COMPUTER-BASED SUBJECT AUTHORITY FILES AT THE UNIVERSITY OF MINNESOTA LIBRARIES

Audrey N. GROSCH: University of Minnesota Libraries

A computer-based system to produce listings of topical subject terms and geographically subdivided terms is described. The system files and their associated listings are called the Subject Authority File (SAF) and the Geographic Authority File (GAF). Conversion, operation, problems, and costs of the system are presented. Details of the optical scanning conversion, with illustrations, show the relative ease of the technique for simple upper case data files. Program and data characteristics are illustrated with record layouts and sample listings.

INTRODUCTION

As a corollary to the creation and maintenance of large library catalogs, it has become necessary for academic or research libraries to maintain authority files of various kinds, such as author name, subject, series. In a manual cataloging system these files serve to unravel the mysteries of form, meaning, and usage to the cataloger. They also serve as a control to help avoid conflicts, synonyms, or overlapping subjects. With a system of decentralized catalogs using different subject entries from a system's union catalog, some method must be derived to preserve such usage for the cataloger. A computer-based subject authority file provides that means.

In January 1970, the University of Minnesota libraries began studying the relationship of subject authority files to both the present manual cataloging system and to a planned mechanized system employing the MARC II format for storage of bibliographic data. Minnesota's subject authority files are divided into two distinct logical files: Subject Authority and Geographic Authority Subdivisions. The Subject Authority File (SAF) contains all topical subject heading terms and their subdivisions down to nine

levels of term, and Geographic main headings, i.e. U.S. with nongeographic subdivisions. Nonterm data such as origin, usage notes, "libraries using," and other kinds of information are contained in the SAF. The Geographic Authority File (GAF) contains topical headings found in the SAF, with geographical place names as subdivisions and indications of direct or indirect terms in geographic heading assignment. Also similar nonterm data as found in the SAF are found in the GAF.

Immediate and long range benefits, together with the cost of conversion versus photocopying showed that greater flexibility would be achieved through the conversion to machine-readable form. Some of the benefits were:

1) immediate assistance to the libraries performing their own decentralized cataloging, while providing cards to the union catalog at Minnesota;
2) future assistance to our coordinate campus libraries should they wish to increase compatibility of their catalogs to the Minneapolis Campus union catalog;
3) future provisions of a machine-readable authority to enable linking of various subject vocabularies together for an on-line controlled vocabulary subject searching system.

When the decision had been made to convert the files to machine-readable form, we tried to determine what others had done regarding this application. Although much previous work has been done on subject analysis, cataloging, vocabulary construction, and mechanization of bibliographic processes, very few designers have developed systems to support thesauri or subject heading files. In 1967 Heald (1) reported on the system for TEST—Thesaurus of Engineering and Scientific Terms. The following year Hammond of Aries Corp. (2) described the NASA Thesaurus and Way (3) outlined in detail the Rand Corporation Library Subject Heading Authority List (SHAL) mechanized using punch cards and computer in 1967. Mount and Kollin (4) described the use of the computer in the updating and revision of the subject heading list for Applied Science and Technology Index.

Of course several famous information systems use mechanized thesauri, among them the National Library of Medicine's MEDLARS System with its MeSH vocabulary and the Department of Defense DDC Descriptors. In addition, the seventh edition of the Library of Congress Subject Headings utilized computer photocomposition.

Another reported work on subject headings in a mechanized system is that of the Library of Congress in which a MARC record for subject headings is discussed. Avram et al. (5) give examples of this record and describe the system now under development at LC. Unfortunately, for us, we completed the work herein reported in 1971, thereby not structuring our file to MARC specifications. We mention this work here, as our file will lend itself to such a conversion, should we later require it.

DATA PREPARATION AND FILE CONVERSION

The SAF and GAF files comprised 59 catalog card drawers of information (about 115,000 lines of typed data). Each file would be converted and maintained separately, but would use the same system design and processing programs. At a later stage, merging the files would be considered. Moreover, the cost of the system would be lower if one design could be used for both files.

Two conversion methods were evaluated, keypunching and optical scanning. Other methods would have lent themselves to this conversion, such as IBM Magnetic Tape Selectric Typewriters (MT/ST) or an on-line system such as IBM's Administrative Terminal System (ATS). However, because of the relatively small file size (under six million characters) and a desire for as economical a conversion as possible, only keypunching and optical scanning input were seriously considered. MT/ST typewriters were ruled out because of cost and lack of locally available tape conversion equipment. Keypunching was considered too slow in relation to typing. Our assessment of optical scanning as the cheapest method was confirmed later after completion of the conversion phase of the project, as an estimated $1800 in total savings over keypunching.

Files were converted without intermediate coding, permitting the typists to transcribe directly from the subject and geographic authority card files. The data preparation was done by the Catalog Division's subject authority coordinator. This librarian edited the file to eliminate ambiguities before the typist received the drawer. Otherwise, except for a quick check of the typist's finished sheets, the data were not examined again until after they were in machine readable form on tape. This procedure worked very smoothly, and caused the staff of the Catalog Division little inconvenience during the conversion phase. Figure 1 shows flow of the complete conversion activity.

Equipment used for preparation of the data consisted of two IBM Selectric typewriters Model 715 with carbon ribbon, dual cam inhibitor, and 065 typing element (Rabinow font). One machine had a pin feed platen. This feature later proved to make no discernible difference in the quality of the typed output, but some typists stated that they preferred the pin feed platen over the standard platen.

The Control Data 915 page reader with a CDC 8092 Teleprogrammer operating under GRASP III software was used for the conversion. Block time was rented at a commercial service bureau for $50.00 per hour. Library Systems Division personnel operated the system during these time periods. Control Data provided a system manual and debugging time in order to prepare for our operation during conversion. However, little assistance in handling the application was actually received from the Control Data personnel, who were familiar only with business data processing.

A stock form, called the CDC 915 page reader form, procured from a

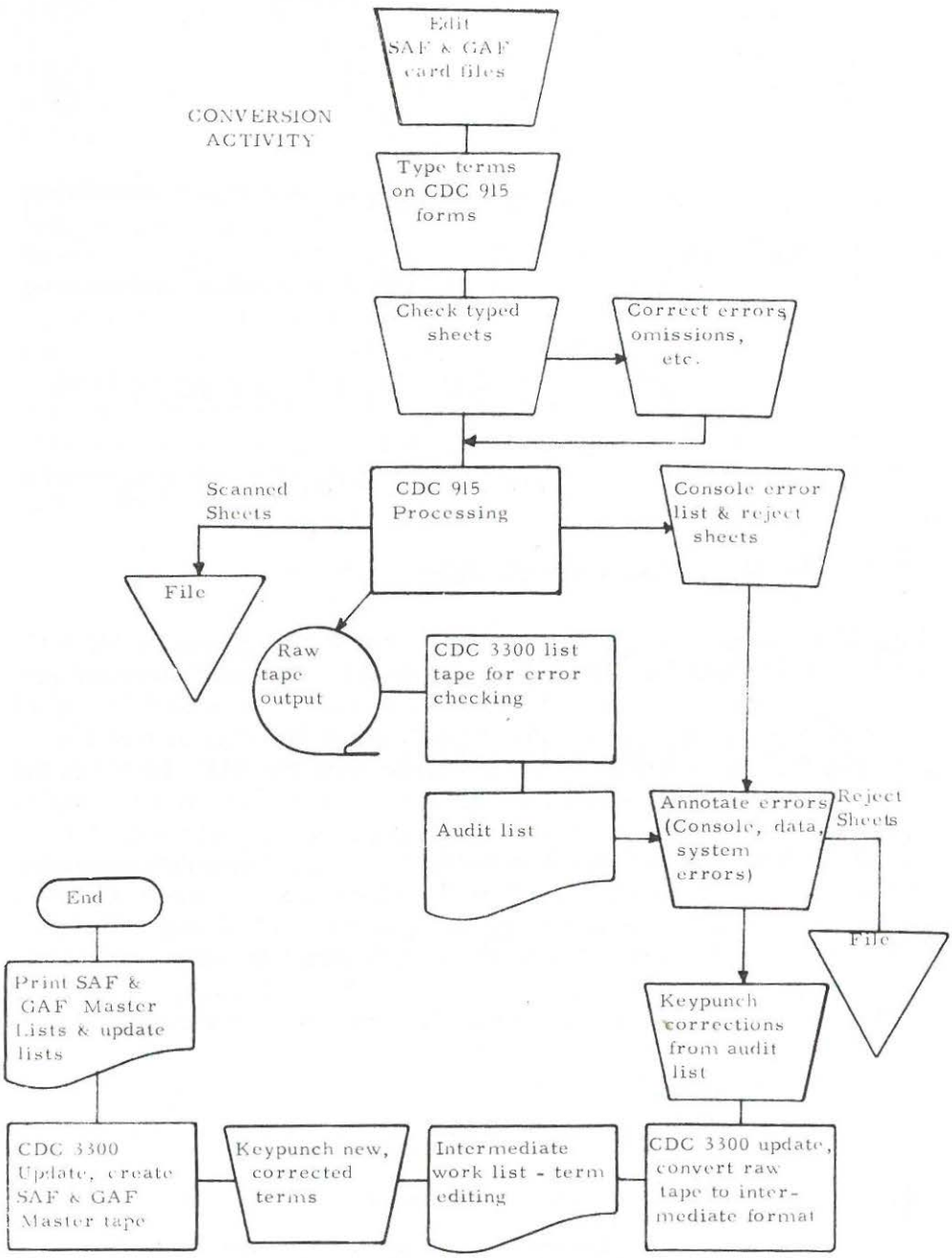

Fig. 1. Conversion Process for SAF and GAF

```
*00A
T SOCIAL SCIENCE RESEARCH$SGFA
T   ESSAYSA
T   PERIODICALSA
D     CL, EA
T SOCIAL SCIENCESA
N   DO NOT SUBDIVIDE FURTHER WITHOUT APPROVAL.A
T   ABSTRACTSA
T     PERIODICALSA
T       1900-    $MNUA
R         MNU PH.D. THESIS SHAW, GEORGE. [illegible]1970.A
N         ADOPTED JUNE 1970 PER RECOMMENDATION OF A.S.A
C         DO NOT DATE SUBDIVIS?DE FURTHER IN MNU CAT.A
#
```

Fig. 2. SAF Input Typing Sample Page

local forms vendor, was used. This form has a typing area of 9¾" x 13" marked off by faint blue lines. Top and bottom alignment areas are provided to check for line skew. Scanner throughput is increased by use of the longest permissible form with as much single line data as possible.

Figure 2 shows a portion of a typed page from the SAF. Line 1 is the format recognition line which was repeated on each sheet as a precaution against its loss by the optical scanner program during processing. Such a loss of the format recognition line would have forced complete rerunning of the job. The remaining lines show the various data elements identified by tag characters. The complete set of tag characters is shown in Table 1.

The end of page symbol # is used on pages which terminate before the last physical line of the page to increase scanner throughput. The ⊥ symbol terminates each line and serves the same speed-increasing function.

Table 1. Conversion Identification Tags

Tag	*Description*
T	Term
D	Departmental catalog in which the term is used
N	Scope note or general note on use of the term
C	Continuation line
R	Reference from which the term was verified if other than LC
Z	Followed by S = See; by SA = See also; by X = See tracing; by XX = See also tracing.
X	Geographic authority file cross reference tracing (implied).

Table 2. Term Subfield Indicators

Indicator	*Description*
$ SGF	Term also entered in GAF
$ DIR	Direct
$ IND	Indirect
$ MNU	Local University of Minnesota subject term
$ PROV	Provisional term
$ MeSH	Medical Subject heading term
$ NAL	National Agriculture Library term

Indentation spaces serve as a flag to the conversion program to show the level of the term or other data element. This technique decreased the number of characters to be typed, yet level errors were easy to detect during proofreading. Subfield indicators for certain nonterm data completed the input format used during conversion. Table 2 describes these indicators and the meaning of each subfield.

The GAF typed input is shown in Figure 3. Note the similarity between the two files, yet the presence of the variant treatment of an older term (SOCIAL SURVEYS IN) from a newer term (SOCIAL SCIENCES). As a result the Catalog Division has now changed these old form terms to conform with Library of Congress subject heading forms.

```
⋈□□↵
T SOCIAL SCIENCES↵
T  HISTORY$DIR↵
T    BYZANTINE EMPIRE↵
T      SOURCES↵
D        ART↵
T SOCIAL SURVEYS IN$DIR↵
X   AFRICA, SOUTH↵
X   ALABAMA↵
BRYNMAWRM?, WALES↵↓↓
X   BRYNMAWR, WALES↵
#
```

Fig. 3. GAF Input Typing Sample Page

During typing, error correction by typists was facilitated by the use of three special characters:

↓ —Delete line

? —Delete preceding character

‡ —Type over a character to delete character without inserting blanks.

A program is typed on an optical scanning sheet in an assembly level language for the CDC 915 page reader. It is then assembled into object code which operates the page reader and its controlling computer. An example of the program used in this conversion is shown in Table 3. Line 1 of this program defines the input-output and control characters together with a coordinate to terminate reading of a line if data are not found on the line. It also defines the special characters described above for error correction, end of line, etc. Line 2 specifies that a stock form (not preprinted) is to be read, giving the left-most and right-most character positions and maximum number of lines per page together with the first line number to establish the scanning area coordinates. These coordinates are expressed as three digit octal values determined through use of a forms grid and ruler. Line 3 describes the tape record format including the field size, the blank fill character, left or right justification, and alphanumeric or numeric only data field content. Line 4 instructs the 8092 teleprogrammer unit to convert certain characters to octal values matching the CDC 3300 computer system which are not identical to the normal 915 page reader octal values. The final E terminates reading of the program sheet. From this sheet GRASP III compiles an object program which is stored in the 8092 teleprogrammer memory, enabling scanner operation.

SYSTEM DESCRIPTION AND OPERATION

The raw data tape created during optical scanning was used to build the SAF and GAF data files. The magnetic tape coding is binary (odd parity) using 800 bpi density. A fixed length record of 20 characters is used with 100 records per physical block. As many 20 character Format C (continuation of data) records are used as needed to achieve variable length logical records. Table 4 shows the three record formats used.

Table 3. CDC 915 Program for Raw Data Tape Creation

```
|CTL|BLK,05|CAN,?|DLT,↓|EOL,^|EOP,#|FMT,⋈|⋈⋈

|STK|027,350,116,004|⋈⋈

|FMT,0|140,B,L,A|⋈⋈

|CNV|{,34|%,16|',56|⋈⋈

E
```

Table 4. SAF and GAF Record Formats

Format A - Control Record

Char. Pos.	Contents	Values
1	Record type	*
2-5	Page number	1-9999
6	Column number	1-3
7-14	File creation date	MM-DD-YY
15	File identification	
	Subj. Auth. (SAF)	S
	Geog. Auth. (GAF)	G
16-18	Columns used (123 standard)	123, 121, 111, 131
19-20	Number of lines per page (75 standard)	80 max.

Format B - Data Record (initial)

Char. Pos.	Contents	Values
1	Record Type	
	Term	T
	Reference term (GAF only)	X
	Reference	R
	Dept. Library	D
	See	1
	See also	2
	See from	3
	See also from	4
2	Level number	1-7
3	Sort exception code	
	Numeric exception	N
	Hyphen exception	H
	Substitution exceo.	S
	U.S. abbreviation	U
	Gt. Brit. abbrev.	G

Format B (Continued)

Char. Pos	Contents	Values
4	Qualification code (6 bit binary)	
	SGF (See Geographic)	1
	DIR (Direct entry)	2
	IND (Indirect ent.)	4
	PRO (Provisional entry)	8
	MNU (Minnesota term)	16
	MESH (Medical subj. heading term)	32
	NAL (National Agri. Library term)	48
	Combinations of these terms are possible. They are stored by adding the above values together, i.e. 17 - /MNU/SGF	
5-6	Number of display lines for item	
7-20	First 14 characters of item	

Format C - Data record (continuation)

Char. Pos.	Contents	Values
1	Record type	blank or .
2-20	Continuation of item.	

To change or modify the file, keypunched cards are used; one transaction card is used for each correction for both SAF and GAF files. Table 5 shows the layout of this card.

Table 5. SAF and GAF Transaction Card

Column	*Contents*	*Values*
1-4	Page of master list	1-9999
5	Column of master list	1-3
6-7	Line of master list	1-80
8-9	Sequence number	00-99 or blank
10	Deck number	0-9
11	Continuation number	blank or 0-9
12	Level number	1-7
13	Transaction type	
	Add	A
	Cancel	C
	Modify	M
14-15	Record type	
	Term	T
	Reference term (GAF)	XT
	Reference	R
	Departmental Library	D
	See	S
	See also	SA
	See from	X
	See also from	XX
16-80	Data	

Catalogers in the Wilson Library (the University's largest and central library) and the Bio-Medical Library use a 3 x 5 card as an input form. This card is filled in and transmitted to the librarian acting as subject coordinator. Then the information is keypunched and prepared for submission to an updating run. The normal schedule as originally planned was to run a cumulative supplement monthly, with a quarterly full updating of the file. However, this schedule has been flexible as the transaction volume has varied considerably from early estimates. Currently updates are run quarterly to produce supplements, with a full listing annually. These updates vary from 5,000 to 14,000 transactions.

The program for the system is written in COBOL for the CDC 3300 computer operating under the MASTER operating system. Upon demand the program performs four basic functions on the data files: 1) creation of a cumulative supplement list from a transaction card deck; 2) updating of the tape files from the transaction card deck; 3) preparation of master lists either during the update process or independently; and 4) querying the file on the basis of user defined search terms.

Parameter cards control the options available when supplements or master lists are to be run. The ACCEPT, DECK, LIST, ABORT, LINE, SPACE, COLUMN parameters provide control over cutoff for new supplement, transaction card list form, termination of job if the number of error cards exceeds a given value, number of lines per page of output, and number of blank lines before and after each transaction on the supplement, and whether a single or double column supplement is to be produced. Figure 4 shows a sample from the SAF Supplement.

The updating phase of the program creates the new master file and produces an update error listing accompanied by a report on composition of the file by level number, kind of data, and logical/physical record counts.

The master list printout is also controlled through parameter cards. The LINE, COLUMN, SELECT options indicate the number of lines of data to be printed in each column, the number of columns per page, and which pages are to be listed. This latter feature permits supplying replacements for pages improperly printed or bound and suppression of printing when a program restart is necessary. Figure 5 shows the most commonly used Master List format.

The file query function is performed upon demand to assist in file revision, to change a term throughout the file, or other special purpose. The search items can be composed of any and/or combinations of record types, record levels, qualification codes, sort exception codes, and key words or phrases.

A keyword search is a character by character search of file items. Thus, by specifying a root word, all derivatives of the word formed by adding prefixes or suffixes will be identified. If these derivatives are not desired, a blank preceding and/or following the root word in the search key will prevent their display. However, the word will not be identified if it is

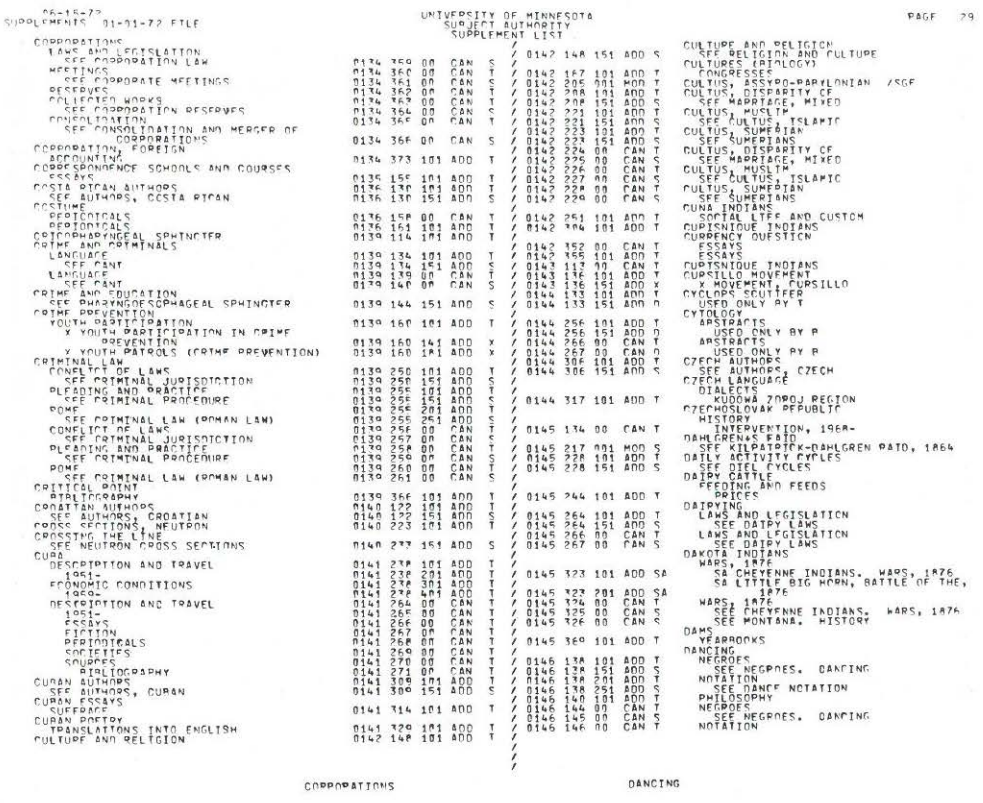

Fig. 4. Subject Authority File (SAF) Supplement List

preceded by or followed by a hyphen or punctuation. Matching on only the initial character positions or matching on exact match (i.e. no other characters in the item) are also available options.

The search terms are defined on the QUERY KEY parameter cards. A logical "and" operation is performed on search conditions on a single card. Up to 64 search terms can be used for a single pass of the query function. An additional option by use of a QUERY ONLY parameter card provides for search conditions applying to all search terms in the job, or to selected levels of the file only. The search output format is that of the supplement list, including an identifying number beside the search term satisfied by the querying function.

The supplement and master lists are run on "no-carbon-required" paper with vertical lines separating the columns, producing six copies. Those copies are sufficient to provide for the Catalog Division and Bio-Medical Library Catalog Division needs.

After a period of parallel maintenance the old card files used were removed from their trays giving one and one-half sections of drawers for

Fig. 5. Master List Format Using 3 Column Standard

expansion of the shelf list. Although file conversion took five months to complete, the program to operate the system was delayed because of termination of the programmer originally assigned to the project. Although the basic program features were ready in about 3-4 months, it was not until January of 1971 that the system was installed. During that year the staff gained experience in the system and cleansed the data of many ancient errors. By the end of the year, the system was an integral part of our Catalog Division support activities.

COSTS

As was pointed out previously, there was consideration given to photocopying the authority files to provide a duplicate set for the Bio-Medical Library. It was determined that this would cost $2,400 (60,000 cards @ $.04 each). This equalled the cost of the typing personnel and rental of optical scanning equipment. Moreover, there would have to be duplicate cards and filing to maintain both files, with no assurance that they would remain exact duplicates of one another. In our opinion the benefits of this

Table 6. Conversion Costs

Item	*Cost*
Senior clerk typists @ $2.40 (2 FTE for 3 mos.)	$1810.56
CDC 915 rental (20.1 hours @ $50 per hour)	1007.50
Typewriter purchase	532.70
Typewriter rental (2 mos.)	60.00
Magnetic tape	74.00
CDC 915 forms	400.00
CDC 3300 computer time @ $95.00/hr.	1411.45
Total	$5296.21

computer-based system offset the additional cost over the photocopying approach. To create these files completely cost $5,296.21 for all direct expenditures for clerical help, scanner time, typewriter purchase and rental, supplies, and CDC 3300 computer time. Table 6 shows the breakdown of these costs. During the conversion and development phase, salaries of the systems personnel were absorbed by the library so that only these direct costs were charged to the project. Also, the library absorbed the Subject Coordinator's time for editing the file of cards prior to typing. Two senior clerk-typists at $2.40 per hour each were employed for three months full time to type the data.

Operating costs are borne by the library, which requires a half time librarian as Subject Coordinator and a student keypunch operator for 15-20 hours per week. The Systems Division provides program maintenance as required. Supplies and computer time require about $2,100 per year if quarterly full lists are used with monthly supplements.

Some idea of the relative processing economy can be shown by examining some typical running times on the computer. The sizes of the SAF and GAF files are respectively 4.35 and 1.75 million characters. A typical supplement with 12,000 transactions takes 45 minutes to print on the CDC 3300 equipped with a 1000 line-per-minute printer for either SAF or GAF. Printing of a full master list for the SAF and GAF is 1 hour 25 minutes and 45 minutes respectively. Updating the files takes about 1 hour 40 minutes for 12,000 transactions. A query of the file takes about 30 minutes. Current computer and channel charges are $95 per hour.

GENERAL OBSERVATIONS

Our experience with this project has shown us the high reliability of the CDC 915 page reader as a conversion device. Less than 1 percent of the total amount of data the page reader scanned was rejected. Those errors rejected were easily spotted and retyped. No scanner-produced errors were found in the data; however, there was an occasional failure to pick up spaces when more than three occurred together. These errors were very infrequent and were discovered in the raw data proofreading. These errors were corrected and, after the final output file was generated, we again

checked for similar conditions and found everything in order with regard to term level indication.

With an upper-case file such as this, use of the CDC 915 is simple and easily accomplished. However, the library should not rely upon a scanner manufacturer or the installation where a unit is being leased to provide all the assistance required. The library will have to design its application and become familiar with the equipment in order to achieve best results.

All optical scanning usage requires that certain care be exercised in the typing operation. Lines must not be skewed, characters must not be blurred, and length of line can be critical even though the scan optics may be opened and closed over longer lines than are intended to be typed. Further, it is imperative that the paper used in the scanning operation meet specifications for use with the chosen scanner. Our experience indicates that a pin feed platen is not necessary to maintain forms alignment if typists use care in initial alignment.

We experienced some operational problems when we actually tried our program on the page reader. Initially, the system would not compile our program. It was not due to a catastrophic error in our program, but rather a hardware fault in the 8092 teleprogrammer. In trying to read the program onto tape after compilation, the system consistently failed. We finally gave up trying and recompiled from the scanned input sheet at the beginning of each conversion run. No one at the data center could explain our failure to load, but we must assume an intermittent or undetected hardware problem. During the job run it was imperative that the scanner be watched closely as occasionally it would stop reading or fail to feed a sheet. These were not difficult problems but did require occasional attention by the center's customer engineer. On one occasion the scanner failed during our run, and we could not achieve a timely repair. We rescheduled for the next week and then experienced no problem.

After our experiences with the 915 page reader at the data center we felt that we knew as much about the equipment as any of the operators we met while doing our production runs. We would not hesitate to use the page reader again for a simple file conversion, and would continue to handle the operation ourselves as the center operators were no better able to run our job.

ACKNOWLEDGMENTS

The author wishes to thank Mr. Eugene D. Lourey for developing the program for this system. Mr. Curt Herbert deserves recognition for the preliminary design for the system and initiating the optical scanning activities. Also, Mr. Carl O. Sandberg, who was responsible for the many details of the conversion portion and who now maintains these programs, contributed many significant design parameters. The staff of the Catalog Division, too, deserve our gratitude for their file cleansing and data editing during and after conversion.

REFERENCES

1. J. Heston Heald, *The Making of TEST—Thesaurus of Engineering Scientific Terms.* (Final Report of Project LEX, [U.S. Office of Naval Research: Nov. 1967] AD 661,001).
2. William Hammond, *Construction of the NASA Thesaurus, Computer Processing Support, Final Report.* (Aries Corp., 1968) N 68-28811.
3. William Way, "Subject Heading Authority List, Computer Prepared," *American Documentation* 19: 188–99, (April 1968).
4. Ellis Mount and Richard Kollin, "Analysis and Revision of Subject Headings for Applied Science and Technology Index," *Special Libraries* 60: 639–46, (Dec. 1969).
5. Henriette D. Avram, Lenore S. Maruyama, and John C. Rather, "Automation Activities in the Processing Department of the Library of Congress," *Library Resources and Technical Services* 16: 195–239, (Spring 1972).