ENTRY/TITLE COMPRESSION CODE ACCESS TO MACHINE READABLE BIBLIOGRAPHIC FILES

William L. NEWMAN: Systems Analyst and Programmer, and
Edwin J. BUCHINSKI: Assistant to the Librarian, Systems and Planning, University of Saskatchewan, Saskatoon.

An entry/title compression code is proposed which will fulfill the following requirements at the Library, University of Saskatchewan: 1) entry/title access to MARC tapes; 2) entry/title access to the acquisitions and cataloguing in-process file; and 3) entry/title duplicate order edit within the acquisitions and cataloguing in-process file. The study which produced the code and applications for the code are discussed.

INTRODUCTION

The determination and design of access points, or keys, to machine readable bibliographic files is a major problem faced by libraries planning computer assisted processing. Alphabetic keys, i.e. truncations of title and/or author variable fields, are inadequate, since minor differences in spelling, punctuation, or spacing between master key and request key cause difficulties in accessing records. Numeric keys, such as Library of Congress card numbers, ISBN, purchase order numbers, etc. are therefore usually employed for searching machine readable library files. More sophisticated means must be developed in order to maximize the usefulness of these files, since a searcher, even with book in hand, may not be able to provide the numeric key necessary to obtain the book's machine readable data.

This problem may be solved through the use of compression codes generated from author/title, or other bibliographic information. Studies

of compression codes and their performance have been reported by Ruecking (1), Kilgour (2), and the University of Chicago (3). This approach has been endorsed by the Library of Congress (4) in the RECON study.

Studies at the Library, University of Saskatchewan, were initiated with the hope of producing a compression code that would provide machine duplicate order edit in the acquisitions and cataloguing in-process file, and retrieve entries, using unverified or verified bibliographic information as input, either from a partially unverified file, such as the acquisitions and cataloguing in-process file, or from an authoritative machine readable data base, such as MARC II. In addition, the desired code would have to minimize errors in punctuation and spelling in order to achieve a high retrieval percentage, yet produce a low volume of duplicate codes for dissimilar works.

CONSTRUCTION OF THE DATA BASE

Since June, 1969, the MARC data base has been used at the Library to generate unit cards which have been used as source data for unit card masters in the cataloguing department. Approximately 300 of these were drawn at random. At the same time the original order request forms for these items were searched. Of the 300 items, 254 requisition forms were found in the manually maintained acquisitions in-process file. The LC card numbers were used to retrieve the corresponding MARC records from the Library's history MARC tape. An additional 4,128 MARC records were placed on the same tape as the 254 MARC records for which order request information existed.

The LC card numbers, order entry, title, and, if present, date of publication were keypunched from the 254 requisition forms. This bibliographic information formed the data base from which search codes were produced for the acquisitions department records.

CODE GENERATION

A computer program performed the following modifications on all input data prior to generating the actual compression codes. First, all the lower-case alphabetics were converted to upper-case alphabetics. Then all punctuation was eliminated from the title field except for periods and apostrophies within a word. A word compaction routine then eliminated periods from within abbreviations and apostrophes from within words.

The entries from the 4,382 MARC records and the 254 requisition forms were categorized according to personal name, and corporate or conference name. The first comma delimited the portion of the personal name to be used in the compression routine. Spaces, diacritics, periods, apostrophes, and hyphens were all eliminated from the personal name.

The first two codes used in the Project were labelled imaginatively Code Type 1, and Code Type 2, where Code Type 1 was a slight modification of the code developed by Frederick H. Ruecking (1). Code Type 2 was

based on a modified University of Chicago Experimental Search Code (3) incorporating ideas from some of Ruecking's studies.

CODE TYPE 1

Title Compression (16 characters)

See Ruecking (1) for the rules which were used to construct the four-character compressions.

Entry Compression (12 characters)

Three four-character compressions were used for corporate or conference names instead of Ruecking's four. One four-character compression was produced for personal names.

Date of Publication (3 characters)

If the year of publication was available, the last three digits were used, otherwise, the date was left blank.

The total length of Code Type 1 is 31 characters.

CODE TYPE 2

Title Compression (6 characters)

1) "A", "an", "the", "and", "by", "if", "in", "of", "on", "to" were deleted from the title.
2) The first word containing two consonants was located and the first two consonants appearing in the word were used for the search code.
3) Step 2 was repeated with a second and third word of the short title, whenever these were available.
4) If three words with two consonants were not available, the balance of the six characters needed for the code were supplied by those characters immediately after the last character used (except for blanks).

Entry Compression (6 characters)

a) Personal name.
   1) Only the surname, or the forename if there was no surname.
   2) If the name had six or fewer characters, the entire name was used. Otherwise, vowels were deleted from the name (working backwards on the name) until the six-character compression was formed, or the second consonant was located.
   3) If the six-character compression was not formed by step 2, then the first four characters and the last two characters were used for the six-character compression.

b) Corporate and conference entries

The rules for title compression to form the six-character code were followed.

Date of Publication (3 characters)

The last three digits of the date of publication, as in Code Type 1, were used.

In either of the codes, if the title was the main entry, a code was generated with the entry field blank.

Examples of Code Generation

Title: Factors in the Transfer of Technology.
Entries: 1) M.I.T. Conference on the Human Factor in the Transfer of Technology, Endicott House, 1966.
2) Gruber, William H.
3) Marquis, Donald George
4) Massachusetts Institute of Technology
Date of publication: 1969
Code Type 1 compressions:
1) FACTTRSFTCHNbbbbMITbCOFRHUMb969
FACTTRSFTCHNbbbbMITbCOFRHUMbbbb
2) FACTTRSFTCHNbbbbGRBRbbbbbbbb969
FACTTRSFTCHNbbbbGRBRbbbbbbbbbb
3) FACTTRSFTCHNbbbbMAQSbbbbbbbb969
FACTTRSFTCHNbbbbMAQSbbbbbbbbbb
4) FACTTRSFTCHNbbbbMATTINTTTCHN969
FACTTRSFTCHNbbbbMATTINTTTCHNbbb
Code Type 2 compressions:
1) FCTRTCMTCNHM969
FCTRTCMTCNHMbbb
2) FCTRTCGRUBER969
FCTRTCGRUBERbbb
3) FCTRTCMARQUS969
FCTRTCMARQUSbbb
4) FCTRTCMSNSTC969
FCTRTCMSNSTCbbb

PROCEDURE AND RESULTS

The two types of codes were generated from the 4,382 MARC records using publication date, short title, main entry, and added entries. Another program was written to generate codes from the acquisitions department data on cards and to write them on a separate tape using publication date if available, entry, and the first four significant words of the title and/or the words of the title up to the first punctuation mark. The two tapes containing codes were sorted in ascending code sequence, then compared. If the code generated from the acquisitions data, hereafter called the unverified code, was exactly the same as the code generated from MARC tape, hereafter called the verified code, the codes and corresponding LC card numbers were printed as a hit. The program then checked the LC card

numbers corresponding to the identical codes. If the LC card numbers were the same, a retrieval was recorded; otherwise, the matching codes were considered a false drop. The program also checked and printed duplicates existing within the verified codes and within the unverified codes.

Since the code formation programs involved string manipulation, they were written in PL/I, while the comparison program was written in COBOL. The programs were run on the IBM/360 model 50 installed at the University of Saskatchewan Computation Centre.

Table 1 and Table 2 give the results. The following is a description of the error types used in evaluating non-retrieval:

A—The unverified entry had only a remote relationship to the verified entry. No retrieval technique would have produced a match.

B—The unverified entry was misspelled.

C—The unverified title had only a remote relationship to the verified title.

D—The unverified title contained misspelled word(s).

E—Only the unverified date of publication was incorrect.

As an immediate consequence of the analysis of Tables 1 and 2, the publication date was eliminated from the codes and the comparison program rerun producing the results given in Table 3.

Table 1. Code Performance

	Retrievals	*False Drops*	*Percent Retrieval*
Code Type 1	200	0	78.74
Code Type 2	206	0	81.10

Table 2. Non-Retrieval Analysis

	Number of Non-Retrievals	
Error Type	*Code Type 1*	*Code Type 2*
A	9	9
B	10	7
C	8	8
D	7	4
E	20	20

Table 3. Code Performance

	Retrievals	*False Drops*	*Percent Retrieval*
Code Type 1A	220	0	86.61
Code Type 2A	226	0	88.98

No duplicate codes existed within the unverified code tape. From the 4,382 MARC records, 6,828 codes were produced for each of Code Type 1A

and Code Type 2A. Works having the same author and title, but different imprint, were not considered duplicates even though the program listed them as such. Seven duplicates, one triplicate and one quadruplicate occurred in Code Type 1A; and eight duplicates, two triplicates and two quadruplicates in Code Type 2A. Government publications were responsible for all but one of the duplicate codes.

CODE TYPE 2B

A graph of the number of duplicate codes vs. the number of source records was drawn for Code Type 1A and Code Type 2A (Fig. 1). As a result of this graph Code Type 2B was proposed. This code employed the same rules for construction as Code Type 2A, except that four significant words from the title and four significant words from corporate or conference entries were used to generate the compression. The total length of Code Type 2B is thus sixteen characters. Six duplicates, one triplicate and one quadruplicate appeared when the comparison program was run using Code Type 2B. Figure 1 is a graph of the result.

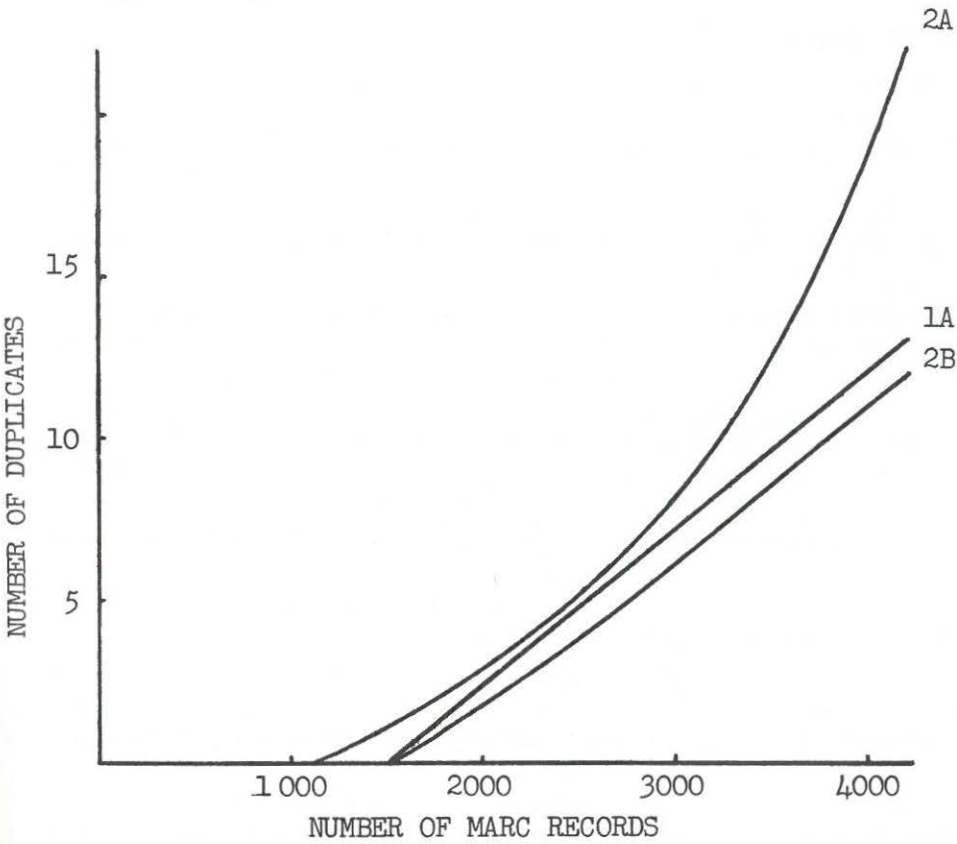

Fig. 1. Number of Duplicates vs Number of Source Records for Code Types 1A, 2A and 2B.

The performance of Code Type 2B is summarized in Tables 4 and 5.

Table 4. Code Performance

	Retrievals	*False Drops*	*Percent Retrieval*
Code Type 2B	223	0	87.80

Table 5. Non-Retrieval Analysis

Error Type	*Number of Non-Retrievals Code Type 2B*
A	9
B	10
C	8
D	4

APPLICATIONS

MARC Tapes

A MARC code tape was recently created and is being maintained at the University of Saskatchewan, as flowcharted in Figure 2. Each record on the tape consists of a compression code and an LC card number. Approximately 100,000 entry/title keys, plus series statement and SBN keys, have been created from the 65,000 records on the current MARC history tape. Figure 3 illustrates how these access points are used to provide unit card printouts.

Figure 4 shows a sample output from the matching step in Figure 3. This printout indicates the results of the search, and serves as a link between the request and the catalog card printed from the MARC tape. In the printout, entry/title requests that have found more than one LC card number do not necessarily indicate a false drop. So far, these multiple finds have resulted from the same publications appearing on MARC with different imprints. It is a simple matter to select the catalog card with the appropriate imprint. The discrepancy in Table 6 between MARC records found and titles verified is due to the above, and to multiple hits on a single record when requests for that record were submitted in more than one form, i.e. S.B.N. and author/title.

Table 6 presents a summary of the results of submitting unverified requests over a four-week period against the MARC code tape. During that time, 563 English language monographs with potential 1969 and 1970 imprints were searched. Desired MARC records were found for 184 titles, or 32.7% of these requests. The source data for the requests was supplied from title-pages and order recommendations. This data was not verified because the compression code access technique partially solves the problem of non-retrieval due to human errors in the submission.

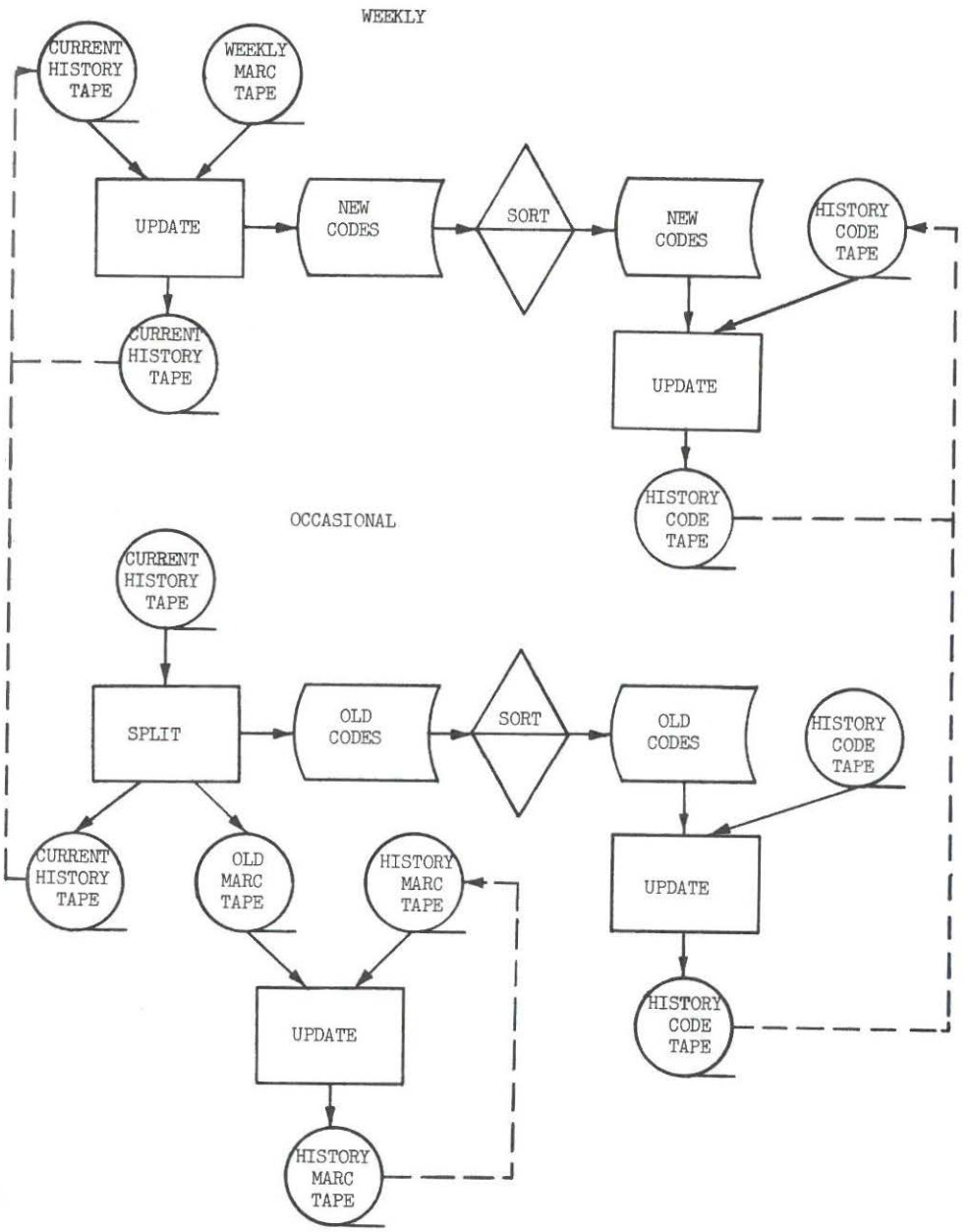

Fig. 2. MARC Tape Processing.

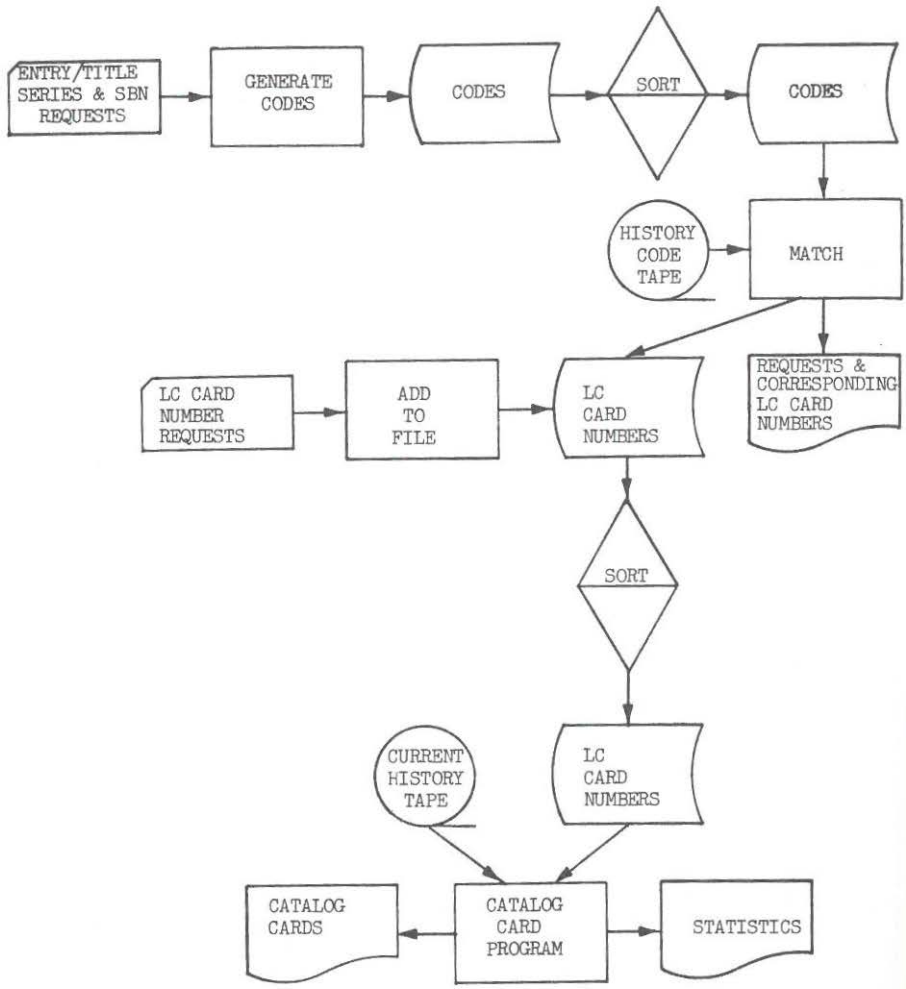

Fig. 3. MARC Access Programs.

Table 6. MARC Retrieval

	Form of Request					
	Author/ Title	*Corporate Author/Title*	*Title*	*SBN*	*Series*	*Total*
Number of Requests	546	29	130	139	11	855
MARC Records Found	173	2	11	36	18	240
Titles Verified	148	2	6	20	8	184
False Drops	0	0	2	0	6	8

LIBRARY - ENTRY/TITLE REQUESTS FOR MARC CATALOG CARDS PAGE 3

REQUEST	ID	SEARCH CODE	LC CARD NUMBER(S)
WILKINSON LATTER MIDDLE AGES IN ENGLAND 1216 1485	CAA	LTMDGSNGWILKSN	NO LC CARD NUMBERS FOUND
SHAH MODERNIZATION OF UNIVERSITY TEACHING	CAA	MDNVTCHISHAH	72902551
TRINKAUS IN OUR IMAGE AND LIKENESS	CAA	MGLKENESTRINKS	NO LC CARD NUMBERS FOUND
TRINKAUS IN OUR IMAGE AND LIKENESS HUMANITY AND DIVINITY IN ITALIAN	CAA	MGLKHMDVTRINKS	70460128
HARRISON MALAYS OF SOUTH WEST SARAWAK BEFORE MALAYSIA	CAA	MLSTWSSRHARRSN	73102868 78462841
MCCLOSKEY META ETHICS AND NORMATIVE ETHICS	CAA	MTTHNRTHMCCLKY	73437142
KEMP INDUSTRIALIZATION IN NINETEENTH CENTURY EUROPE	CAA	NDNNCNRPKEMP	74452199
KEMP INDUSTRIALIZATION IN 19TH CENTURY EUROPE	CAA	ND19CNRPKEMP	NO LC CARD NUMBERS FOUND
VALE ENGLISH GASCONY	CAA	NGGSCONYVALE	NO LC CARD NUMBERS FOUND
VALE ENGLISH GASCONY 1399 1453	CAA	NGGS1314VALE	NO LC CARD NUMBERS FOUND
HARRISON ENGLISH HOME	CAA	NGHME HARRSN	NO LC CARD NUMBERS FOUND
MYRES ANGLO SAXON POTTERY AND SETTLEMENT OF ENGLAND	CAA	NGSXPTSTMYRES	70455773
SPECK AN ANONYMOUS PARLIAMENTARY DIARY	CAA	NNPRDRY SPECK	NO LC CARD NUMBERS FOUND
SPECK AN ANONYMOUS PARLIAMENTARY DIARY 1705 6	CAA	NNPRDR17SPECK	NO LC CARD NUMBERS FOUND
KING ANTARCTIC	CAA	NTARCTICKING	79435704
JOHNSON INTRODUCTION TO SOVIET LEGAL SYSTEM	CAA	NTSVLGSYJOHNSN	78455132
TURNER POLITICS AND MULTI NATIONAL COMPANY	CAA	PLMLNTCMTURNER	NO LC CARD NUMBERS FOUND
SKELTON POEMS	CAA	PMS SKELTN	NO LC CARD NUMBERS FOUND

Fig. 4. Results of Entry/Title Search for MARC Unit Card Printouts.

Manual searching of the NUC catalogs was employed to verify titles that could not be located on the MARC tape. Ten titles were found with MARC notations after failing to be retrieved by compression code matching. Type A and type C errors were primarily responsible for this non-retrieval. However, two of these titles could not be retrieved from the MARC tape following manual verification, since the verified entries in the NUC preceded their counterparts on the MARC tape. Thus the performance of the compression codes can be evaluated as 184 of a possible 192 hits, or a 95.8% retrieval rate.

During the four-week period the keypunchers formulated 52% more entry/title requests than there were titles for verification. This is due mainly to the need for submitting more than one author/title request whenever the portion of the title which comprises the short title is in doubt, since the code is formulated from the short title only. Additional experience should decrease the number of redundant requests.

Only 8 false drops have been received in the above submissions. Retrieval of series entries is likely to engender the greatest number of false drops because series statements are treated as titles in the code generation procedure.

Acquisitions and Cataloguing

During the past two years, the Technical Services Department at the

Library and the Computation Centre have designed, and are currently testing TESA I (Technical Services Automation—Phase I), an automated acquisition and cataloguing system (5), the primary objectives of which were to pursue a total library system concept and to provide for conversion from a batch system to an on-line operation when sufficient computer facilities become available.

At the same time that work proceeded towards these objectives, status codes and receiving reports were employed as used in Washington State University's LOLA system (6) and (7). However, MARC tapes and compression codes comprise an integral part of the system. If a MARC record can be located before an order is entered, a tremendous amount of keying is saved. One 64-character in-process transaction will supply the ordering information and transfer the bibliographic data from the current MARC history tape to the direct access acquisitions and cataloguing in-process file (IBM's Basic Direct Access Method). Minimal cataloguing updates are necessary before catalog card sets can be produced. Entry/title access ensures that only a small percentage of needed MARC records will slip through TESA I's fingers at order initiation time.

Another code application as illustrated in Figure 5 will exploit the fact that the same code construction rules are used in the MARC system as in TESA I. Items requiring bibliographic information will be flagged in the in-process file. When a new MARC tape arrives, the in-process code file (IBM's Index Sequential Access Method) will be automatically matched with the MARC codes created from the new weekly tape. A sample printout from these matches is provided in Figure 6. After verifying which MARC records are needed, MARC bibliographic information will be transferred to the appropriate in-process records.

Each record in the direct-access ISAM compression code file consists of a compression code (or SBN or LC card number) and the key (purchase order number) to the corresponding in-process record. A threaded list structure exists within the in-process file to handle the possibility of one code accessing several items. Thus an in-process record may be directly accessed by entry/title, series statement, SBN, LC card number or purchase order number.

A fast edit routine built into the direct-access write detects whether or not the compression code about to be written is a duplicate of a code already in the file. If the code is unique the code record is written on disc and a single item list is created within the corresponding in-process record. If the code is not unique, the code record cannot be written. In this case the list structure for the code is updated to include the key of the in-process record being added. A message, together with the purchase order numbers of items which may be duplicates, is printed to warn the acquisitions staff that a potential duplicate is being added to the in-process file. Traditional duplicate checking of in-process items thus becomes an exception.

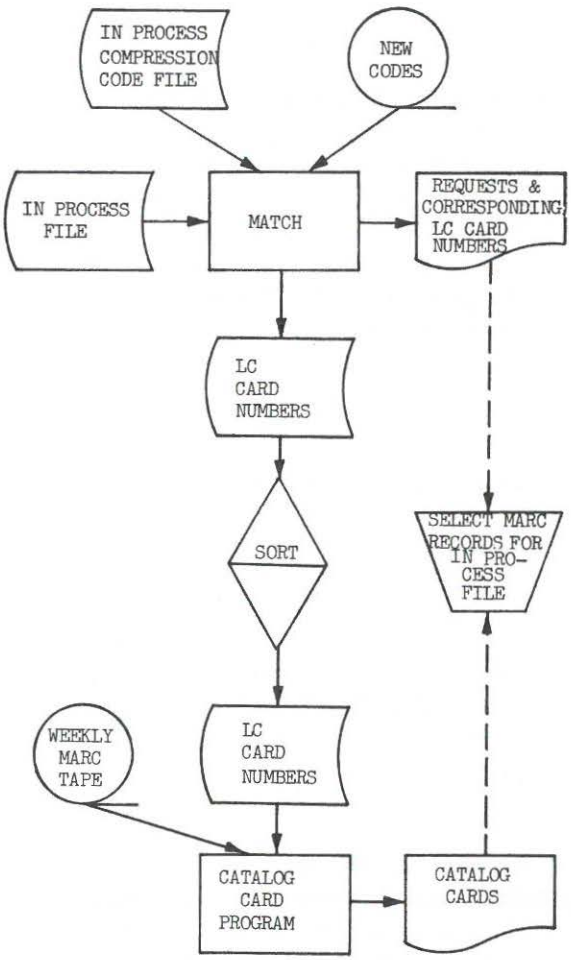

Fig. 5. Search of Weekly MARC Tape for Records Needed in the In-Process File.

```
     LIBRARY - IN PROCESS ITEMS FOR WHICH BIBLIOGRAPHIC INFORMATION MAY EXIST ON THE NEWLY ARRIVED MARC TAPE    PAGE  1
     ------------------------------------------------------------------------------------------------------------------

ITEM NUMBER: 100152C0 MAY CORRESPOND TO LC CARD NUMBER:    77111204
AUTHOR: KENYON, ROBERT LLOYD,
TITLE:  KENYON'S GOLD COINS OF ENGLAND.

ITEM NUMBER: 100153A0 MAY CORRESPOND TO LC CARD NUMBER:    77123648
AUTHOR: SHAPO, MARSHALL S.
TITLE:  PRODUCTS AND THE CONSUMER: DEFECTIVE AND DANGEROUS PRODUCTS,

ITEM NUMBER: 100150G0 MAY CORRESPOND TO LC CARD NUMBER:    78117811
AUTHOR: HUGHES, ERNEST RICHARD,
TITLE:  CHINA, BODY & SOUL:

ITEM NUMBER: 100151E0 MAY CORRESPOND TO LC CARD NUMBER:    79906166
AUTHOR: SEMINAR ON PANCHAYATI RAJ, PLANNING AND DEMOCRACY, JAIPUR, 1964.
TITLE:  PANCHAYATI RAJ, PLANNING AND DEMOCRACY; <PROCEEDINGS>.
```

Fig. 6. In-Process Items for Which Bibliographic Information May Exist on the Newly Arrived MARC Tape.

Remote Access to MARC

An experiment was conducted in which entry/title requests were submitted from the IBM S360/40 computer at the University of Saskatchewan, Regina Campus, Computer Centre, over a communication link to the Saskatoon Campus IBM S360/50 computer. The MARC access program was read into the Regina computer, sent to Saskatoon's computer, spooled in the Saskatoon job queue and executed; then the results of the search were sent to Regina to be printed. The entire process took approximately the same time as if the program had actually been executed in Regina. No data transmission errors were encountered in transmitting either the requests or the retrieved MARC unit cards over this 150-mile communication link.

CONCLUSION

There is an inverse relationship between retrieval performance and number of duplicate codes produced. A high retrieval code such as Code Type 2A results in more duplicates than a code such as Ruecking's, which has a slightly lower retrieval performance.

Code Type 2B fulfills the requirements for a code short in length and easy to construct that produces a low number of duplicates and has high retrieval capability. For an index to a library holdings file, or to a national data base, a code such as Ruecking's, with four or more significant words from title and corporate or conference entries, and with different rules for personal author compression, would perhaps be suitable.

ACKNOWLEDGMENTS

The authors thank the Library staff for their assistance in the study. They are also grateful to the Library and Computation Centre administrations, in particular, D. C. Appelt, G. C. Burgis, and N. E. Glassel for the allotment of computer time and their encouragement.

REFERENCES

1. Ruecking, Frederick H. Jr.: "Bibliographic Retrieval from Bibliographic Input; The Hypothesis and Construction of a Test," *Journal of Library Automation*, 1 (December 1968), 227-238.
2. Kilgour, Frederick G.: "Retrieval of Single Entries from a Computerized Library Catalog." In American Society for Information Science, Annual Meeting, Columbus, O., 20-24 Oct. 1968: *Proceedings*, 5 (1968), 133-136.
3. "University of Chicago Experimental Search Code." In Avram, Henriette D.; Knapp, John F.; Rather, Lucia J.: *The MARC II Format: A Communications Format for Bibliographic Data* (Washington, D.C., Library of Congress, 1968), pp. 129-131.
4. "Computer Requirements for a National Bibliographic Service." In *RECON Working Task Force: Conversion of Retrospective Records to Machine-Readable Form* (Washington, D.C.: Library of Congress, 1969), pp. 183-226.

5. Newman, W. L.: *Technical Services Automation—Phase I Acquisitions and Cataloguing* (Computation Centre, University of Saskatchewan, Saskatoon, November, 1969), mimeographed.
6. Burgess, T.; Ames, L.: *LOLA; Library On-Line Acquisitions Sub-System* (Pullman, Wash.: Washington State University Library, July, 1968).
7. Mitchell, Patrick C.: *LOLA, Library On-Line Acquisitions Sub-System,* (Washington State University, June, 1969) unpublished.