Item Indexing by Automated Processes

By RUSSELL M. SMITH

Manuscript Division, Library of Congress

THE Presidential Papers program of the Library of Congress was launched in 1957 with the enactment of legislation authorizing the Library to arrange, microfilm, and index the 23 Presidential collections in its holdings. In 1958 funds were made available, the Presidential Papers Section of the Library's Manuscript Division was organized, and the project got underway.

Item indexing has been a traditional method of providing guides to personal papers, arranged as a rule in chronological order or chronologically within series; and the Library's Manuscript Division had in its Reading Room calendars and card indexes that followed this technique. Several were incomplete indexes to Presidential collections, unfinished because of the limitations of staff and staff time. Item indexing, then, was the natural approach to the task. The Presidential indexes were to be name indexes of correspondents and not subject indexes.

Upon considering the size of the task, with an estimated total of two million manuscripts to be indexed, the Library decided to automate the indexing process to facilitate handling and sorting the large volume of entries. The Library at that time was using unit-record or mechanical equipment for automatic data processing of its business operations. The Presidential Papers Section rented two printing card-punch machines and a card sorter. Another sorter and a tabulator for producing print-outs were available for the section's use in the Library's Tabulating Division.

Techniques for adjusting our indexing to machine methods had to be established. The index entry was confined to a form that could be placed on a standard 80-column punchcard. The entry consisted of seven parts or fields: a number designating the Presidential collection to which the entry belonged, the writer-recipient field, the date field, and other fields for series number, page count, additional information, and card count. The most important of these were the writer-recipient field and the date field. These were the fields to be sorted for arrangement of the entries in the published index.

The author is head of the Presidential Papers Section, Manuscript Division, Library of Congress. His paper is based on one read by him before the Society of American Archivists at Atlanta, Ga., on Oct. 7, 1966. Readers of Mr. Smith's paper may wish to refer to one by Fred Shelley on "The Presidential Papers Program of the Library of Congress," in *American Archivist,* 25:429–433 (Oct. 1962).

The first and last fields were for bibliographical control, the others for bibliographical data. In indexing three large 20th-century collections, an eighth field for numbers of subject files or case files was formed by taking five columns from the additional-information field.

After arranging the manuscripts, the staff worked through a series, writing index entries on 3″×5″ slips. The slips then were given to a card-punch operator for punching on a card. In later years some indexers with typing ability learned to index on the punchcard machines, bypassing the handwritten slip. To take advantage of their skills the section leased two more card-punch machines. The cards were listed by the tabulator, and the lists were edited. Edited lists were used to find and correct cards containing errors. With the entries in the same order as the manuscripts, a check against the document could easily be made. These listings were called "shelflists," a term borrowed from our librarian colleagues.

After editing at the shelflist stage, the cards were sorted alphabetically by the name appearing first in the writer-recipient field. Sorting in this field gave us problems since the names were of varying length and the unit-record sorter sorts a deck of cards column by column. We usually sorted the first 12 to 14 columns, which meant in short names, such as *Smart, G.,* more than the first name sorted; frequently the "to" or "fr" (for from) part of the entry and part of the second name became involved in the alphabetical sorting. In longer entries, such as *Cordoza de Oliviera, S. M.,* not all of the name sorted. We found it necessary to re-sort (by machine or manually) about 10 percent to 15 percent of the deck of cards. After the writer-recipient fields had been sorted, the date field was sorted to put multiple entries under a given name in chronological order. Sorting the two fields required that each card pass through the machine 25 to 30 times. Even without human error or machine breakdown, it was a tedious, time-consuming, and noisy operation.

With entries in alphabetical order, a second listing was run off on the tabulator, and a second editing was done. At this stage names, spelling, and corporate entry forms could more easily be standardized. After the alphabetical editing a third print-out was made on unruled sheets of paper carrying 115 lines per column. These sheets were mounted on boards, two columns of sheets per page, and were our printer's copy for photo-offset printing of the index. In summary, we had, after indexing, three listings: a "shelflist" for editing, an alphabetical list for editing, and an edited and reproduced alphabetical list for printer's copy.

In 1963 the Library placed with a computer firm, as is necessary,

a letter of intent to lease a computer, chiefly for its business operations. Realizing that when the computer arrived we would of necessity be using it, we made plans for a smooth transition to computer operation. The head of the Library's Data Processing Office (formerly the Tabulating Division) wrote a program for taping, sorting, and listing our index entries. The program was tested on another agency's computer, and our Benjamin Harrison index of 77,000 cards was taped and sorted on a Government Printing Office computer before the Library acquired its own computer.

When the Library's computer arrived in January 1964, three other programs were written for us: a shelflisting program, an edit program, and a program to produce printer's copy. The sort and edit programs were the major programing jobs. The records in the writer-recipient field were of variable length, and they included data to be excluded from the sorting process. Achieving a correct sorting with these factors present was the chief problem to be solved in the sort program.

It was solved by the creation of a "sort key." In this program the computer is instructed to search each entry as it is recorded on tape for a "to" or a "fr" preceded and followed by a blank space. The entry up to that point is recorded on an adjacent fixed-length tape record of 44 characters called the sort key. Also recorded in the sort key is the date, with numeric counterparts substituted for the alphabetic month abbreviations. The process is illustrated by the example on page 298.

Creation of the sort key also provided a means of following a few filing rules for final arrangement of the index entries. In the name *O'Bryan* the apostrophe has been omitted and characters to the right of the apostrophe moved one position to the left to make the name file as if it were *OBryan* with no apostrophe. In the entry for *Mrs. Walen,* the "Mrs." was dropped from the sort key, and the initials were moved four positions to the right. In addition, a "9" was placed in the sort key after the name in "Mrs." entries, forcing the *Mrs. Walen* entry to file after entries in her husband's name. The sort key contains only data to be sorted, with value substitutions for some characters to ensure proper filing.

The 124-character records, the sort key, with the index entry, are sorted with the result shown below. After sorting, the computer assigns each entry a unique "accessions" number, in our case an 8-digit number, creating a 132-character tape record. From this record an 88-character record, the index entry with its accessions number, is printed to make up our alphabetical listing for editing.

Editing from this point is done through the accessions number

SORT KEY GENERATED BY COMPUTER		CARD IMAGE	
WRITER OR RECIPIENT	DATE	WRITER OR RECIPIENT	DATE
MABE D E	19120916	27 MABE D E FR WW	1912 SE 16
MACCOOLEY J	191209 6	27 MCCOOLEY J FR WW	1912 SE 6
MACDOUGALL R V	19120916	27 MACDOUGALL R V FR WW	1912 SE 16
MACINTYRE J	19120916	27 MCINTYRE J FR WW	1912 SE 16
MADISON C W	19120916	27 MADISON C W FR WW	1912 SE 16
MADISON C W	19121099	27 MADISON C W	1912 OC–DE
OBEAR A C	19120916	27 OBEAR A C FR WW	1912 SE 16
OBRYAN G S	19120916	27 O'BRYAN G S FR WW	1912 SE 16
OBRYAN G S	9 19120114	27 O'BRYAN MRS G S TO WW	1912 JA 14
OLEARY J	19120916	27 O'LEARY J FR WW	1912 SE 16
OWEN W A	19120916	27 OWEN W A FR WW	1912 SE 16
PARSONS B C	00000001	27 PARSONS B C	SEE
SCOTT C R	00000001	27 SCOTT C R	SEE
WALEN J A	18901231	27 WALEN J A TO WW	1890 DE 31
WALEN J A	18901288	27 WALEN J A TO WW	1890 DE
WALEN J A	18908888	27 WALEN J A TO WW	1890
WALEN J A	18909988	27 WALEN J A TO WW	1890–91
WALEN J A	99999999	27 WALEN J A TO WW	ND
WALEN J A	99999999	27 WALEN J A TO WW	*ND1890
WALEN J A	9 19120916	27 WALEN MRS J A FR MRS L M RYAN	1912 SE 16

in very much the same way that a payroll or account record is updated by searching for the employee number or account number and changing the tape record accompanying it.

A change slip and a punched change card are used in the edit program. On the change slips an editor writes the accessions number of an entry to be edited and the correct data for only the field to be corrected. Corrections also may be new entries or deletions of whole entries. The data are punched on the change card with a code punch in column two indicating the type of change: "1" for a new entry, "2" for a deletion, and "3" for a change within an entry. Accessions numbers have a progressive 80-digit jump between entries to allow interfiling of new or changed entries between them. A changed entry may be moved to a new location within the alphabetical sequence by assigning to it a new accessions number that forces it to file in its proper place between original entries.

The punched change cards are used first to create an edit report consisting of the original entry from the tape and the corrected entry printed side by side, in the manner shown by the tabulation on the facing page. Only after the edit report is checked for ac-

TYPE CHANGE	ACCESSION NUMBER	PRES. NO.	NAME OF WRITER OR RECIPIENT	DATE OF DOCUMENT	NO. OF SERIES	NO. OF PAGES	ADDITIONAL INFORMATION	CROSS REFERENCE
CHANGE	01720300	7	SMITH B F TO AJ1	1818 JL 1	5	2	V11–P1 RECEIPTED ACCT	NO
TAPE	01720400	7	SMITH B L TO AJ1	1818 JL 1	5	2	V11–P1 RECEIPTED ACCT	NO
DELETE	01720480	7	SMITH B L TO AJ1	1818 JL 1	5	2	V11–P1 RECEIPTED ACCT	NO
CHANGE	01721200	7	SMITH D FR R C FOSTER	1806 JA 3	1	4		NO
TAPE	01721200	7	SMITH D FR P C FOSTER	1806 JA 3	1	4		NO
CHANGE	01723140	7	SMITH G TO AJ1	1813 NO 23	1	4		NO
TAPE	01723840	7	SMITH G O TO AJ1	1813 NO 23	1	4		NO
CHANGE	01726485	7	SMITH J–DESERTION	1817 JE 12	5	2	V10–P39 DESCRIPTION	NO
TAPE	01728560	7	SMITH J–DESERTION	1817 JE 12	5	2	V10–P39 DESCRIPTION	NO
CHANGE	01740100	7	SOMMERVILLE J TO*AJ1	1823 JE 7	1	2	RECEIPT	NO
TAPE	01740320	7	SOMMERVILLE J H TO*AJ1	1823 JE 7	1	2	RECEIPT	NO
DELETE	01740640	7	SOSA D TO PENSACOLA COMMRS	1821 AG 21	3	2	VM–P215	NO
CHANGE	01743700	7	SOUTHERN MIL DIV EASTERN SECT–TROOPS	1819 NO 30	5	3	V11–P103 RETURN	NO
TAPE	01742640	7	SOUTHERN DIV EASTERN SECT–TROOPS	1819 NO 30	5	3	V11–P103 RETURN	NO
CHANGE	01742970	7	SOUTHERN DIVISION–SICK TROOPS	1819 JE 30	5	2	V11–P61 RETURN	NO
TAPE	01742720	7	SOUTHERN DIV–SICK TROOPS	1819 JE 30	5	2	V11–P61 RETURN	NO
CHANGE	01743210	7	SOUTHERN MIL DIV EASTERN SECT	1819 JL 19	1	4	GENERAL ORDER	NO
TAPE	01742880	7	SOUTHERN DIVISION–EASTERN SECTION	1819 JL 19	1	4	GENERAL ORDER	NO
CHANGE	01742580	7	SOUTHARD S L FR AJ1	1827 MR 6	6	8		NO
TAPE	01744160	7	SOUTHWARD S L FR AJ1	1827 MR 6	6	8		NO
ADDS	01749940	7	STANTON H FR A HENNEN	1827 MY 26	1	1	BOOK LIST	NO
CHANGE	01749825	7	STANTON H FR A J DONELSON	1821 MY 1	3	1	VL–P217	NO
TAPE	01753280	7	STAUNTON H FR A J DONELSON	1821 MY 1	3	1	VL–P217	NO
CHANGE	01749830	7	STANTON H FR A J DONELSON	1821 MY 7	3	1	VL–P219	NO
TAPE	01753360	7	STAUNTON H FR A J DONELSON	1821 MY 7	3	1	VL–P219	NO
CHANGE	01755040	7	STEELE M TO R HAYS	1816 JL 23	1	2		NO
TAPE	01755040	7	STEELE M TO R HAYES	1816 JL 23	1	2		NO

curacy are the changes made on the tape record, with the change cards acting as instructions to the computer.

The computer has eliminated months of manhours in card handling. It sorts in hours what formerly required weeks of machine and manual sorting, and the computer does a better job. We once estimated that it would require a minimum of 8 man-months to sort the half million cards to be compiled for the William Howard Taft index. The computer can do this job within several days. By saving the cost and time of this type of card handling, the computer has enabled our staff to concentrate on the job of indexing and editing. It has not eliminated our indexers and editors. They have been "programed" in history classes and on the job for years, and no computer available to us today can equal their personal memory bank and their judgment.

Thus far we have published 16 indexes totaling 467,000 index entries. We have on tape and are editing approximately 800,000 more entries, with an additional 100,000 on cards and not yet taped.

In 1966 the Library substituted a new computer system, which necessitated changing our taped data file of index entries from a seven-track record to a nine-track record. This was a minor revision compared to conversion from unit-record equipment to the computer, but it reinforces our conviction that any automatic data processing project extended over a number of years must adjust to the rapid advances in equipment available for such projects.

Looking into the future affords some fascinating possibilities. This year the Government Printing Office, which prints our indexes, will begin using an electronic composing system. This system produces printer's copy for photo-offset from coded magnetic tape at a maximum rate of about a thousand characters a second. We expect to print our Theodore Roosevelt index of 244,000 entries by sending three reels of coded tape to the Government Printing Office. According to preliminary information given us, the system can print these index entries in three 500-page volumes in about a day.

Two large companies are developing microfilm information retrieval systems, wedding the computer to the microfilm camera, a development particularly pertinent to a program like ours, which combines an item index with a microfilm publication. With this system microfilm reels carrying coded index terms on each exposure may be searched for individual items and a quick print of that item made within a few seconds. Already the techniques used in our program are becoming obsolete!

The Presidential Papers program is primarily a data processing program rather than an information retrieval system. The pub-

lished indexes are the information retrieval tools with which the searcher retrieves the information he is seeking. At the end of the program, we shall have 1½ million index entries on tape. It is not too impractical to assume that a system could be instituted to retrieve selected entries or groups of entries from the tape record. Since the indexes are alphabetical, such retrieval might be by date, by series number, or by a combination of each. In this way, a different approach to the masses of index entries could be provided for the researcher.

The factors considered by the Library of Congress in automating the Presidential Papers program are similar to those that any item indexing project must consider: cost, utility, and justification of this form of guide.

To date, in our program, complete processing of an item has cost 72c. This cost includes the entire process: arranging, indexing, editing, microfilming, and publishing. Microfilming costs are about 15 percent of the total, or 11c. Approximately 50c of the 72c is the cost of indexing and editing. In the past 2 years, as we have moved into large 20th-century collections and have had the use of the computer, our cost-per-item has been reduced to slightly over 50c for this period. It may rise somewhat when we process two smaller and earlier collections.

Customer reaction to the products of the program is not wholly discernible. We have had few complaints. Users of the film and indexes usually have commented favorably. The Library, thus far, has sold over 40,000 reels of our microfilm to purchasers in 43 States and 4 foreign countries, and the Library's service copies of the microfilm are constantly circulated on interlibrary loan. We feel that the known reaction to our product is a justification for the program and that this form of justification will be stronger as more of our indexes and microfilm publications become available.

Item indexing in large collections (50,000 manuscripts or more) brings out information and correspondence relationships not previously known. For instance, comparatively little research has been done in the general correspondence of the Taft papers, which consists of approximately 250,000 manuscripts. The size of this series makes searching it a fearful task for any researcher without extensive funds and patience. We know now, for instance, that there are at least 241 letters to or from Warren G. Harding in the Taft papers, and we know exactly where to find them. Our Theodore Roosevelt index will reveal that this President corresponded with the widest variety of prominent men of any President

since Thomas Jefferson—among them Winston Churchill, Rudyard Kipling, Bat Masterson, John L. Sullivan, John Burroughs, and Thomas A. Edison. We believe item indexing will lead to more research, and more fruitful research, in collections of comparable size and quality. Chronological arrangement with item indexing preserves the picture of day-by-day historical development and at the same time gives ready access to individual manuscripts filed anywhere in the collection.

In justifying automated item indexing our experience has led us towards these standards for such a project: (1) the collections should be of major importance and should have a sustained high subject quality; (2) the collections, for maximum technical efficiency, should be fairly extensive so that the project can include enough items to make it a production job rather than a custom job; and (3) the collections should be related to one another in some way to increase efficiency in standardization of indexing techniques. They might center in a historical period, or around a subject, or relate to a geographical area, or have some other characteristic in common. If a series of collections meets these standards, automated item indexing is justified.