BIBLIOGRAPHIC RETRIEVAL FROM BIBLIOGRAPHIC INPUT; THE HYPOTHESIS AND CONSTRUCTION OF A TEST

Frederick H. RUECKING, Jr.: Head, Data Processing Division,
The Fondren Library, Rice University, Houston, Texas

A study of problems associated with bibliographic retrieval using unverified input data supplied by requesters. A code derived from compression of title and author information to four, four-character abbreviations each was used for retrieval tests on an IBM 1401 computer. Retrieval accuracy was 98.67%.

Current acquisitions systems which utilize computer processing have been oriented toward handling the order request only after it has been manually verified. Systems, such as that of Texas A & I University (1), have proven useful in reducing certain clerical routines and in handling fund accounting (2). Lack of a larger bibliographic data base and lack of adequate computer time have prevented many libraries from studying more sophisticated acquisitions systems.

At the time the MARC Pilot Project (3) was started, the Fondren Library at Rice University did not have operating computer applications in acquisitions, serials, or cataloging. The University administration and the Research Computation Center provided sufficient access to the IBM 7040 to permit the study of problems associated with bibliographic retrieval using input data which has varying accuracy.

In 1966, Richmond expressed the concern of many librarians about the lack of specific statements describing the techniques by which on-line retrieval could be accomplished without complicating the problems presented by the current card catalog (4). She had previously described some of the problems created by the kind and quality of data being utilized as references by library users (5).

An examination of the pertinent literature indicates that most of the current work in retrieval, while related to problems of bibliographic retrieval, does not offer much assistance when the input data is suspect (6, 7,8). Tainiter and Toyoda, for example, have described different techniques of addressing storage using known input data (9,10).

One of the best-known retrieval systems is that of the Chemical Abstracts Service, which provides a fairly sophisticated title-scan of journal articles with a surprising degree of flexibility in the logic and term structure used as input. Comparable systems are used by the Defense Documentation Center, Medlars Centers, and NASA Technology Centers. These systems have one specific feature in common: a high level of accuracy in the input data.

USER-SUPPLIED BIBLIOGRAPHIC DATA

The reliability of bibliographic data supplied to university libraries from faculty and students has long been questioned (5). Any search system which accepts such data must be designed 1) to increase the level of confidence through machine-generated search structures and variable threshholds and 2) to reduce the dependence upon spelling accuracy, punctuation, spacing and word order.

The initial task of formulating an approach to this problem is to determine the type, quality, and quantity of data generally supplied by a user. To derive a controlled set of data for this purpose, the Acquisition Department of the Fondren Library provided Xerox copies of all English language requests dated 1965 or later and a random sample of 295 requests was drawn from that file of 5000 items.

This random sample was compared to the manually-verified, original order-requests to determine 1) the frequency with which data was supplied by the requestor and 2) the accuracy of the provided information. Results of this study are given in Table 1.

Table 1. Level of Confidence in the Input Data

Data Elements	Times Given	Times Correct	Accuracy	Level of Confidence
Edition	295	294	99.6	99.6
Title	295	292	99.0	99.0
Author	290	264	91.0	82.7
Publish.	268	218	81.3	73.9
Date	265	215	81.1	72.8

The results suggest that edition can have great significance when specified and should be used as strong supporting evidence for retrieval. It should not necessarily be a restrictive element because of the low-order magnitude of actual specification, which was five times in the sample. (Unstated editions were considered as first editions, and correct.)

Title is the most significant and most reliable element. As Richmond indicates, use of the entire title for searching would present distinct problems for retrieval systems (4). Consequently, an abbreviated version of the title must be derived from the input data which will reduce the impact and significance of the problems described by Richmond (5).

THE HYPOTHESIS

It is hypothecated that retrieval of correct bibliographic entries can be obtained from unverified, user-supplied, input data through the use of a code derived from the compression of author and title information supplied by the user. It is assumed that a similar code is provided for all entries of the data base using the same compression rules for main and added entry, title and added title information.

It is further hypothecated that use of weighting factors for individual segments of the code will provide accurate retrieval in those cases when exact matching does not occur. Before the retrieval methodology can be described, it is necessary to outline the compression technique to be used with author and title words.

TITLE COMPRESSION

To gain some understanding of the problems to be faced in compressing title information, a random sample of 500 titles was drawn from the first half of the initial MARC I reel (about 4800 titles). Each of these titles was analyzed for significant words and tabulations were made on word strings and word frequencies. The following words were considered as non-significant: *a, an, and, by, if, in, of, on, the, to.*

The tabulated data, shown in Table 2, contain some surprising attributes. Approximately 90% of the titles contain less than five significant words, which suggests that four significant words will be adequate to match on title.

Table 2. Significant Word Strings in Titles

	Length of Word String					
	1	2	3	4	5+	Total
Number of titles	42	151	179	76	52	500
Percentage	8.4	30.2	35.8	15.2	10.4	100.0
Cumulative Percentage	8.4	38.6	74.4	89.6	100.0	

Letting n stand for the corpus of words available for title use, the random chance of duplicating any specific word in another title can be stated as $\frac{1}{n}$. When a string of words is considered, the chance of randomly selecting the same word string may be considered as $\frac{1}{n^a}$, where 'a' is the number of words in the string.

Certain words are used more frequently than others, and the occurrence of such words in a given string reduces the uniqueness of that string. The curve displayed in Figure 1 shows the frequency distribution of words in the sample. The mean frequency of words in the title-sample is 1.33.

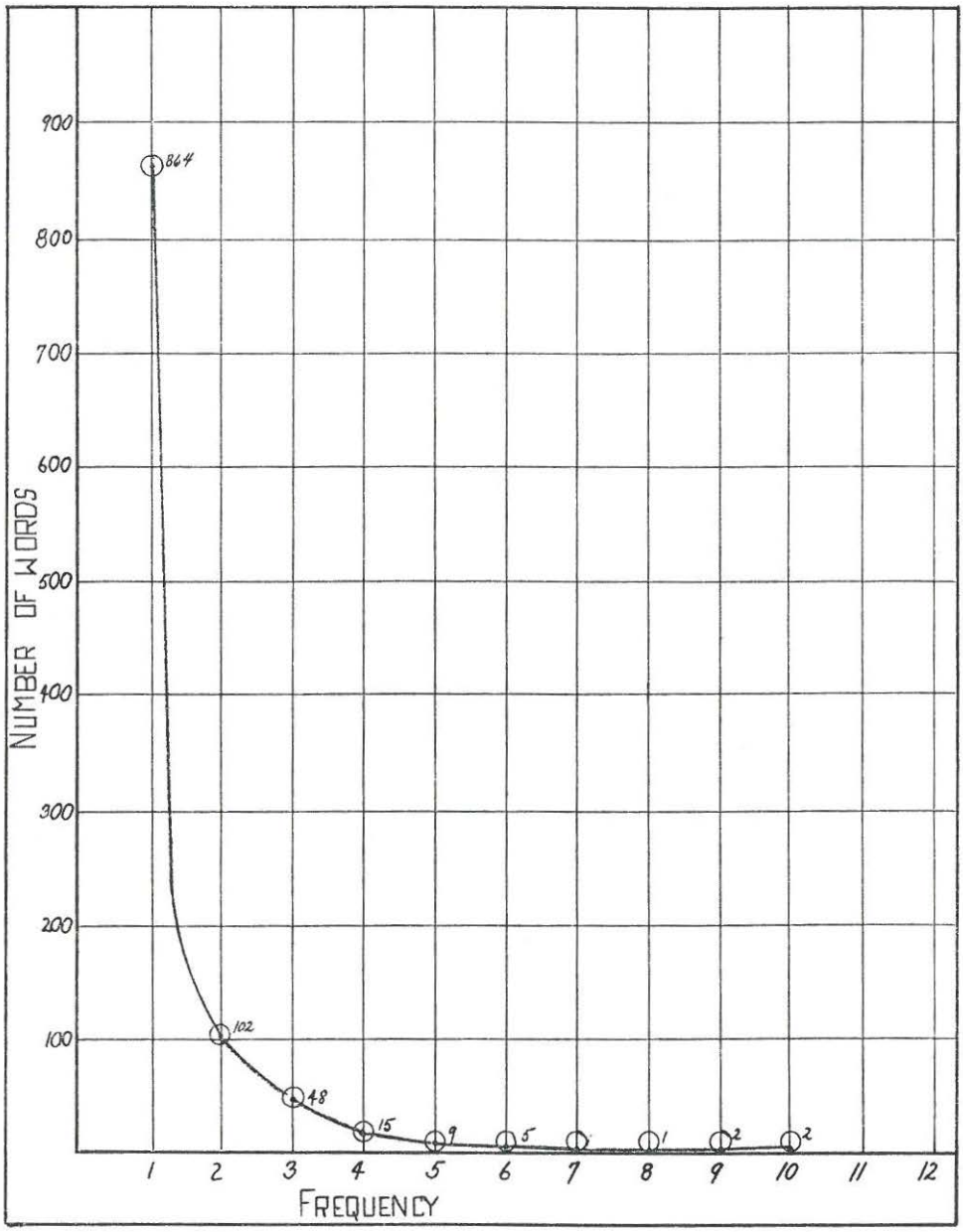

Fig. 1. Frequency Distribution of Words in Sample.

Therefore, the chance of selecting an identical word string can be more accurately expressed as:

$$\frac{1.33^a}{n^a}$$

An examination of word lengths, as shown in Table 3, shows that 95% of the significant title words contain less than ten characters. An examination of the word list revealed that some 70% of the title words contain inflections and/or suffixes. If these suffixes and inflections are removed, approximately 43% of the remaining word stems contain less than five characters and 59% contain less than six.

Table 3. Distribution of Character Length and Stem Length

Length in Characters	Total Words	Different Words	Percent	Stems	Percent
1	7	5	0.5	5	0.8
2	25	14	1.3	14	2.3
3	87	48	4.6	48	7.9
4	172	117	11.1	196	32.3
5	229	163	15.5	92	15.2
6	198	153	14.5	94	15.5
7	202	159	15.3	64	10.6
8	158	122	11.6	45	7.4
9	121	102	9.7	15	2.5
10	84	69	6.6	8	1.3
11	54	48	4.6	7	1.2
12	38	28	2.7	2	0.3
13	14	12	1.1	2	0.3
14	6	4	0.4	0	0.0
15	3	3	0.3	0	0.0
16	2	2	0.2	0	0.0
Summary	1400	1049		592	

The reduction of word length does affect the uniqueness of the individual word, merging distinct words into common word stems at a mean rate of 2.5 to 1.0. In Table 3 the difference between 1049 words and 592 stems reflects the reduction of similar words into a common stem; for example: America, American, Americans, Americanism, etc., into Amer. Thus, the uniqueness of a string of title words is reduced to the following chance of duplication:

$$\frac{(2.5 \times 1.33)^a}{n^a} \text{ or } \frac{3.3^a}{n^a}$$

An analysis of consonant strings made by Dolby and Resnikoff provides frequencies of initial and terminal consonant strings occurring in 7000 common English words (with suffixes and inflections removed) (11,12, 13). These frequency lists clearly show that the terminal string of consonants has considerable information-carrying potential in terms of word identification. The starting string also carries information potential, but significantly less than the terminal string. By combining the initial and terminal strings, it is possible to generate an abbreviation which has adequate uniqueness and reduces the influence of spelling.

The high percentage of four-character word stems and the fact that the maximum terminal string contains four consonants suggest the use of a four-character abbreviation. To compress a title word into four characters, it is necessary to specify a set of rules. The first rule will be to delete all suffixes and inflections which terminate a title word. The second rule will be to delete vowels from the stem until a consonant is located or the four-character stem is produced. The suffixes and inflections deleted in this procedure are contained in Table 4. When the stem contains more than four characters, the third compression rule states that the four-character field is filled with the terminal-consonant string and remaining positions are filled from the initial - character string.

Table 4. Deleted Suffixes and Inflections

-ic	-ive	-in	-et
-ed	-ative	-ain	-est
-aged	-ize	-on	-ant
-oid	-ing	-ion	-ent
-ance	-og	-ation	-ient
-ence	-log	-ship	-ment
-ide	-olog	-er	-ist
-age	-ish	-or	-y
-able	-al	-s	-ency
-ible	-ial	-es	-ogy
-ite	-ful	-ies	-ology
-ine	-ism	-ives	-ly
-ure	-um	-ess	-ry
-ise	-ium	-us	-ary
-ose	-an	-ous	-ory
-ate	-ian	-ious	-ity
-ite			

The relative uniqueness of the generated abbreviation can be calculated using the data supplied by Dolby and Resnikoff. For example, Carter and Bonk's *Building Library Collections* would be abbreviated - BULD, LIBR,COCT. The random chance of duplicating any abbreviation can be stated as consisting of the product of the random chance of duplicating the initial string and the random chance of duplicating the terminal string:

$$\frac{f_i}{n_i} \times \frac{f_t}{n_t} \times 3.3^2$$

The frequencies listed by Dolby and Resnikoff may be substituted in the above equation producing the following chances for duplication:

$$\frac{324}{6800} \times \frac{63}{6800} \times 10.89 = \frac{1}{208} \quad \text{for BULD}$$

$$\frac{288}{6800} \times \frac{1}{6800} \times 10.89 = \frac{1}{14745} \quad \text{for LIBR}$$

$$\frac{277}{6800} \times \frac{16}{6800} \times 10.89 = \frac{1}{1041} \quad \text{for COCT}$$

The random chance of duplicating this string of three abbreviations can be calculated by multiplying the individual calculations, which yields the random chance of 1 in 32×10^8. This high uniqueness declines rapidly when the title contains less than three significant words and contains high frequency words, such as the title *Collected Works*, for which the same uniqueness calculation produces the random chance of 1 in 44×10^4.

To increase the level of uniqueness on short titles, like *Collected Works*, it becomes necessary to provide supporting data to the title information. It is clear that the supporting data must come from supplied author text.

AUTHOR COMPRESSION

The same compression algorithms can be used for both personal and corporate names with some modifications. The frequent substitution of "conference" for "congress" and "symposia" for "symposium" suggests that meeting names should be considered as a secondary sub-set of non-significant words. Names of organizational divisions, such as bureau, department, ministry, and office, can be considered as part of the same sub-set.

The rules which govern the deletion of inflections, suffixes and vowels can be used for corporate names, but personal author names must be carried into the compression routine without modification. Only the last name of an author would be compressed into a code.

CONSTRUCTING THE TEST

Four, four-character abbreviations are allowed for title compression and four for author. Rather than use a 32-character fixed field for these codes, the lengths of the input and main-base codes are variable, with leading control digits to specify the individual code sizes for the title and author segments.

Provision is made for the inclusion of date, publisher and/or edition in the search-code structure although these were not implemented in the test performed.

At the time the input data is read, the existence of title, author, edition, publisher and date is indicated by the setting of indicators which control the matching mask and which, in part, control the specification of the retrieve threshhold. The title indicator specifies the number of compressed words in the supplied title which must be matched by the base code.

A simple algorithm is used to calculate the threshhold values given in columns two through four of Table 5. Columns five through seven are obtained by adding two to the calculated threshholds. Each agreement within the mask adds to a retrieve counter the values indicated in the last five columns of Table 5, the values of X and Y being the number of matching code words in the title and author segments respectively.

CONDUCTING THE TEST

As mentioned above, the initial tests of the retrieve were based upon title and author matching exclusively and required three runs on the Fondren Library's 1401 computer. The first loaded 2874 original order-requests, generated a search code utilizing the rules specified in this paper and created an input tape. The second run extracted title and author data from the MARC I data base, created multiple search codes for title, main entry, added title and added entry. Both tapes were sorted into ascending search-code sequence.

The final run was the search program which attempted to match input codes with the MARC I base codes. When there was agreement based on relationship of threshhold and retrieve counter, the printer displayed threshhold, short author and short title on one line, and retrieve value, input author and title on the next line as illustrated in Figure 2. The printed results were compared to validate the accuracy of the retrieve. This comparison was cross-checked against the results of the acquisition department's manual procedures.

The search program also provided for an attempt to match titles on the basis of a rearrangement of title words. In such attempts the retrieve threshhold was raised.

ANALYSIS OF RESULTS

The raw data obtained from this experimental run are shown in Table 6. Of the 2874 items represented in the input file, 48.4%, or 1392, were actually found to exist in the data base. Of those actually present 90.4%, or 1200, were extracted with an overall accuracy of 98.67%.

An examination of the sixteen false drops revealed several omissions in the compression routines for the input data and for the data base. One of the more significant omissions was failing to compensate for multi-character abbreviations, particularly 'ST.' and 'STE.' for 'Saint.' A subroutine for acceptance of such abbreviations added to the search-code generating program would increase the retrieve accuracy to 99%.

Table 5. Values for Variable Threshhold

Data Given	Threshhold Values						Agreement Values				
	Full-Code Test			Individual Code Test			Title	Author	Edition	Publish.	Date
TAEPD	3 or 4	2	1	3 or 4	2	1					
XY111	12	8+2Y	4+2Y	14	10+2Y	6+2Y	4X	2Y	3	2	1
XY110	12	8+2Y	4+2Y	14	10+2Y	6+2Y	4X	2Y	3	2	1
XY101	12	8+2Y	4+2Y	14	10+2Y	6+2Y	4X	2Y	3	2	1
XY100	12	8+2Y	4+2Y	14	10+2Y	6+2Y	4X	2Y	3	2	1
XY011	12	8+2Y	4+2Y	14	10+2Y	6+2Y	4X	2Y	3	2	1
XY010	12	8+2Y	4+2Y	14	10+2Y	6+2Y	4X	2Y	3	2	1
XY001	12	8+2Y	4+2Y	14	10+2Y	6+2Y	4X	2Y	3	2	1
XY000	12	8+2Y	4+2Y	14	18+2Y	6+2Y	4X	2Y	3	2	1
XO111	12	11	7	13	12	7	4X	2Y	3	2	1
X0110	12	11	7	13	12	7	4X	2Y	3	2	1
X0101	12	11	7	13	12	7	4X	2Y	3	2	1
X0100	12	11	7	13	11	7	4X	2Y	3	2	1
X0011	12	10	6	13	11	7	4X	2Y	3	2	1
X0010	12	10	6	13	Not permitted		4X	2Y	3	2	1
X0001	12	9	5	13	Not permitted		4X	2Y	3	2	1
X000	12	Not permitted		Not permitted							

```
10 AMERAMBRHCHS
10 AMERAMBRHCHS

06 AMERBOLL
06 AMERBOLL

12 AMERBUSQSHOWZIEN
12 AMERBUSQSHOWZEIN

12 AMERCNTRCAMPBRTH
12 AMERCNTRCAMP

12 AMERJEWSISRLISCS
14 AMERJEWSISRLISCSHALDR

12 AMEROCCPSTCTBLAU
14 AMEROCCPSTCTBLAU

12 AMEROCCPSTCTDUNN
12 AMEROCCPSTCTBLAU

12 AMERPARTSYSMCHRS
14 AMERPARTSYSMCHRS

10 AMERPREDWARN
10 AMERPREDWARN

10 AMERSCHKBLCK
10 AMERSCHKBLCK

10 AMERSCHOSEXN
10 AMERSCHOSEXNPATCCAYO

12 AMERSPACEXPRSHEN
14 AMERSPACEXPRSHEN

12 AMERTHETTODADOWR
14 AMERTHETTODADOWR

12 AMERTHTRAS  SEENBRWN
16 AMERTHTRAS  SEENMOSSMONSJ

12 AMERTHTRAS  SEENMOSS
18 AMERTHTRAS  SEENMOSSMONSJ

12 ANAZPHPHARGUMCGL
12 ANAZPHPHARGUMCGFJAN PHIP

12 ANCIHUNTFAR WESTPOUD
18 ANCIHUNTFAR WESTPOUD
```

Fig. 2. Sample of Retrieved Citations.

```
HEINRICHS, WALDO H.          AMERICAN AMBASSADOR  JOSEPH C. GR
HEINRICHS‡                   AMERICAN AMBASSADOR‡

BOSWELL, CHARLES.            THE AMERICA  THE STORY OF THE WORL
BOSWELL‡                     THE AMERICA. THE STORY OF THE WORLD

ZIEDMAN, IRVING.             THE AMERICAN BURLESQUE SHOW.
ZEIDMAN‡                     THE AMERICAN BURLESQUE SHOW‡

BOSWORTH, ALLAN R.           AMERICA-S CONCENTRATION CAMPS  BY
CLAY, C. T.‡                 AMERICA-S CONCENTRATION CAMPS‡

ISAACS, HAROLD ROBERT,       AMERICAN JEWS IN ISRAEL  BY HARO
ISAACS, HAROLD R.‡           AMERICAN JEWS IN ISRAEL‡

BLAU, PETER MICHAEL.         THE AMERICAN OCCJPATIONAL STRUCTUR
BLAU‡                        THE AMERICAN OCCUPATIONAL STRUCTURE

DUNCAN, OTIS DUDLEY,  JO     THE AMERICAN OCCUPATIONAL STRUCTUR
BLAU‡                        THE AMERICAN OCCUPATIONAL STRUCTURE

CHAMBERS, WILLIAM NISBET     THE AMERICAN PARTY SYSTEMS  STAGES
CHAMBERS‡                    THE AMERICAN PARTY SYSTEMS. STAGES

WARREN, SIDNEY,  1916 -      THE AMERICAN PRESIDENT  READINGS
WARREN‡                      THE AMERICAN PRESIDENT‡

BLACK, HILLEL.               THE AMERICAN SCHOOLBOOK.
BLACK‡                       THE AMERICAN SCHOOLBOOK‡

SEXTON, PATRICIA CAYO.       THE AMERICAN SCHOOL  A SOCIOLOGIC
SEXTON, PATRICIA CAYO‡       THE AMERICAN SCHOOL. A SOCIOLOGICAL

SHELTON, WILLIAM ROY.        AMERICAN SPACE EXPLORATION  THE F
SHELTON‡                     AMERICAN SPACE EXPLORATION. THE FIR

DOWNER, ALAN SEYMOUR,        THE AMERICAN THEATER TODAY,  EDITE
DOWNER‡                      THE AMERICAN THEATER TODAY‡

BROWN, JOHN MASON,  1900     THE AMERICAN THEATRE AS SEEN BY IT
MOSES, MONTROSE J.‡          THE AMERICAN THEATRE AS SEEN BY ITS

MOSES, MONTROSE JONAS,       THE AMERICAN THEATRE AS SEEN BY IT
MOSES, MONTROSE J.‡          THE AMERICAN THEATRE AS SEEN BY ITS

MCGREAL, IAN PHILIP,  19     ANALYZING PHILOSOPHICAL ARGUMENTS
MCGREAF, JAN PHILLIP‡        ANALYZING PHILOSOPHICAL ARGUMENTS.

POURADE, RICHARD F.          ANCIENT HUNTERS OF THE FAR WEST,
POURADE‡                     ANCIENT HUNTERS OF THE FAR WEST‡
```

Table 6. Table of Results

Retrieve Values	Total Hits	Correct Hits	False Hits	Percentage Correct
6	14	14	0	100
8	0	0	0	0
10	311	311	0	100
12	264	248	16	93.3
14	232	232	0	100
16	118	118	0	100
18	260	260	0	100
20	1	1	0	100
Totals	1200	1184	16	98.7

Table 7. Distribution of Errors

	Title Errors		Author Errors				
No. of Codes	Title Error	Spelling	Author Lacking	Author Error	Spelling	Other	Total
1	2	3	10	12	27	4	58
2	2	6	17	26	60	23	134
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
Total	4	9	27	38	87	27	192

The occurrence of titles with the words "selected" or "collected," etc., produced additional false drop when the title word string exceeded two words. A modification to the search program to raise the threshhold when the input data contain codes such as 'SECT,' 'COCT' would increase the retrieve accuracy to 99.17%

The presence of personal names in titles, such as 'Charles Evans Hughes' and 'Franklin Delano Roosevelt' caused seven additional false drops. At present it seems unlikely that a simple method to prevent them can be included.

CONCLUSION

The experimental results indicate that the hypothesis suggested is valid. Use of multiple codes for added entry, added title in addition to the main entry, and main title data are clearly necessary. Approximately 10% of the correctly retrieved items were produced by the existence of an added entry code.

The influence of spelling accuracy was lessened by use of a compression technique. An inspection of extracted titles revealed the existence of 43 spelling errors which did not affect retrieval. Thus, the search code reduced the significance of spelling by some 30%.

Utilizing table search followed by table look-up and linking random-

access addresses, should enable the search code approach to bibliographic retrieval to provide rapid, direct access to the title sought.

ACKNOWLEDGMENT

This study was supported in part by National Science Foundation grants GN-758 and GU-1153 and by the Regional Information and Communication Exchange. The assistance of the Acquisitions Department staff, the Research Computation Center staff and the staff of the Fondren Library's Data Processing Division is gratefully acknowledged.

REFERENCES

1. Morris, Ned C.: "Computer Based Acquisitions System at Texas A & I University," *Journal of Library Automation,* 1 (March 1968), 1-12.
2. Wedgeworth, Robert: "Brown University Library Fund Accounting System," *Journal of Library Automation,* 1 (March 1968), 51-65.
3. U. S. Library of Congress: *Project MARC, an Experiment in Automating Library of Congress Catalog Data* (Washington: 1967).
4. Richmond, Phyllis A.: "Note on Updating and Searching Computerized Catalogs," *Library Resources and Technical Services,* 10 (Spring 1966), 155-160.
5. Richmond, Phyllis A.: "Source Retrieval," *Physics Today,* 18 (April 1965), 46-48.
6. Atherton, P.; Yorich, J. C.: *Three Experiments with Citation Indexing and Bibliographic Coupling of Physics Literature* (New York, American Institute of Physics, 1962).
7. International Business Machines Corporation: *Reference Manual, Index Organization for Information Retrieval* (IBM, 1961).
8. International Business Machines Corporation: *A Unique Computable Name Code for Alphabetic Account Numbering* (White Plains, N.Y.: IBM, 1960).
9. Tainiter, M.: "Addressing Random-Access Storage with Multiple Bucket Capacities," *Association for Computing Machinery Journal,* 10 (July 1963), 307-315.
10. Toyoda, Junichi; Tazuka, Yoshikazu; Kasahara, Yoshiro: "Analysis of the Address Assignment Problems for Clustered Keys," *Association for Computing Machinery Journal,* 13 (October 1966), 526-532.
11. Dolby, James L.; Resnikoff, Howard L.: "On the Structure of Written English Words," *Language,* 40 (Apr-June 1964), 167-196.
12. Resnikoff, Howard L.; Dolby, James L.: "The Nature of Affixing in Written English, Part I," *Mechanical Translation,* 8 (March 1965), 84-89.
13. Resnikoff, Howard L.; Dolby, James L.: "The Nature of Affixing in Written English, Part II," *Mechanical Translation,* 9 (June 1966), 23-33.