TECHNIQUES FOR MACHINE-ASSISTED CATALOGING OF BOOKS

B. A. Lipetz, D. E. Sparks, and P. J. Fasana

Research Division
Itek Corporation
Lexington 73, Massachusetts

Special Report No. 3 (revised)
Contract AF 19(604)8438

October 1962

Prepared for

ELECTRONIC RESEARCH DIRECTORATE
AIR FORCE CAMBRIDGE RESEARCH LABORATORIES (AFCRL)
AIR FORCE OFFICE OF AEROSPACE RESEARCH (AFOAR)
UNITED STATES AIR FORCE
BEDFORD, MASSACHUSETTS

TABLE OF CONTENTS

1. INTRODUCTION

This report summarizes techniques for the automatic production of catalog card sets from machine-interpretable records. The automatic techniques were developed to facilitate increased book cataloging without increasing manpower requirements at the Air Force Cambridge Research Laboratories (AFCRL) Research Library; the machine-interpretable records format was designed to be compatible with the requirements of other Air Force libraries and information systems as well as the AFCRL Research Library.

Appendix A described the AFCRL Research Library's process flow. The recommended cataloging procedures described below are designed to be compatible with this, yet take advantage of labor-saving machine techniques. The recommended cataloging procedures preserve all current controls on the content and accuracy of the AFCRL Research Library catalog cards, and produce sets of catalog cards which are similar to those filed in the catalog to date.

2.CATALOGING ACTIVITIES

Blank cataloging worksheets are inserted into uncataloged books. Any information pertinent to the cataloging process discovered during earlier processing routines is noted on the worksheet.

Each book is systematically inspected by the cataloger for the purpose of developing cataloging data. Four types of activities are distinguished:

1. Development of call number: a unique call number, comprising a class number and an author number, is created and assigned to each book. Class numbers are taken from printed schedules; author numbers are generated by applying special frequency scales to the main entry for the book.

2. Subject analysis: Subject headings which best describe the subject content of the book are selected from a standard list of subject headings and listed on the worksheet.

3. Descriptive cataloging: Descriptive cataloging data is recorded on the worksheet in a form compatible with accepted authorities.

4. Completing the worksheet: A special worksheet is used to record cataloging information. It is designed to mechanically sequence and position blocks of bibliographical data into an encoded format. Every space on the worksheet is keyed to this encoded format and must be completed.

Upon completion, the worksheet is again inserted in the book, and sent to a tape-typewriter operator for input keying.

3. CATALOG CARD FORMAT DESCRIPTION

The content and structure of the conventional unit catalog card used in libraries is highly standardized. A unit catalog card is a bibliographic record which uniquely describes a bibliographic item (i.e., a book, serial, document, etc.); it is made up of strictly defined types of information, recorded in strict sequence in a fixed format.

The unit catalog card is the basic bibliographic record from which a series of secondary control records are made. The most important secondary record created is the catalog card set, which consists of a number of 3-by-5 inch cards which reproduce all of the information contained on a unit card, plus a unique tracing printed out at the top.

Tracings are file headings which are generated by extracting certain types of information from pre-defined statements contained in the body of the unit catalog card and printing them out at the head of every card (excepting the unit card) in the catalog card set.

4. MACHINE-INTERPRETABLE NATURAL FORMAT

Before any store of information can be automated, two basic machine requirements must be satisfied. Information must be machine-accessible and machine-interpretable. Making information machine-accessible is simple, requiring only that data from written or printed records be translated into a machinable form (i.e., paper tape, magnetic tape, or punched cards). This is accomplished by using such devices as key punches and tape typewriters.

Making information machine-interpretable is more complex, requiring considerable preliminary planning and analysis. In the past, machine interpretability has been achieved in one of two ways:

1. By fixed-symbol coding, where each item of information is tagged with a special symbol; each symbol is used to signify one type of information (e.g., Δ = title-page information; ↑ = subject information; etc.).

2. By fixed-field coding, where particular kinds of information are associated with a fixed position in the machine medium; each position is used to signify one type of information (e.g., columns 20 through 30 = subject information; etc.).

Neither method is completely satisfactory. In the first method, a special symbol must be devised and used to identify every type of information to be encoded. The generation of these symbols becomes extremely complex when many different kinds of information are to be distinguished. A library catalog card, for example, contains 30 or more distinguishable types of information (e.g., class number, book number, title, imprint, etc.,). In addition, the symbols themselves are difficult to generate since the number of printing codes available on standard key-input equipment is severely limited (47 printing codes for IBM equipment, and 88 printing codes for tape typewriters). Fixed-symbol coding of data creates additional problems in printout. If the symbol codes are not printed out for each item of information, proofreading routines become cumbersome, requiring continual reference to some master record; if they are printed, the symbols create a cluttered record which reduces legibility.

In fixed-field coding, the amount and form of information included is critically affected by the constraints of the media being used. For example, the IBM card has 80 columns and only one printing character is allowed per column. Catalog cards contain from 300 to 500 characters, which means that if all information is to be printed in full, there will be a minimum of four

punched cards for every catalog card. This means, usually, that information is drastically abridged in order to save punched cards. In either case, a dedicated field is the basis of machine identification; therefore, that field must always be reserved for that type of information on every card regardless of whether a particular item has that information. In catalog cards for example, less than forty percent of the books cataloged have series statements. In fixed-field coding this position would have to be left vacant for more than sixty percent of the cards.

In both methods of making information machine-interpretable, processing of information for input is complex and places a great many restrictions on both the cataloger and the input keyer.

The machine-interpretable natural format combines the best features of both methods, resulting in an encoding method especially suitable for library purposes. Natural typing manipulations (i.e., carriage returns, tabulate shifts, and spaces) are used to delimit and identify data, thereby avoiding the rigidity of fixed-field coding. These typing manipulations are formalized in the format and are called "boundary markers." Information is recorded in a fixed sequence, thereby avoiding the complexities of fixed-symbol coding. These features reduce input keying routines to the level of ordinary typing, thereby reducing the usual cost of input processing. Figure 1 shows an example of a library catalog card translated into the machine interpretable natural format. This bibliographic adaptation is designed to fulfill three basic requirements:

1. To simulate the appearance of the conventional catalog card.
2. To allow interfiling compatibility with existing card catalogs.
3. To encode bibliographical data for machine manipulation and interpretation.

Two devices are used to structure the format and identify the nature of the data: sequential position (or the sequence of paragraphs) and boundary markers. In the following sections a detailed description is given of the bibliographic format showing how these devices are used.

5. SEQUENTIAL POSITION

Four paragraphs are used in the format, being distinguished by combinations of carriage returns, tabulate shifts, and spaces. Paragraphs are typed in a fixed sequence. Each paragraph is defined so as to correspond with a block of bibliographic information on a Library of Congress card. (cf. Figure 1).

Within each paragraph there are a controlled number of statements which are distinguished by a series of three spaces. Within statements, there can be any number of phrases, which are separated by two spaces.

The following is a brief identification of the four paragraphs used in the format, showing paragraph and statement sequences: (cf. Figure 1).

1. Call number paragraph
   - a. Class number statement
   - b. Author number statement
   - c. Main entry statement
2. Descriptive paragraph
   - a. Title and title-page transcription
   - b. Imprint statement
3. Collation and notes paragraph
   - a. Collation statement
   - b. Notes statement
4. Tracing paragraph
   - a. Subject statement

b. Added entry statement

c. Series statement

6. TRACING AND NON-TRACING MODE

Of the ten statements distinguished in the encoded format, only four are used for tracing purposes: 2a, 4a, 4b, and 4c. These statements are said to be in the "tracing mode." When processed on the Crossfiler or a computer, a tracing is automatically made for each complete phrase included in each of these statements. All other statements are in a "non-tracing" mode, and no tracings are pulled from them. Figure 2 shows a catalog card set produced by the Crossfiler, with the tracings automatically pulled and reformatted.

7. BOUNDARY MARKERS

Cards, paragraphs, statements, and phrases are separated and distinguished by conventional typing operations, which serve both as normal punctuation in the printout and as machine-interpretable boundary codes in the punched tape record. The boundary combinations are reserved for these used exclusively:

1. Cards are introduced or separated by a sequence of at least five carriage returns.
2. Paragraphs are separated by a sequence of one carriage return and two tabulate shifts.
3. Statements are separated by a sequence of three spaces.
4. Phrases are separated by a sequence of two spaces.

Figure 3 shows the encoded format with the boundary markers used to define cards, paragraphs, statements, and phrases.

8. PERMUTING AND ABRIDGING TRACINGS

Two special characters are provided to permute and abridge any statement in a tracing mode. The permuting device is a non-printing symbol that rearranges and pulls an additional heading for any statement (or phrase) in a tracing mode. This symbol is used

most frequently in the title statement to create shortened tracings or catchword titles. Figure 4 shows how the permuting bar is used to create additional tracings.

The semicolon is used to abridge any statement in the tracing mode, and frequently in the descriptive paragraph to set off incidental title-page information. Figure 4 shows how the semicolon is used.

9. THE CROSSFILER

The crossfiler is a computer-like device designed and built especially for cataloging automation, based on the concept of the machine interpretable natural format. Its primary function is to read, interpret, and manipulate bibliographic data from a properly-formatted punched paper tape, which represents a unit catalog card, to produce a paper tape representing a set of diversely-headed catalog cards. It does this by processing the punched paper tape representation of a unit catalog card and punching out a secondary tape representing a completed catalog card set. The expanded output tape is then loaded into a tape typewriter and typed out on continuous-form card stock.

The unexpanded input tape may be re-used to produce other library records, such as accessions lists, book-form indexes, book-pocket information, circulation records, etc. Since the tape format is designed to be compatible with general-purpose digital computers, a mechanized retrieval file is automatically accumulated as a by-product of the cataloging activity.

Operation of the Crossfiler is extremely simple, requiring only that input tapes be loaded into the reader and start buttons be pushed. The machine then automatically processes the original tape document to produce an expanded output tape.

10. MACHINE SYSTEM FOR CATALOGING AUTOMATION

A number of input processing routines have been developed to be used with the Crossfiler and its encoded fc at. The following description represents a practical application of this machine system and is presently being implemented in the AFCRL Research Library:

1. A specially designed worksheet (see Figure 6) is inserted in every book being processed before it passes to cataloging. The cataloger transcribes bibliographic information into the appropriate blocks of the worksheet. Completing the worksheet amounts to format translation, since the worksheet is designed to sequence and position blocks of bibliographical data for the encoded format.

2. Completed worksheets are then passed on to the tape-typist who transcribes the data from the worksheet. The necessary format boundary codes are inserted into the tape as part of the typing routine.

3. Initial typing is done on continuous form paper. Copy is read and corrected, and necessary corrections are made in the paper tape, which is then ready to be processed on the Crossfiler.

4. Corrected tapes are expanded by the Crossfiler, and these tapes are typed out on Flexowriters using continuous card stock, with top and bottom cuts and side perforations. Original input tapes are stored for future use.

5. Catalog card sets are assembled and sent to be filed in the appropriate catalogs.

11. CATALOGING COSTS

It is difficult to give exact cataloging costs for the machine system described above, since the system is still undergoing testing. However,

statistics accumulated during the experimental phase indicate that significant savings in terms of time and money will be realized once the system is fully operative. In the following sections, a number of statistics are presented and discussed, showing how and where these savings will be effected.

12. CATALOG CARD PRODUCTION COSTS

The clerical cost of preparing and filing a catalog card in the AFCRL Research Library is approximately 20 cents per card, or $1.40 per card set (an average of seven cards per set). This figure is based on using Library of Congress cards and includes the cost of ordering, materials, typing headings, proofing, sorting, and filing. It does not include any cataloging costs. The cost per card in the machine system is 12 cents per card, or $2.5 cents per card set. This includes the cost of typing, materials, sorting, and filing. (See Chart 1 for a detailed breakdown of these figures.) This figure does not include machine or overhead costs.

13. MACHINE COSTS

The machine system for catalog card production as described is amenable to both computer and Crossfiler application. However, since the Crossfiler was built for location within a library and operation by unskilled personnel, it does a limited number of functions at a lower cost and with greater efficiency than a general-purpose digital computer. In the following sections the machine costs for both applications are computed and compared.

It must be stressed that the cost figures quoted are approximate and must be considered as being merely indicative of what the real cost would be.

14. CROSSFILER COSTS

The 12 cents per-card cost figure for Crossfiler-produced cards quoted above does not include machine costs. The Crossfiler is a prototype machine; it is impossible, therefore, to give an exact cost figure for this first model since the research and development costs

would have to be included. The feasibility of producing commercial Crossfiler models is being studied, and preliminary estimates indicate that the price of production models will be in the twenty-to-thirty thousand dollar range. This figure does not include the price of the tape typewriter.* If one uses $25,000 as an approximate price, the cost per card for a library to produce card sets for 14,000 titles per year would be increased 2.4 cents per card, making the total cost per card 14.4 cents,the total cost per card set 99.3 cents. (See Chart 2 for a cost analysis of Crossfiler produced cards.) This figure represents using the Crossfiler at about 15% of its total processing capacity, or an actual processing time of one hour per day. This means that a single Crossfiler could be used by several co-operating libraries to handle their total cataloging load, thereby reducing the initial machine expense for any one library.

15. COMPUTER COSTS

Computer costs are usually figured on the basis of rental time rather than purchase. The cost, therefore, will vary from one computer service to another. The following cost figures are based on a rental charge of $75 per hour, which is an average cost for a medium to large sized computer. The rental charge does not include the cost of writing the computer program or the cost of a technician to operate the computer. Since these costs are variable, they have not been used in computing the following costs.

The tape representation of several thousand card sets were made using a PDP-1 computer.** The average processing time per card set was 1.5 minutes, or 100 card sets in 2 hours and 30 minutes. The cost per card set was computed to be $2.70 in this experiment, or 39 cents per card. It should be stressed that this figure represents the cost for a particular kind of computer and should be used only as a rough comparison figure. (See Chart 3 for breakdown of this figure.)

*A variety of tape typewriters are currently being marketed, ranging in price from $2,100 to $5,000.

** Manufactured by Digital Equipment Corporation, Maynard, Mass.

16.OTHER OPERATING EXPENSES

The same amount of time is needed to type out expanded tapes produced either on the Crossfiler or computer. A run of 100 card sets takes approximately 10 hours to type out on a tape typewriter, averaging 6 minutes per card set. Typing out of cards on the tape typewriter is an "unattended" clerical operation. This means that the typist loads the punched paper tape into the tape typewriter and is then free to do other work. The only additional attention required of her is to check the cards occasionally as they are being typed and then to burst the continuous-form card stock into card sets. These routines require about 30 minutes per 100 card sets. The clerical operating expense for this operation is, therefore, negligible.

17.CARD HANDLING TIME

The machine system promises to effect an appreciable savings in terms of card handling time. At present, the Library spends an average of 32 minutes of clerical time to prepare, process, and file a card set of LC cards.

The clerical handling time for machine-produced card sets is cut in half. The amount of time required to process, prepare, and file a card set using the Crossfiler is approximately 17.5 minutes. This time includes initial input typing, proofing, processing on the Crossfiler, and filing. The savings in time is primarily a result of cards being immediately ready for filing once they have been typed out on the tape typewriter.

18.ADDITIONAL FEATURES

The potential worth of the machinable input record produced as a by-product of the machine system has not been discussed in this paper. However, this machinable record can be used to produce a variety of bibliographic products. A number of applications using the Crossfiler are presently being experimented with. These include:

1. The production of announcement and accessions lists by automatically abridging and rearranging input records for various output printing formats.

2. The production of control records for bindery routines and circulation activities.

3. The production of intermediate data to be sorted and published as book-form catalogs with alphabetical indexes.

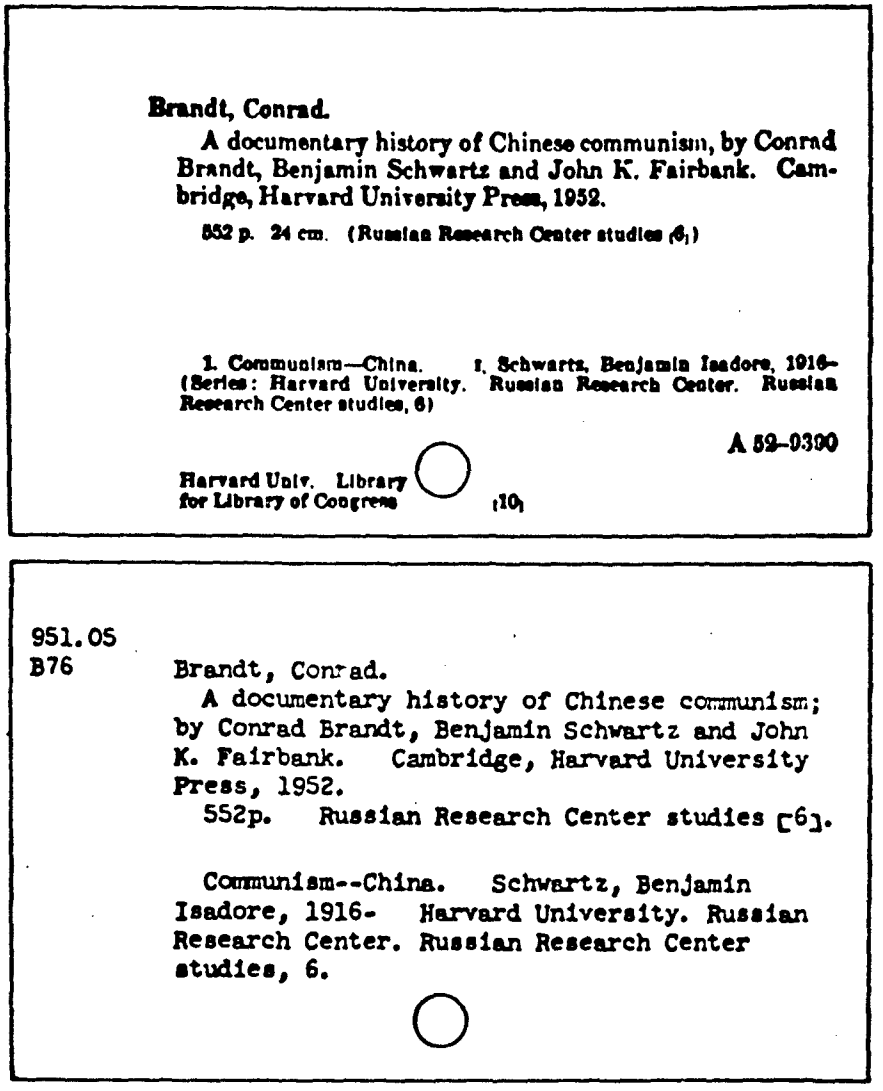
4. The production of selective lists according to input category codes or class numbers.

A number of applications using a computer are also being experimented with, including;

1. Generation of a machine-searchable file to produce demand bibliographies.

2. Automatic compilation of printed subject bibliographies with various indexes (i.e., author, title, subject, etc.)

3. Automatic compilation and typesetting of book-form catalogs with various indexes.



Brandt, Conrad.

A documentary history of Chinese communism, by Conrad Brandt, Benjamin Schwartz and John K. Fairbank. Cambridge, Harvard University Press, 1952.

552 p. 24 cm. (Russian Research Center studies 6)

1. Communism—China. I. Schwartz, Benjamin Isadore, 1916- (Series: Harvard University. Russian Research Center. Russian Research Center studies, 6)

A 52-0300

Harvard Univ. Library
for Library of Congress (10)

951.05
B76 Brandt, Conrad.
A documentary history of Chinese communism; by Conrad Brandt, Benjamin Schwartz and John K. Fairbank. Cambridge, Harvard University Press, 1952.
552p. Russian Research Center studies [6].

Communism--China. Schwartz, Benjamin Isadore, 1916- Harvard University. Russian Research Center. Russian Research Center studies, 6.

Fig. 1. Examples of Library of Congress catalog card layout and Crossfiler *machine-interpretable natural format.*

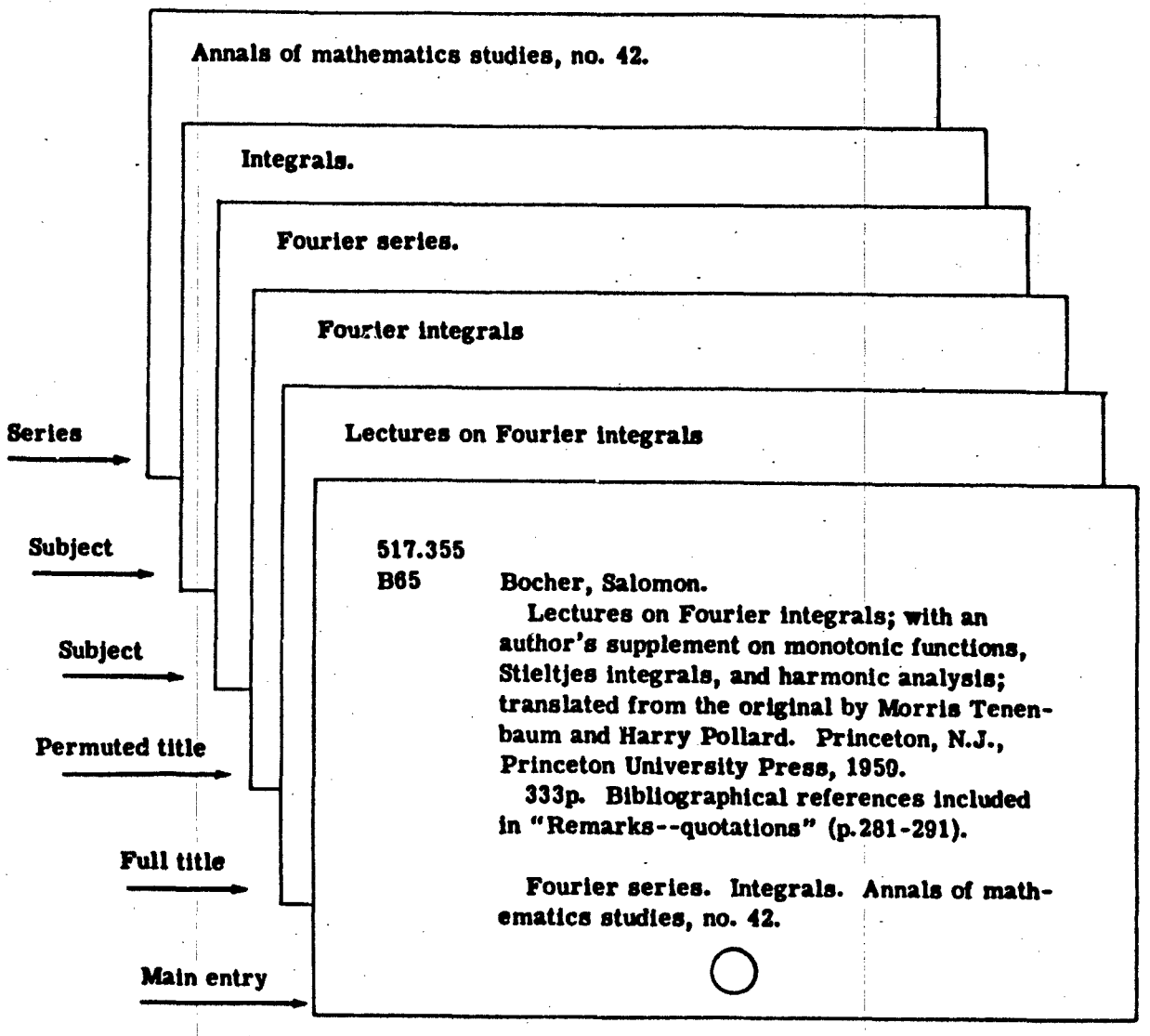


Fig. 2—Crossfiler-generated catalog card set

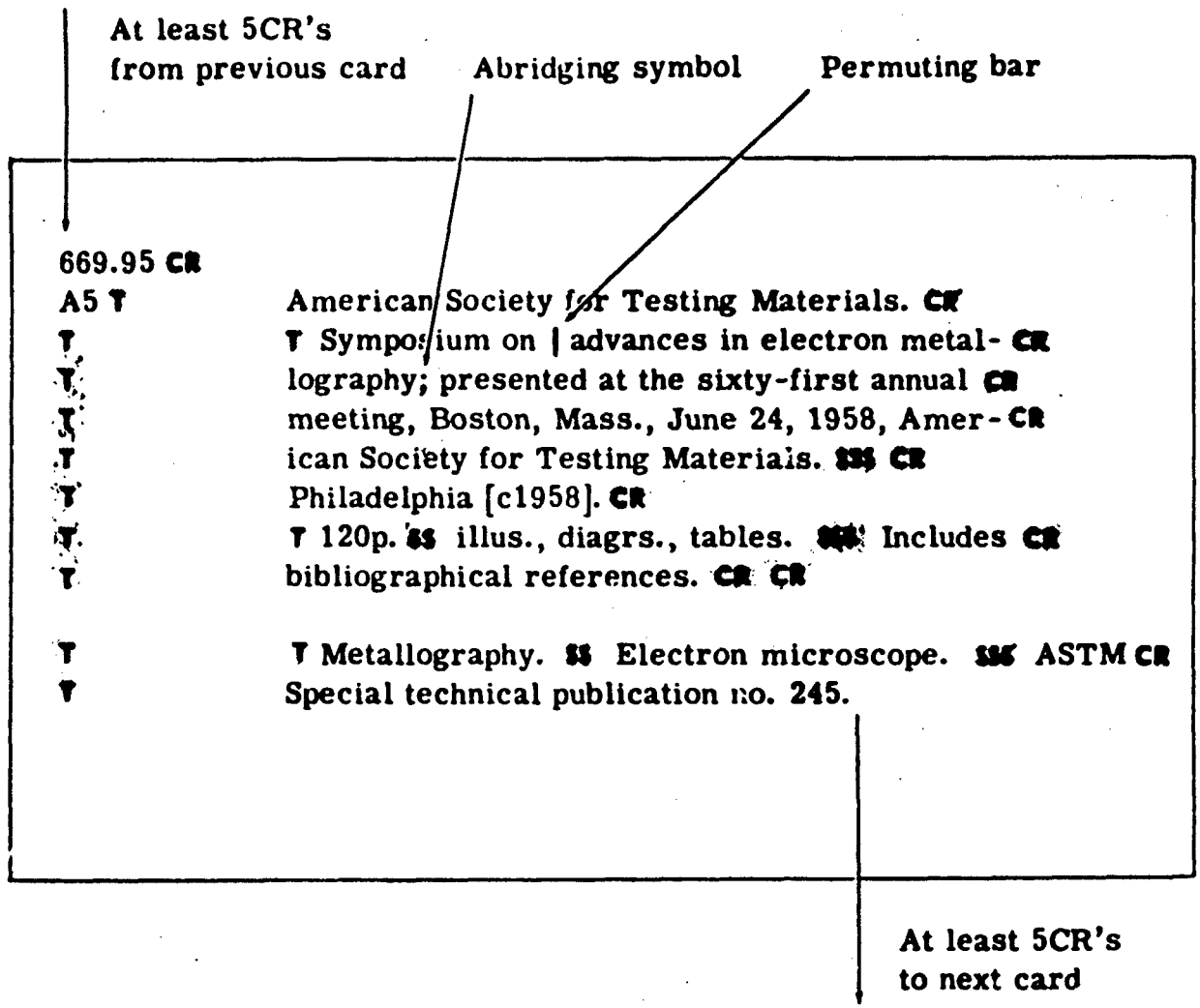


Fig. 3—Crossfiler *machine-interpretable natural format* showing typing manipulations used as boundary codes

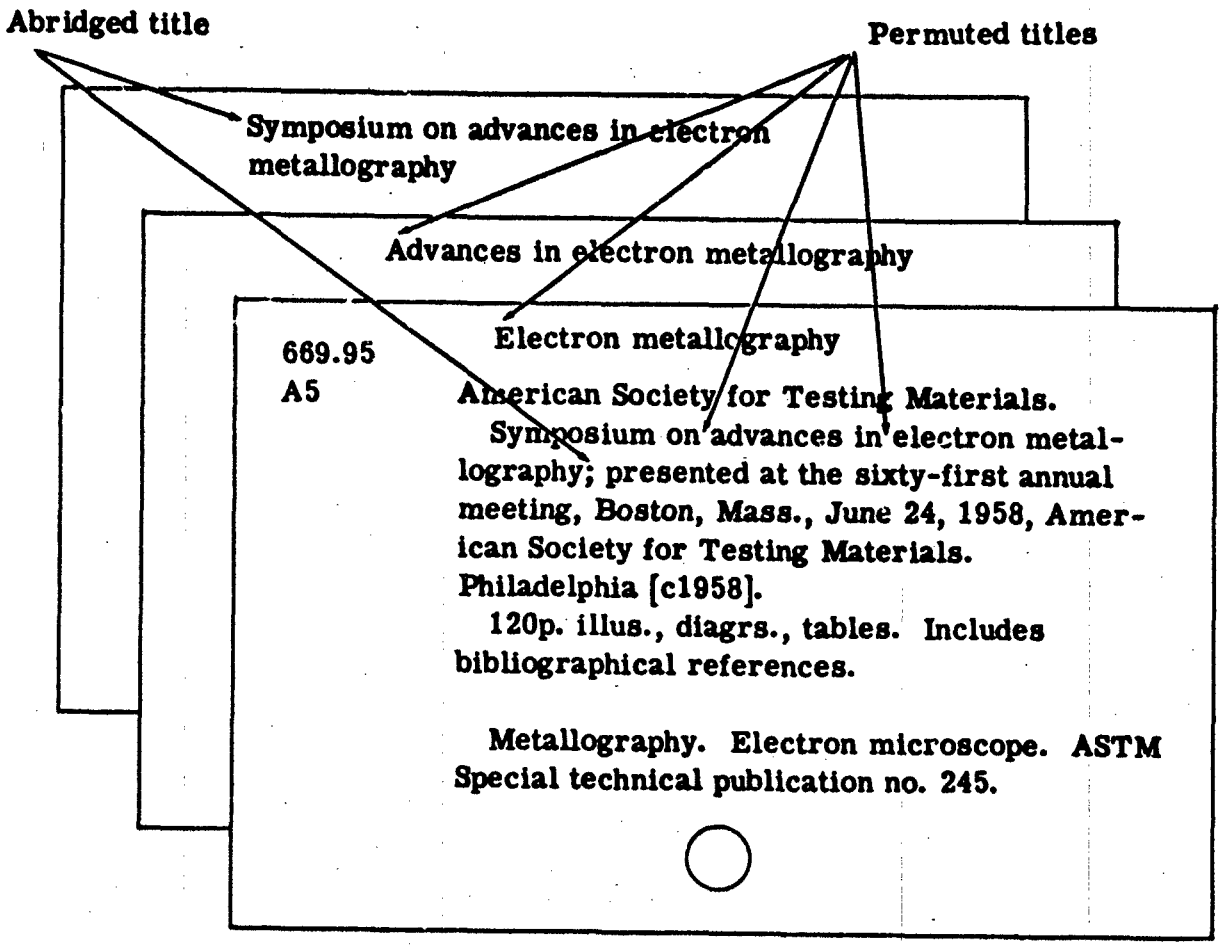


Fig. 4—Abridged and permuted title entries automatically generated and reformatted by Crossfiler

THE CROSSFILER SYSTEM: CATALOG CARD PRODUCTION

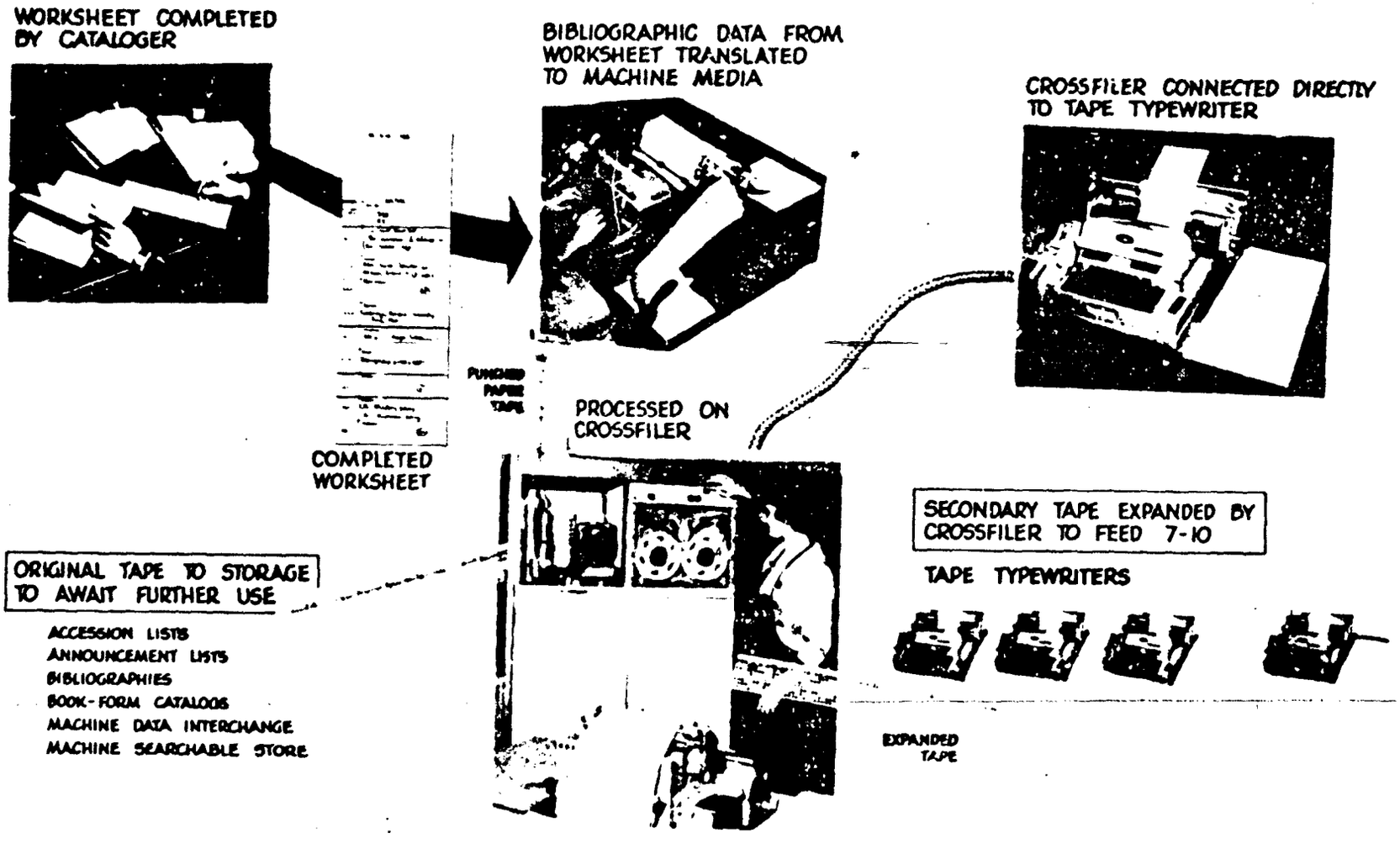


Fig. 5

BOUNDARY	DATA FIELD
CCCCC	Call Number 669.95 A5
T	Main entry American Society for Testing Materials. null
CTT	Title and Title-Page Transcription Symposium on advances in electron metallography; presented at the sixty-first annual meeting, Boston, Mass., June 24, 1958, American Society for Testing Materials.
SSS	Imprint Philadelphia [c1958]
CTT	Collation 120 p. illus., diagrs., tables.
SSS	Notes Includes bibliographical references. null
CTT	Subjects Metallography Electron microscope
SSS	Added Entries null
SSS	Series ASTM Special technical publication no. 245. null

Fig. 6—Worksheet for monograph cataloging

CHART 1

TIME-COST ANALYSIS FOR CATALOG CARD SET PRODUCTION BY MANUAL AND MACHINE METHODS IN THE AFCRL RESEARCH LIBRARY*

	Manual Methods				Machine Methods**	
	Using duplicate (LC) cards		Typing complete card sets			
Task Description	TIME	COST	TIME	COST	TIME	COST
Ordering routines for Library of Congress cards (includes receiving and checking)	960 min (16 hrs)	32¢ per card set	N/A	N/A	N/A	N/A
Typing of headings on LC unit cards to make card sets (includes time for proofreading, correcting)	840 min (14 hrs)	28¢ per card set	N/A	N/A	N/A	N/A
Hand typing of first unit card (does not include time for proofreading or correcting)	N/A	N/A	5 min per card	16.6¢ per card	5 min per card	16.6¢ per card
Proofreading of hand typed unit cards	N/A	N/A	1.8 min per card (1 person)	6¢ per card	2.4 min per card (2 people)	24¢ per card

CHART 1—*Continued*

Task Description	Manual Methods				Machine Methods**	
	Using duplicate (LC) cards		Typing complete card sets			
	TIME	COST	TIME	COST	TIME	COST
Hand typing of card sets from original unit card	N/A	N/A	30 min (6 × 5 min)	99.6¢ (6×16.6¢)	N/A	N/A
Proofreading of hand typed card sets	N/A	N/A	10.8 min (6 × 1.8 min)	36¢ per card set (6×6¢)	N/A	N/A
Production of the tape representation of a complete card set	N/A	N/A	N/A	N/A	Using the Crossfiler: 1 min per card set Using a computer: 1.5 min per card set	See Charts 2 & 3 for cost analysis
Typing out of card sets on tape typewriters from machine generated punched tape	N/A	N/A	N/A	N/A	6 min per card set	Unattended operation; cost not computed
Materials:						
(1) Library of Congress cards	N/A	56¢ per card set	N/A	N/A	N/A	N/A
(2) 3" × 5" card stock	N/A	N/A	N/A	5.1¢ per card set	N/A	N/A
(3) Continuous form card stock	N/A	N/A	N/A	N/A	N/A	11.9¢ per card set
Sorting and filing (per card set)	9.1 min.	30¢	9.1 min	30¢	9.1 min	30¢
TOTALS (per card set)	27.1 min	$1.46	56.7 min	$1.93	23.5 min	82.5¢

* Figures are based on processing 100 card sets, with an average of 7 cards per set.
**Machine costs and times are the same for Crossfiler and computer application except if noted otherwise.

CHART 2
COST ANALYSIS OF CROSSFILER PRODUCED CARD SETS*

Percentage of total Crossfiler processing capacity	100%	50%	25%	12.5%	6.25%
Total number cards produced in 1 year	806,400	403,200	201,600	100,800	50,400
Total number of card sets produced in 1 year	115,200	57,600	28,800	14,400	7,200
Machine cost per card	.3	.6¢	1.2¢	2.4¢	4.8¢
Final cost of machine produced cards	12.3	12.6¢	13.2¢	14.4¢	16.8¢
Final cost of machine produced card sets	82.7¢	86.7¢	90.9¢	99.3¢	$1.16

* Figured on using a Crossfiler exclusively for card set production. Based on a $25,000 machine cost and 10 year depreciation.

CHART 3
COMPARISON OF COMPUTER AND CROSSFILER PROCESSING TIMES AND COSTS FOR CATALOG CARD SET TAPE PRODUCTION

	Crossfiler	Computer*
Input reading speed	110 characters per second (cps)	200 cps
Output punching speed (paper tape)	110 cps	66 cps**
Average processing time required to produce one tape representation of a catalog card set	1 min (or 60 per hour)	1.5 min (or 40 per hour)
Total annual processing capacity	115,200 card sets (based on a 240 day work year)	Not computed; see below
Cost per card set	82.5¢ (see Chart 2)	$2.70 (based on a service rate of $75 per hour)

* Figures are for Digital Equipment Corporation's Standard 4-K PDP-1, with disc memory.
**High speed printing equipment is available and used with many computers, but these devices are capable of printing in upper-case letters only.