of optical character readers available: 1) document readers, which read only a limited number of lines per document, e.g., a credit card; and 2) page readers, which are capable of reading an entire page. Because of the input requirements of bibliographic data, document readers were not considered in this study.

An OCR device accomplishes its recognition by flooding the document with light and analyzing the reflection. Light patterns are captured in a photomultiplier and converted into an electronic signal. In general, these signals are matched against either memory or logic circuitry, and a corresponding code configuration is output onto the desired medium, e.g., disk or tape. Each manufacturer has specific requirements for the type of paper used and style of printing recognized.

Machines were considered as possible candidates if they were capable of processing uppercase and lowercase alphabetic as well as numeric characters, standard punctuation, and some special symbols. The special characters available, which vary among the manufacturers, can be separated into two categories: edit function characters, and those characters that cannot be categorized as alphabetic, numeric, or standard function characters, e.g., "&" or "[]."

Equipment produced by the following manufacturers was considered for the initial evaluation:

Information International
Mergenthaler-Linotype
Control Data Corporation—915 Page Reader
IBM Corporation—Model 1287
Farrington—Models 3050 and 3030
Scan-Data—Models 100 300, 200
Philco-Ford, Inc.—General Purpose Reader
Recognition Equipment, Inc.—Retina
CompuScan—Model 370

Although Mergenthaler-Linotype and Information International did not have any device commercially available at the time, each did have a machine in production. Two of the companies subsequently gave up their efforts toward direct-reading of the LC printed card because of the complexity of the data on the card.

Investigation of the remaining devices revealed that all except the CompuScan and Scan-Data, required keying of data before reading with the scanner. Some manufacturers claimed that their devices had the potential to read the LC printed card directly but required substantial funding for hardware and or software development that was out of range for the RECON Pilot Project. As a result, tests could only be conducted on the CompuScan and Scan-Data.

CompuScan Optical Character Reader

The Model 370 CompuScan is a computer-directed flying-spot scanner which matches the scanned portions of a character with an electronic character held in the core memory of the computer. The record set would have to be microfilmed according to the specifications required by the scanner. Since the scanner operates with negative film, a very dark background with a very clear, white image is necessary.

The manufacturer examined a sample of LC printed cards selected at random covering a 10-year period and concluded that although the hardware would be sufficient to read the record set optically, a rather significant software effort would be necessary.

The LC record set is not entirely composed of "mint" cards (cards printed from the metal of the original Linotype composition) but instead is a mixture of originals and reprints of the original. When the stock of the original printing is close to depletion, the card is reprinted by photographing the card and making duplicates by a photo-offset process. As this cycle is repeated, the card for any one title could be several generations removed from the original. In some instances, a microscopic examination of the cards seemed to indicate that the matrices used on the Linotype were worn. Thus, what might appear as the same character to the naked eye would present a different pattern configuration to the scanner.

The coarseness of the surface of the card itself may cause variations in the same character. To achieve the archival standards re-

quired by libraries, LC cards are printed on high-rag-content stock. The rough surface of the card does not affect readability for a human but may cause variations in a given character. Software must be written to handle these variant characters and to match them with characters in the core memory of the scanner.

Another significant problem in dealing with LC cards concerns touching characters, a connection between what are intended to be distinct characters but read by the scanner as one. For example, if a lowercase "n" were next to a lowercase "t" and the cross bar on the "t" touched the "n," the scanner would consider the combination of the "n" and the "t" as one character. Another module of software is required to set an allowable limit for reading a single character so that the machine will recognize touching characters as separate entities. When this limit is exceeded, the pattern must be divided and each section matched against a single character pattern held in core. A machine decision must then be made to identify the two patterns.

When variant or touching characters occur, the output on magnetic tape is flagged for later spot checking. In this way, the scanner can continue to operate at throughput speeds without human intervention. The resultant magnetic tape would serve as input to the format recognition program to reformat the scanner's output into the MARC format. It has been estimated that the throughput speed of the CompuScan would be in the vicinity of 1,800 cards per hour.

The manufacturer offered to use originally printed LC cards to test the device without expending funds for software modifications. Twenty-five letterpress LC cards representing English language titles and containing no diacritical marks were sent to the firm for input. Since existing CompuScan software was used for the test, only the portion of the LC card containing fonts already built into the existing configuration could be used. All data except the title paragraph (title through imprint) were blocked out before microfilming for subsequent scanning.

Operator intervention was required for 1 to 25 percent of the characters on each card. In addition to the problems described above, fine lines in certain characters caused a misreading of the character by the scanner, the letter "e," for example, being interpreted as the letter "c." CompuScan felt that this problem might be resolved by increasing the size of the comparison matrix of the hardware. Another problem encountered was that a period was generated in the middle of a word due to the coarseness of the card stock.

Scan-Data

Dissly Systems has effectively modified the Scan-Data optical character reader, via software, to read 55 different type fonts. The various fonts are recognized by a "best compare" technique using six stored fonts to match against the remaining 49. The manufacturer claims that direct-read is accomplished with accuracy levels of approximately 95 percent. Errors are flagged during a proofing cycle after the record is in machine-readable form and corrected in the machine data base.

The Scan-Data equipment does not have a transport for a 3 x 5 document, and the LC cards must therefore be attached to an 8 x 14 sheet for scanning. Since the manufacturer cannot return these cards to the Library, they would have to be taken from stock rather than from the record set. This constraint places severe limitations on the application of the Scan-Data since many cards are out of stock and those that are in stock may be second- or third-generation cards which, as indicated above, are not ideal candidates for direct-read scanning.

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

Cathode Ray Tubes

Cathode ray tubes were not considered for original input but rather as an aid for the MARC correction and verification cycles. The CRT devices essentially fall into two categories: graphic terminals and alphanumeric terminals. A graphic terminal, best described as a line drawing unit, is used primarily in applications involving the drawing of plans, schemat-

ics, etc., with a minimum of associated text. An alphanumeric terminal is similar to a typewriter in that it can display alphanumeric and special characters.

All CRT devices have the same basic operating components: screen, memory, keyboard, character generator, and a set of electronics that tie the components together in a unit capable of communication. Various options are available, including additional peripheral devices, expanded character sets, and editing features. Some devices are linked to minicomputers that allow data manipulation at the terminal.

CRT specifications were developed for the LC MARC character set, both for keying and display, and for editing functions which would allow insertion and deletion of characters, words, or lines. The device would have to communicate with the LC hardware configuration and be adaptable to other major manufacturers' hardware in case of changes in the configuration in the future. Minimum requirements were established for viewing areas, character size, character capacity of the screen, etc.

Equipment was evaluated by matching the capabilities of a device against the specifications. Few devices met the requirements, with the primary limitation being the character set. Most of the displays have a 64-character capability, some offer 96 characters as an option, and a few are capable of expansion to 128 characters.

The two devices that conformed most closely to LC specifications were studied in greater detail. The first, the Irascope Model LTE, built by Spiras Systems, Inc., was developed in conjunction with the Ohio College Library Center. The limitation of the Irascope was the size of the character set. The device permits keying and transmission of 155 characters, but only 128 characters can be displayed.

The second device, the PDS-1, is manufactured by the IMLAC Corporation and has both graphic and alphanumeric capabilities. The standard equipment includes a 4K miniprocessor; through the use of software, characters of any shape can be displayed. There are 196 characters that can be keyed, and the resultant 196 unique codes can be translated into any 196 shapes for display. Although precise figures are not yet available, it was determined that the cost of the PDS-1 is higher than that of the Irascope LTE. The final decision was to acquire the Irascope for use in MARC correction cycles.

Mini-Computers

The Library also conducted an investigation to determine the feasibility and desirability of using a mini-computer on-line for MARC input functions (both original input and corrections). This study was performed with contractual support. Benefits which might result from converting MARC to an on-line system for input included, in addition to increased productivity:

1) Improved timeliness of data released to the MARC Distribution Service.

2) Savings in IBM 360 computer time required to process MARC records from the point of input to the stage at which they are declared error free and transferred to the master data base. Assumption of various input functions by the mini-computer would relieve the main computer of these functions, and the 360 would thus be required only to process verified records on the master data base.

This survey, conducted in late 1969, was not intended to be all inclusive. Time and funding were limited, and since the mini-computer field is expanding rapidly, it was not possible to have surveyed the totality at any given cut-off point. Inquiries were directed to seven firms known to manufacture and market mini-computers. Six of the firms responded with descriptions of devices that were considered potentially applicable to MARC operations. These included the Burroughs TC-500, Digital Equipment Corporation PDP-8/I, Honeywell DDP-516, IBM 1800, Interdata Model 4, and XDS Sigma 3. Of these, the DEC PDP-8/I and the Honeywell DDP-516 had the greatest potential for meeting the requirements for MARC input processing.

Most manufacturers offer programming support on an individually negotiated contract basis. All of the mini-computer manufacturers covered in this study supplied an assembler as well as debugging and editing routines. Several provided a FORTRAN compiler and an operating system; however, the minimum cost of a

system that supports compilers and operating system storage is approximately $10,000. In addition, the mini-computer may include only a few standard features in its basic system, and addition of the optional features necessary for a given application can make the price substantially higher than that quoted for the basic system.

On the basis of this study, it was concluded that although the use of a mini-computer is technically feasible in performing MARC input functions, addition of a mini-computer to the present LC hardware configuration would not result in either technological or economic gains. Specifically, the processing load removed from the 360 computer by the mini-computer would not be sufficient to justify the added cost imposed by the latter system.

The experience of processing MARC records in the LC environment during the past several years has indicated that there is no gain in original input on-line but a great deal to gain with on-line correction procedures. This fact weakens considerably the argument in favor of devoting a separate mini-computer subsystem to original input and correction procedures, since the correction aspect alone represents a much smaller load factor.

In addition, the procedures under consideration include corrections to both the working files and the MARC master data base. Since the Library's requirements for handling large files and sophisticated access structures are beyond the capabilities of a basic mini-computer at the present time, the extent to which the mini-computer could handle correction procedures is quite limited.

Under the MARC input procedures in effect at the time of the mini-computer study, editing of the records was carried out manually. Because editing is now being accomplished by means of the format recognition program, a reassessment of the mini-computer may be in order. Since the success of format recognition depends on accurate typing, greater flexibility in correcting simple typing errors before processing would promote greater accuracy in machine editing.

Notes

[1] RECON Working Task Force. *Conversion of Retrospective Catalog Records to Machine-Readable Form; a Study of the Feasibility of a National Bibliographic Service* (Washington, Library of Congress, 1969), p. 52–55.

CHAPTER 8

Microfilming Techniques

As part of the RECON Pilot Project, microfilming techniques and their associated costs were investigated in cooperation with the staff of the Library of Congress Photoduplication Service. The possibility of obtaining cost estimates for commercial microfilming was considered but was finally rejected on the grounds that spending staff time to explain the project to a contractor could not be justified when the Library had its own fully qualified photoduplication laboratory.

The RECON feasibility report recommended that priority be given to the conversion of records for English-language monographs from 1960 to 1968. It was noted, however, that certain problems arise in connection with the use of the record set for any category of materials, since this file is arranged by card series (year) and by sequential number within each series. The file can be readily divided into one-year segments from 1898 through 1968,[1] but the card numbering system does not lend itself easily to a division of the file by language or form of material.

The RECON report recommended that the record set be divided into categories according to conversion priority, the cards filmed, and the file then reassembled. It was considered that selection of categories for conversion before filming would be more efficient since fewer cards would have to be filmed. Further study during the pilot project indicated that the entire record set containing the category of material chosen for conversion should be microfilmed and then culled for the titles to be converted. Although this method results in the filming of more cards, it presents the following advantages:

1) The microfilm copy, containing the records for all languages and forms of material in the series chosen, can be used again for any other category of conversion. The need for returning to and disrupting the arrangement of the record set to select another category is thus eliminated.

2) The records can be filmed as found, with a minimum of intervention by the operator. Selection of a particular language or form of material would require an individual with a knowledge of bibliographic data, and the filming would be slowed down by the selection process.

3) The microfilm can be retained as a security copy of the record set.

4) The figure for the number of cards printed in a given year is more accurate than the figure for the number of records representing a category of material for the same time period; hence, more reliable cost estimates could be established.

Certain ground rules were established for the actual filming process. The selected drawers of the record set would be "frozen" for a day or two before filming, i.e., cards known to be out of the file would be refiled, and no cards would be removed from the file while filming was in process. The filming would take place during normal work hours.

Once the decision was made to film first and select later, it was necessary to ascertain the volume of cards to be used as a basis for cost estimates. Since Photoduplication Service cost estimates are firm for only a one-year period because of the effects of increases in salaries, cost of materials, etc., and because there would

be paper handling problems in managing large quantities of worksheets for conversion, it appeared reasonable to assume microfilming rates in proportion to conversion rates rather than attempting to project cost estimates for filming the entire record set.

A volume of 100,000 cards for the year 1965 wa chosen as a base figure for computation. It was estimated that one operator could film approximately 5,000 cards per day, and approximately 20 working days would be required to film the collection of cards representing one year of the record set. In preparing the cost estimates, it was assumed that quality control would be limited to inspection for technical requirements only, with a spot check about every 300 images for camera operator errors. It would probably be less expensive to correct other errors as they were discovered during the conversion process. It was anticipated that these errors would not exceed one percent of the total number of exposures. There would be no inspection for bibliographic content nor would any attempt be made to guarantee file sequence, i.e., card numbers could be missing.[2]

Based on the method of filming before selection and on the volume of cards cited above, cost estimates for the following alternative techniques were derived:

1) Microfilming for direct-read optical character reader specifications.

2) Microfilming for reader/printer specifications.

3) Microfilming for reader specifications.

4) Microfilming for Xerox Copyflo printout of LC printed cards overlaid on 8 x 10½ inch worksheet.

The following definitions are given to help in understanding the techniques described:

1) *Planetary camera*—a microfilm camera in which, during exposure, the film is held stationary in a horizontal plane parallel to the item being copied.

2) *Rotary camera*—a microfilm camera in which loose-sheet documents are transported on the surface of a rotary drum past a lens which records the document on a roll of film moving synchronously with the rotation drum at a speed equal to the reduction ratio. Although the unit cost per exposure is less for a rotary camera, the quality of the image may be inadequate for some purposes.

3) *Film*—in all four techniques, the film used is 16-mm negative microfilm.

4) *Reduction ratio*—a numerical expression of the number of times a copy is smaller in size, linearly, than the original from which it was made; expressed either in diameters (e.g., 5X, 14.5X, 20X) or as a ratio (e.g., 1:5, 1:14.5, 1:20).

5) *Image position*—the orientation of images on a roll which can be controlled by turning either the document or the camera head and adjusting the reduction ratio accordingly. There are two basic positions: horizontal (1A), with the head of the image to the left of the frame, and vertical (1B), with the head of the image at the top of the frame.

6) *Feed*—the method of transporting the document to be filmed to the camera head.

7) *Paper stock*—the type of paper used in restoring images to eye-legible copy (hard copy).

8) *Rate per exposure (microfilm)*—unit price per image for microfilming.

9) *Rate per exposure (print)*—unit price per image for restoring film to eye-legible copy (hard copy).

Microfilming for OCR Specifications

The study on input devices (Chapter 7) demonstrated that the present state-of-the-art is such that a direct-read OCR cannot be used to scan LC printed cards. The microfilming technique for the OCR is included in the present comparison on the chance that the capabilities of these devices may improve significantly in the future and to isolate problems that might arise in using the OCR for a large retrospective conversion project. It is assumed that procedures for the use of the OCR would be as follows:

1) Microfilming of LC cards.

2) Automatic reading of the film and transfer of the data in digital form to a magnetic tape.

3) Use of a format recognition technique to create a machine record in the MARC format.

4) Printing of a bibliographic record on the computer printer for proofing.

5) Selection of records for the category to be converted.

6) Comparison of the computer-produced hard copy record with the main entry in the Official Catalog and updating of the record where required.

7) Proofing of the computer-produced record against a hard copy source document for format recognition errors.

8) Subsequent file maintenance procedures.

It should be noted that point 7 above assumes a source document in hard copy form for human readability. Comparison of the computer-produced proofsheet with the microfilm copy of the LC card by using a microfilm reader would place such a significant burden on the editor that it seemed unrealistic to consider this procedure.

CompuScan specifications for a density range of 1.6 to 1.8 were used as the norm for the requirements of OCR devices. Serious problems would arise in using the same film on Xerox Copyflo or even for contact printing to positive film because the density of 1.6 to 1.8 is not ideal for reproducing LC printed cards. The existence of heavily inked small characters and fine lines on the cards requires holding density to the 1.3 to 1.35 range to reproduce all text. Film suitable for OCR requirements would thus have little value for printout purposes, and a second filming would be necessary to provide hard copy. The cost estimate given below does not include the cost of this second filming.

It would not be feasible to employ a rotary camera for production of film suitable for OCR devices since it would not be possible to ensure alignment of each image. In fact, there would be no guarantee that even a small portion of the images would be at a right angle to the edge of the film if a rotary camera were used. It thus appears that OCR requirements demand stop-motion photography and single-card exposure, using a planetary camera.

Camera	Planetary
Film	16 mm
Reduction	20X
Image position	1A
Feed	Hand
Rate per exposure for negative	$.02
Cost for 100,000 cards	$2,000.00

Microfilming for Reader/Printer Specifications

This method assumes that the hard copy produced from the microfilm would be in the form of a MARC/RECON input worksheet. During the filming process, a form must be imposed on the image in such a way that the resultant film copy has the 3 x 5 card positioned to the right of the worksheet. The use of a reader/printer involves the following procedures:

1) Microfilming of LC cards.

2) Reading, via the reader, to select records for conversion; printing, via the printer, of selected records as hard copy source documents.

3) Comparison of the hard copy source document with the main entry in the Official Catalog and updating of the record as required.

4) Keying of the record.

5) Use of format recognition to create a machine record in the MARC format.

6) Printing of a bibliographic record on the computer printer for proofing.

7) Proofing of the computer-produced record against the hard copy source document for typing or format recognition errors.

8) Subsequent file maintenance procedures.

A rotary type camera would not be suitable for this technique since it does not provide the means for controlling the image position or for superimposing an input worksheet form on each image. The use of a stop-motion camera, with each image overlaid with an input worksheet form, seems appropriate.

Camera	Planetary
Film	16 mm

Reduction	16X
Image position	1B
Feed	Hand
Rate per exposure for negative	$.0235
Cost for 100,000	$2,350.00

Microfilming for Reader Specifications

This method does not provide hard copy, and its use would be unlikely because it makes keying and proofing extremely difficult. The following procedures would be required:

1) Microfilming of the LC cards.

2) Reading, via the reader, to select records for conversion; keying of selected records directly from the screen of the microfilm reader.

3) Use of format recognition to create a machine record in the MARC format.

4) Printing of a bibliographic record on the computer printer for proofing.

5) Comparison of the computer-produced record with the main entry in the Official Catalog and updating of the record where required.

6) Proofing of the computer-produced record against the Official Catalog record for typing or format recognition errors.

7) Subsequent file maintenance procedures.

This method requires keying both before and after catalog comparison. Additional keying may also be necessary because of changes made to the record in the Official Catalog. Because there is no hard copy source document, the proofing must be done against either the Official Catalog record or the copy displayed on the microfilm reader. Since an input worksheet is not produced in this method, a rotary camera may be used. A person reading the record on the microfilm reader would not be seriously hampered by uneven placement of the image on the screen.

Camera	Rotary
Film	16 mm
Reduction	20X
Image position	1A
Feed	Automatic
Rate per exposure for negative	$.004
Cost for 100,000 cards	$400.00

Microfilming for Xerox Copyflo Printout

This method of microfilming is employed for the sole purpose of providing hard copy source documents (in the form of the MARC/RECON input worksheets). The following procedures apply:

1) Microfilming of LC cards and production of worksheets.

2) Selection of records for conversion from the worksheets.

3) Comparison of the worksheet with the main entry in the Official Catalog and updating of the record as required.

4) Keying of the record from the worksheet.

5) Use of format recognition to create a machine record in the MARC format.

6) Printing of a bibliographic record on the computer printer for proofing.

7) Proofing of the computer-produced record against the hard copy source document for typing or format recognition errors.

8) Subsequent file maintenance procedures.

The use of a rotary camera would not be practical for the same reasons as discussed in connection with the microfilming for reader/printer specifications.

Camera	Planetary
Film	16 mm
Reduction	16X
Image position	1B
Feed	Hand
Paper stock	20-lb sulfite
Paper size	8 x 10½ overall
Rate per print (8 x 10½), including microfilming	$.07
Cost for 100,000	$7,000.00

Investigation of a Technical Alternative

A prototype mechanism developed by Developtron, Inc., was investigated and evaluated for purposes of retrospective conversion. The device consists of a scissor-type rig, approximately 5 feet long and 3½ feet high, with a Leica 35-mm camera mounted at the top. The

apparatus is placed on a table top, and the operator sits at the scissor point with a tray of catalog cards aligned along the lower blade. The camera, which is mounted on the upper blade, is either lowered into the tray or positioned immediately above the tray at the option of the operator. Cards can be filmed in place or raised out of the tray for filming.

The vendor suggested that a 24X Kodak RV2-Starflite camera head be substituted for the Leica since 16-mm unperforated film used with the RV2 would result in better resolution. Thirty-five mm perforated film at a 4X reduction was used in the demonstration.

Although the device appeared to function adequately, its use did not appear to offer any cost advantage over conventional microfilming. The savings in microfilming and hard copy costs would be offset by the slowness of the process and the fact that it would have to be repeated each time another category of records was selected from the same segment of the file. The cost of mounting the hard copies on worksheets would also have to be taken into account.

Conclusion

Despite the higher unit cost, it appears that the best alternative is to film all cards in a given series against a worksheet form to produce hard copy (Xerox Copyflo printout) and then to select the desired subset for conversion. Since the use of projected microfilm images as source documents is impractical, the reader-only option can be eliminated. The OCR method could not be employed unless a device were developed that could accurately scan LC printed cards. Use of the reader/printer method is a possibility, but the quality of hard copy print would be inferior to that obtained in the Xerox Copyflo printout. The cost of the reader/printer method, which does not include the cost of hard copy, varies with the device selected and could well approach that of the Xerox Copyflo method.

Notes

[1] From 1969 until early 1972, cards in the record set were arranged not by specific calendar year but rather by numbers in the 7 series, with the second digit being a check digit. The year-series numbering was resumed in February 1972.

[2] Gaps in the sequence of card numbers sometimes exist because certain numbers given to a publisher before publication are not used; however, a gap could exist in the file because a card was actually missing.

APPENDIX I

MARC Decisions for Retrospective Cataloging

After the analysis of 5,000 research titles was completed, problems concerning cataloging and MARC editing procedures were brought to the attention of appropriate personnel at the Library of Congress. Based on the discussions with these staff members, decisions were made to handle the problems as follows:

LC Card Numbers

1) A single dagger after a card number, e.g., 10-4173†, should be deleted.

2) A second hyphen followed by a digit, e.g., 1-6360-1 or 1-6360-1 Revised, should be deleted and input as 1-6360//r38 (or whatever date appears with the revision symbol).

3) When "Revised" (but no revision date) follows the card number, use the date of cataloging found on the verso of the Official Catalog card.

4) When an asterisk follows the number, e.g., 8-30156*, delete the asterisk.

5) LC card numbers such as F-3144 should have an Ø1 added after the "F," e.g., FØ1-3144, since 1901 was the year in which they were printed.

6) For the present, card numbers with a leading digit greater than "7," e.g., 99-1974, cannot be input because a check digit error message is generated. These records should be held aside until the programs have been modified to accept these card numbers.

Main Entry

1) Single surnames without forenames will be transcribed with three spaces following the comma (△ = 1 space), e.g., Dezauche, △△△ When such names appear as added entries, they will be transcribed with three spaces rather than with a long dash, e.g.. Dezauche, ——— would become Dezauche, △△△.

Title Added Entry Indicator

1) Before 1912, the printed cards contained no indication of whether the titles should be traced. Although titles have been traced after 1912, these records do not have as many title tracings as they might under current practice. Input these records without adding any title added entry indicator.

Title Statement

1) Ellipses occurring at the beginning of the title should be removed unless they are printed as bold dots. Ellipses in the "c" subfield of the title should also be removed. All other ellipses will be included in the record.

2) Line endings, used to distinguish two editions of a rare book, are indicated as two vertical lines. Replace with △/△.

3) Input superscript or subscript alphabetic characters as regular lowercase alphabetic characters, except in formulas, e.g., $A - B^{(n-1)}$, which is input as A — B '[superscript n]-1)

4) An asterisk and a single dagger are used to indicate birth and death dates of a person, e.g., *Chiquinquira, 21 de mayo de 1857. Usiacuri, 7 de febrero de 1923, or von Norbert Klüken und Karl Hoffmann†. If there is a birth/death date phrase, delete it from the title statement.

If there is only a single dagger following a personal name in an author statement, delete the dagger.

Imprint Statement

1) When the reprint statement has the appearance of a double imprint, e.g., Bonnae, Apud Henry & Cohen, 1856; Frankfurt/Main, Minerva, 1967, the actual reprint statement should be separated from the imprint by a period instead of a semicolon. The reprint statement will not be considered part of the imprint field.

2) In cases where place of publication has the appearance of a street address, e.g., 72-Souligne-sous-Ballon, l'auteur, 1968, "72" is actually the zip code zone and Souligne-sous-Ballon is the name of the town. The two together should be considered as the place of publication.

3) If neither a date of publication nor the abbreviation [n.d.] is present, e.g., Paha, SNTL, this is an error. Refer this record to the cataloger.

4) When an incomplete place name is given as the place of publication, e.g., Rio, Editôra Simões, 1956., this is an error. Refer such a record to the cataloger. (The example cited should read: Rio [de Janeiro].)

5) When two places of publication are separated by "and," e.g., New York and London; by "und" or "u.," e.g., München u. Hannover; or by a hyphen, e.g., Paris-Bruges or Milano-Roma-Napoli, delete the conjunction or hyphen and add a comma to separate the two place names. On German records, separation of two place names by a hyphen generally indicates that one place is located near a larger, better known place, e.g., Hamburg-Altona; such entries should be considered as a single place of publication. Occasionally, the hyphen is used on German records to indicate two places of publication. Such records should be given to a supervisor to check. On German records when two place names are separated by "bis" or "b.", e.g., Ratingen b. Düsseldorf, they constitute a single place of publication.

Collation Statement

1) When the statement "Cover title" is included in the collation, e.g., Cover title, 36 p. 21 cm., delete "Cover title" from the collation and make it the first note.

2) If a statement of illustration is given in the title paragraph but not in the collation, e.g., Hrsg. von Norbert Klüken. Mit 24 Abbildungen und 13 Tabellen., leave the collation statement as it is and do not add illustration codes to the fixed field.

3) When illustrative information is given as a general note, e.g., Maps on lining papers, do not add maps to the collation statement and do not add an illustration code to the fixed field.

4) When a statement of reprint is included in the collation, e.g., reprint: 2 v. in 1. 29 cm., delete "reprint" from the collation.

Series Statement

1) Many older records carry information in the series statements which is not used in the *Rules for Descriptive Cataloging in the Library of Congress* or *Anglo-American Cataloging Rules*, e.g., Half-title: Library of philosophy. Ed. by J. H. Muirhead. Delete the words "Half-title:" or the editor phrase from the series statement.

2) Some older records have what appears to be a "bound with" note transcribed in the series statement position and a series statement transcribed in a note position, e.g.:

xxiv, 466 p. 17 cm. [With, as issued: Manetho, the historian. Manetho. Cambridge, Mass., London, 1940] Greek and English on opposite pages.
Half-title: The Loeb classical library . . . Manetho. Ptolemy, Tetrabiblos

Tag a field according to what it is rather than where it is.

General Notes

1) Complex notes, e.g.:

The following information regarding dates of publication of each volume is supplied by Dr. C. Wardell Stiles of the U.S. Department of Agriculture:
v. 1. Jan. 1886–6 May 1886.
v. 2. 13 May 1886–28 Oct. 1886.

v. 3. 4 Nov. 1886–21 Apr. 1887.
v. 4. 28 Apr. 1887–13 Oct. 1887.
v. 5. 20 Oct. 1887–5 Apr. 1888.
v. 6. 12 Apr. 1888–27 Sept. 1888.
[etc.]

Refer these records to a supervisor, who will in turn take them to the principal cataloger for a decision.

Subject Headings

1) Subject headings from other libraries (old cooperative copy): With the exception of those records that contain the legend "Shared Cataloging with DNLM" or "Shared Cataloging for DNAL," only LC subject headings (including those for children's literature) will be used. Subject headings from other libraries should be deleted.

a. Older records sometimes contain subject entries that are composites of headings from the Library of Congress and other libraries. The other libraries' headings are in brackets. Delete subject headings or parts of subject headings that are enclosed in brackets, e.g.:

[1. Labor supply—Stat.—Russia] Delete the entire heading. 1. Fruit[—Hardiness] Delete only [—Hardiness]

This rule does not apply to LC children's headings or cards with the legend: "Shared Cataloging with DNLM" or "Shared Cataloging for DNAL."

b. Some retrospective records contain portions of subject headings enclosed in subscript parentheses, e.g.:

1. Wages—(Furniture workers)—United States.

Delete only the subscript parentheses. Retain the data within.

c. Some retrospective subject headings contain both bracketed portions and portions enclosed within subscript parentheses, e.g.:

1. Spraying and dusting residues (in agriculture) [—Testing]

Delete the subscript parentheses around "in agriculture," but retain the data; also delete [—Testing], e.g.:

1. Spraying and dusting residues in agriculture.

2) Personal names without dates used as subject headings:

a. The ALA *Cataloging Rules for Author and Title Entries* contains a list of personal names that may be used as subject headings without dates. Since this practice is no longer followed, the following names should have dates added when they are used as subjects.

Ariosto, Lodovico, 1474–1533.
Bach, Johann Sebastian, 1685–1750.
Bacon, Francis, viscount St. Albans, 1561–1626.
Balzac, Honoré de, 1799–1850.
Beethoven, Ludwig van, 1770–1827.
Boccaccio, Giovanni, 1313–1375.
Browning, Robert, 1812–1889.
Bunyan, John, 1628–1688.
Burns, Robert, 1759–1796.
Byron, George Gordon Noël Byron, 6th baron, 1788–1824.
Carlyle, Thomas, 1795–1881.
Cervantes Saavedra, Miguel de, 1547–1616.
Chaucer, Geoffrey, d. 1400.
Colombo, Cristoforo. [no dates on authority record]
Corneille, Pierre, 1606–1684.
Cromwell, Oliver, 1599–1658.
Dante Alighieri, 1265–1321.
Dickens, Charles, 1812–1870.
Eliot, George, pseud., i.e., Marian Evans, afterwards Cross, 1819–1880.
Goethe, Johann Wolfgang von, 1749–1832.
Goldsmith, Oliver, 1728–1774.
Hawthorne, Nathaniel, 1804–1864.
Heine, Heinrich, 1797–1856.
Hugo, Victor Marie, comte, 1802–1885.
Ibsen, Henrik, 1828–1906.
Irving, Washington, 1783–1859.
Lessing, Gotthold Ephraim, 1729–1781.
Lincoln, Abraham, Pres. U.S., 1809–1865.
Longfellow, Henry Wadsworth, 1807–1882.
Luther, Martin, 1483–1546.
Marie Antoinette, consort of Louis XVI, King of France, 1755–1793.
Milton, John, 1608–1674.
Molière, Jean Baptiste Poquelin, 1622–1673.
Mozart, Johann Chrysostom Wolfgang Amadeus, 1756–1791.
Napoléon I, Emperor of the French, 1769–1821.
Petrarca, Francesco, 1304–1374.
Pushkin, Aleksandr Sergeevich, 1799–1837.
Racine, Jean Baptiste, 1639–1699.
Rousseau, Jean Jacques, 1712–1778.
Ruskin, John, 1819–1900.
Schiller, Johann Christoph Friedrich von, 1759–1805.
Scott, Sir Walter, bart., 1771–1832.
Shakespeare, William, 1564–1616.
Spenser, Edmund, 1552?–1599.
Tasso, Torquato, 1544–1595.
Tennyson, Alfred Tennyson, baron, 1809–1892.
Thackery, William Makepeace, 1811–1863.
Tolstoĭ, Lev Nikolaevich, graf, 1828–1910.
Voltaire, François Marie Arouet de, 1694–1778.
Wagner, Richard, 1813–1883.
Washington, George, Pres. U.S., 1732–1799.

Title Added Entries

1) On some retrospective records, titles have been inverted, e.g., I. Title: Retail Terms, A manual of. These will be tagged as titles traced differently (tag 740). Such a record will not have a title added entry generated from the title field.

Series Tracings

1) Before 1947, series statements were not traced on the printed cards. The tracing (if one were present) was recorded on the main entry card in the Official Catalog. During catalog comparison, check the verso of the main entry Official Catalog card for a series tracing for all records cataloged before 1947 and all records cataloged after 1947 that do not have an * after the card number. If the verso of the main entry card contains a series tracing, transcribe it on the input worksheet.

2) Limited cataloging records do not contain series tracings. They can be identified by a double dagger after the card number, e.g., 54-49564‡. Leave these records as they are. The double dagger following the card number will be deleted and a "/L" substituted in its place.

Full Name Notes (also Secular Name, Name Originally, etc.)

1) Some older retrospective records have full name notes recorded on the right hand side of the card between the tracings and the card number, e.g.;

1. London--Description. I. Title.
[Full name: William Richard Gladstone Kent]
37-28551

Delete these notes.

2) Asterisks preceding added entries indicate that the personal name has been revised. If this name were used as a main entry, a name-originally note would be present. Delete asterisks before personal name added entries.

Copy Statement

1) Copy statements without call numbers have been written or typed on some Official Catalog main entry cards, e.g.,

— — Copy 2
— — Copy 3

Do not transcribe copy statements that are not printed on the printed card.

Copyright Number

1) Copyright numbers have been added to the printed cards at various times. They are recorded in the lower left hand corner, below the Library of Congress legend, e.g.:

Copyright A 29724

Delete the copyright number.

Diacritics

1) Old German uses a small e instead of an umlaut (¨) over a, o, and u, e.g., päbstern, königen, fürsten. Convert the e's to umlauts (¨).

INDEX

☆U.S. GOVERNMENT PRINTING OFFICE: 1972 O—469-083