

# THE APPLICATION OF MECHANIZATION TO MANUSCRIPT CATALOGUE PRODUCTION IN THE PUBLIC ARCHIVES OF CANADA

by  
J. ATHERTON\*

"The electronic computer is a radically new type of machine - not simply an extension of our limbs and our senses. It can be conceived, in a sense, as an appendage to our central nervous system. Automatic data processing machines assist us in remembering and organizing external stimuli, in co-ordinating the tools which give us power over our environment, and in probing more deeply into the foundations of the sciences."<sup>1</sup> This definition clarifies somewhat the role of electronic data processing in twentieth century life. How can it be applied to the work of an archives? Probably most people who have had contact with that dusty institution resembling at the same time a library, a museum, and a mausoleum would consider the prospect of mixing antiquarianism with the most controversial product of our era as dangerously near the brink of insanity. The following paragraph, however, contains a clue as to why such a combination can make good sense:

"With the advent of the industrial revolution, the ratio of clerical personnel to directly productive labor has been growing continually. The invention of calculating equipment has, to a large extent, been motivated by the need for faster, cheaper, more efficient methods of processing data.

"Although the automatic computer was designed mainly in response to scientific needs, the requirements of business and government data handling have been a major stimulant to the future development of this machinery. The fact that about 1.5 billion lead pencils are sold yearly is perhaps some indication of the enormous area of information-processing activities which have yet to be mechanized."<sup>2</sup>

Any archivist charged with the efficient production of accurate finding aids for large units of manuscript material cannot help but feel that here indeed is one "area of information processing activities" which can be mechanized. A significant first step is the extensive Presidential Papers project undertaken by the Library of Congress. In this system, archivists prepare handwritten slips, the information from which is then put on key punch cards. These cards and shelf-list printouts taken from them are then proofread against both slips and, where necessary, original documents, "an arrangement of the cards in alphabetical order is accomplished by a noisy but efficient sorting machine,"<sup>3</sup> and run-offs are produced, forty-eight per cent of the original size, through a photo-offset process. The final product is an alphabetical author list giving for each document the names of writer and recipient, the date, the extent in pages, and occasional miscellaneous descriptive data such as notations to the effect that the document is a draft, extract, memorandum, or clipping, that it is written in French, and so on. In some units documents have been numbered and these numbers appear as aids to location. However, no subject descriptions are used. The limitations which such an omission imposes on the system are obvious. One's mind boggles at the thought of flipping through the index to the Abraham Lincoln Papers and gazing at over five hundred separate entries for correspondence from or to William H. Seward, in chronological order but with no subjects indicated anywhere.

---

\* Mr. Atherton of the Public Archives of Canada presented the paper at the annual meeting of the Archives Section, Canadian Historical Association, in Vancouver, B.C., June 9th, 1965.

1 William H. Desmond, *Computers and Their Uses*, Prentice-Hall, 1964, p.1.

2 *Ibid*, p.2.

3 Fred Shelley, "The Presidential Papers Program of the Library of Congress", *American Archivist*, XXV, 4 (October), p.432.

The Public Archives is concerned with the necessity for providing both author and subject lists, each containing sufficient detail to be of specific value to the historical scholar and each keyed to page numbers for rapid location of references. Our usual system involves fairly obvious and acceptable steps. Cards are prepared containing the author's name, the date, subjects dealt with, information as to occurrence of an enclosure or reply, indication if the document is in a form other than correspondence, and page numbers. Clerks hand sort these cards alphabetically and type author lists. They then re-sort to produce, eventually, a subject index. This step often necessitates production of additional cards, where more than one subject occurs in a single document.

Although the author sort may be started while descriptive work is in progress, the subject sort and actual typing cannot begin until the archivists have completed their work. This sorting-typing period which follows description inevitably postpones considerably the day when usable finding aids are available, especially in the case of large units. For example, the Robert Borden papers now being indexed contain 150,000 to 200,000 subject entries. Employment of one clerk in the sorting-typing phase would mean a delay of at least six years from the time when the archivists' descriptive work is completed until the finished indexes appear. With two clerks the delay would be three years; with three clerks, it would be two. The result, when it finally did appear, would be an author card index, giving no information other than page numbers, and a detailed subject list.

What will electronic data processing do for such a subject? Stated simply, it will provide us with more accurate and more complete finding aids at lower cost and in a fraction of the time (literally no time at all). Information given to key punchers on a regular basis during the course of professional descriptive work is placed on tape, and will be ready for sorting immediately upon completion of description. Programming also having been done ahead of time, the full sorting and printing process will produce, instead of one and a half full finding aids, **three** (sorted by author, subject, and date), within two weeks rather than at the least two years. Mainframe time (that is, time used for the actual sorting phase) on the Borden Papers project will be about ten hours; speed of printing is 600 lines (about 5,400 words) per minute.

Having eliminated the human error factor inherent in manual sorting and typing techniques, we should expect a much more accurate final product. Such accuracy, however, depends upon the careful application of rigid standardization. We must remember that the information produced by mechanized sorting is exactly the same in content as that supplied at the outset. It will have been sorted into a different order, of course, but the actual content will be the same. Since key punchers necessarily are trained to copy exactly what is in front of them, interpreting nothing, we must be certain our information for them is both accurate and consistent. This accuracy and consistency can be achieved by using a standard source document and abiding by a set of unbendable rules. Our source document has been designed to fit a standard twelve-characters-per-inch typewriter. Its spacing corresponds basically to that of a normal I.B.M. punch card, with a total of eighty spaces divided into suitably sized fields to take the various pieces of information. Thus the transfer of information from cards to the source document (Cataloguing Transcription) form is a clerical operation.

This source document has been organized so that all alphabetic information (which is entered in block letters) appears first and all numeric, last. This simplifies the operation for both our typists and the key punchers. Information as to author is entered in the first twenty spaces according to rigid rules. The surname always appears first, beginning in space 1. Two spaces then are interposed between surname and initials which, by the way, are always used in place of full Christian names - for the sake of standardization. Titles are kept to a minimum, the major acceptable ones being Sir, Lord, Lady, Duke of, and Mrs. Common civil or military distinctions such as Doctor, Judge, Honourable, Major, or Captain are never used. In the rare cases where a title does appear, two spaces intervene between initials and title. We never use periods or commas anywhere on the source document.

Standard archives indexing procedure decrees the necessity for providing some means of indicating that the author's name is not obvious, but was supplied **by the archivist**, from either internal or external evidence. Normally this can be accomplished through the use of square brackets. However, such a

APPLICATION OF MECHANISM

CATALOGUING TRANSCRIPTION

AUTHOR	EST A	SUBJECT	UNIT DESIC.	YR	YRDY	EST D	PAGE TO PAGE	
							69	70
CASGRAIN TC		WW I - CAN ARMED FORCES	H	14	165		15629	15621
CHRISTOPHER FS		MONUMENTS R	H	14	023		9869	9871
CONSERVATIVE MPS		BORDEN RL SIR - RESIGNATION RO	H	19	82		3812	
FOSTER GE SIR		WW I - CAN WAR EFFORT	H	14	09		16204	16205
HUGHES WM		IMP RELS - DEF	H	16	57		7976	7979
HARRIS AH		EXPORTS E	H	16	37		19261	19264
UNKNOWN		RAILWAYS - CPR RO	H	14	09	*	13082	
DO		WW I - CAN WAR EFFORT RO	H	DO		*	DO	
HIGNAULT PE		FISHERIES - N ATLANTIC ARB 1909 P	H	16	00		111	
MONTRL CLUB GARTIER		CAB APPOINTMENTS E	H	14	295		13688	13689
DO		CASGRAIN TC E	H	DO			DO	DO
DO		BLONDIN PE E	H	DO			DO	DO
COCHRANE F		RAILWAYS - CAN NORTHERN - OPERATIONS	H	16	27	*	13006	13007
DO		RAILWAYS - CAN GOVT	H	DO		*	DO	DO
REID JD		RAILWAYS - CAN NORTHERN - OPERATIONS M	H	00	000		13004	
DO		RAILWAYS - CAN GOVT M	H	00	000		DO	
PRIVY COUNCIL		RAILWAYS - CAN NORTHERN - FIN M4	H	00	000		12996	13003
DO		RAILWAYS - DULUTH WINNIPEG & PACIFIC M4	H	00	000		DO	DO
DO		RAILWAYS - CAN N COAL & CRE DOCK CO M4	H	00	000		DO	DO
DO		RAILWAYS - CAN N TOWN PROPERTIES CO M4	H	00	000		DO	DO
BOVILLE TC		FISCAL POL MRC	H	13	268		2408	
HOPPER J		AVIATION - CIVIL P	H	16	000		3579	3580
FIELDING WS		FINANCE - BUDGET P	H	01	079		3481	
BORDEN RL SIR		FINANCE - BUDGET P	H	01	083		3482	3483
DO		DO P	H	02	114		3486	
DO		DO P	H	02	065		3484	3485
DO		EDUC - SEPARATE SCHOOLS - MAN P	H	07	254		3483	
ROSS C SIR		WW I - MUNITIONS M	H	17	041		6663	
MENICHOLOS JT		MARRIAGE LAW P	H	08	000		1	
HOLMSTED GS		DO P	H	12	000		2	
		DO P	H	08	001		3	
HUGHES S SIR		NAVAL POL M	H	12	185		8504	8505
ROCHE WJ		R COMM ON PUB SERVICE 1912 M	H	12	083		8384	8389

technique creates problems in this case, because square brackets would interfere with electronic data sorting. In their place we make use of an asterisk which, when needed, can be placed in space 20.

The subject field fills half of the eighty spaces on our source document and, eventually, on the punch card. Nevertheless, we still make use of certain obvious abbreviations, in order to insure that we **never** run over the allotted forty spaces. For such well-known organizations as Trans Canada Airlines, the Canadian Pacific Railway, Canadian Broadcasting Corporation, or the United Nations, we always use their initials. All states in the U.S. and all provinces in Canada also appear in their abbreviated forms. Of course, the major problem of standardization appears when we try to work subjects into a mechanical sorting system. To insure consistency, we have established a master subject list which is kept up to date and consulted whenever the slightest doubt occurs.

We indicate the existence of replies or enclosures and describe items which are not correspondence in the subject field, spaces 59 to 61 inclusive, through the following abbreviations: E (Enclosure); R (Reply); RO (Reply Only); M (Memorandum); P (Pamphlet); and C (Clipping).

Column 62 on our source document contains an indication of the unit being indexed. Thus we can work on several sets of papers at one time, with the various units being separated automatically during the course of sorting. Since it is the Borden papers which now are being processed we are using H, the designating letter for these papers in our Manuscript Group 26.

To indicate the date, we use the "year-yearday" system. Since the productive period for any author, so far as a manuscript unit is concerned, never exceeds ninety-nine years, we need to use only two figures to represent the year. For example, in the case of Sir Robert Borden, whose papers extend from 1893 to 1937, 1893 need appear only as 93, 1906 as 06, 1937 as 37, and so on. To obtain a specific date within a year we simply number all the days of the year, allotting thirty-two numbers for each month, draw up a table, and hand the table to the clerks who type up our source documents. We use the same number of days for each month, even though all are not always needed and, therefore, "dead spots" will occur, simply for the sake numerical standardization. Each month may be thought of in terms of a multiple of thirty-two. But why thirty-two, if the longest months contain only thirty-one days? The extra number provides us with something to represent the month only when we may not be able to determine the exact day. Thus January is simply 001, January 1 is 002, etc. By placing the month only designation at the beginning of each month's range of numbers we are simply following standard archival sorting procedure, that is, moving from general to specific. The following dates, therefore, are in correct order within both the common and the "year-yearday" systems:

	1912	12000
	January, 1912	12001
	1 January, 1912	12002
	20 January, 1912	12021
	February, 1912	12033
	1 February, 1912	12034

Of course, we do not intend to force the users of our finding aids into a position where they must work backwards to obtain a sensible date from an odd-looking set of five numbers. Our computer will perform this operation for us and on the printouts they will be perfectly logical. As in the case of the author, we use an asterisk (in column 68) to indicate that the date has been supplied by the archivist.

Page numbers are entered in spaces 69 to 80, left-justified (that is, working left to right) from spaces 69 and 75.

Where more than one subject appears in one piece of correspondence it is necessary to repeat the author, date, and page numbers for each subject entry, because each entry requires a separate I.B.M. card. This duplicated information we indicate by placing the abbreviation "DO" in the first two spaces under the author, date, and pages for each subsequent time the same information appears. Should we ever find it necessary to repeat the subject, we simply place "DO" in spaces 22 and 23.

Our completed source documents are delivered on a regular weekly basis to the Department of National Revenue's Taxation Data Centre in Ottawa. Here the information is key punched onto I.B.M. cards, then transferred from cards to electronic tape. Although it sounds unnecessary, this card-to-tape conversion is a very practical step. One of the largest costs in an electronic data processing system is the charge for time used by the mainframe computer for the actual sorting operation. This mainframe time must be considered separate from the time for printing, which is done from sorted tape on a much simpler machine. Obviously, the faster we can bring information into the machine, the less will be our cost. To prove the point - "a punched-card reader attached to a computer reads in at the typical rate of 250 cards per minute, or 20,000 characters a minute. On the other hand, magnetic tape . . . can be brought into an electronic data processing machine at speeds in the order of tens of thousands of characters a second."<sup>4</sup>

After having been placed on tape in an unsorted, but numbered-as-received condition, a supply of data is then run off and a print-out returned to the Archives for checking and correcting of errors. The lines of information will appear in the print-out in numerical order. The archivist simply circles in red the numbers for the lines in which errors occur and corrects the mistakes in block letters. This corrected print-out is returned to the Taxation Data Centre with the next returns. Key punchers now make two new cards from the returned print-out. One is a correct entry to replace the one in which the error occurred. This entry receives a new number further on in the sequence of accumulated data and is sent back to the Archives in the next run-off, where it is automatically checked again as a new piece of information. The other new card contains nothing but the number of the incorrect entry. Cards such as these are kept separate. When all key-punching has been completed and all the information is on tape in numerical sequence, before any sorting takes place, the tape and these accumulated error cards will be put through a computer in parallel fashion. Whenever a numbered card matches a numbered entry on tape, the tape entry will simply be erased.

The Public Archives has tentative plans to use this process on other Prime Minister's papers. Depending on the success of electronic data processing on the Borden Papers, two years from now we may be involved in electronically producing finding aids for the papers of Sir John A. Macdonald, Sir John Thompson, Arthur Meighen, and W. L. Mackenzie King - all at the same time.

Eventually, with proper standardization of procedures and subjects, there is no reason why we could not produce a master finding aid for most, if not all, of the Prime Ministers' papers in our custody. A project as involved as this, previously unthinkable, is now nothing more than a logical extension of mechanical sorting techniques.

---

<sup>4</sup> *Desmond*, p. 123