Mechanization of the Manuscript Catalogue at the Public Archives of Canada

By JAY ATHERTON

Manuscript Division
Public Archives of Canada

JUST over 2 years ago the Public Archives of Canada completed a preliminary study and began to process data for the production of catalogues of the papers of Canada's Prime Ministers through an electronic data processing system. We had, of course, studied the indexes produced by the Library of Congress for its Presidential Papers. Our opinion was that, while these publications represent an extremely valuable first step, they are inadequate for our own particular needs. One important piece of information, which is the cornerstone of our cataloguing procedures, is missing from the Library of Congress prototype: the indication of subjects. One can open the author index to the Abraham Lincoln papers and find over 500 entries for correspondence from or to William H. Seward, in chronological order but with no subjects indicated anywhere. Even if we are to assume that most researchers come to these papers wanting to see correspondence from individuals, the omission of subjects imposes an obvious limitation on the usefulness of the catalogue. As it turns out, most historians using our Prime Ministers' Papers appear to be interested in material relating to specific subjects. Therefore the most useful type of finding aid that we could provide obviously would be one either arranged by subjects or at least indicating them.

With these needs in mind we approached various experts within the federal civil service and requested an estimate of the feasibility of our having such needs provided by machine sorting and printing. We then approached the Machine Branch of the Taxation Data Centre, Department of National Revenue, and gained the use of a small portion of their substantial array of key-punch apparatus and operators, computers, and programming staff.

Two basic factors must be considered when deciding upon a system involving the mechanical production of manuscript cata-

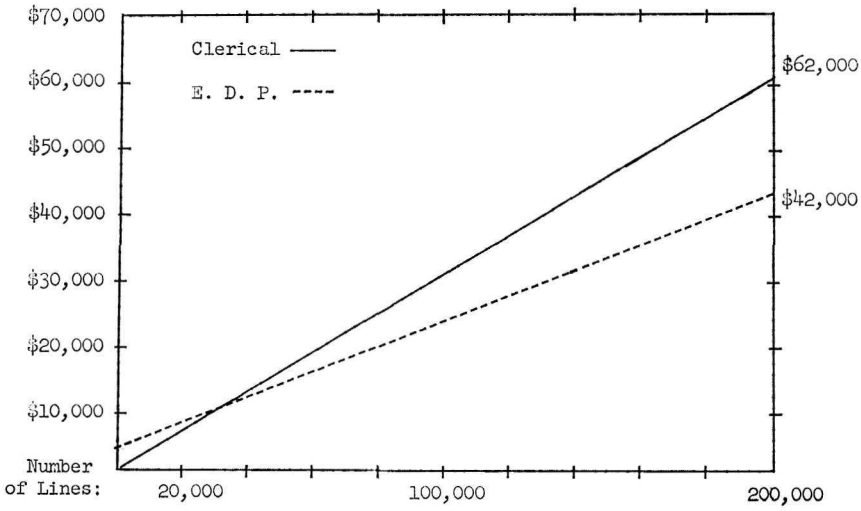
A member of the staff of the Manuscript Division of the Public Archives of Canada for nearly 6 years, the author now heads the division's Post-Confederate Section. His section has responsibility for the care, custody, and servicing of private manuscripts dated subsequent to 1867. Mr. Atherton's article has been especially prepared for this issue of the *American Archivist.*

logues or indexes: time and cost. For the sake of illustration let us look at these two factors in relation to a specific unit—in this case the papers of Canada's first Prime Minister, Sir John A. Macdonald. The Macdonald papers contain an estimated 150,000 lines.[1] Our statistics suggest that probable time for clerical hand sorting and typing of three complete finding aids (arranged by author, subject, and date) would be about one clerk-month per 1,000 entries. Of course, little of this operation could begin until all professional description had actually ceased. The production of catalogues to the Macdonald papers, therefore, would occupy one clerk for 150 months, or about 12 years. With two clerks this time would be 6 years, with three it would take 4 years, and even with a half dozen clerks working full time on sorting and typing 2 full years would elapse before we had our three finding aids ready for the use of researchers. By bringing mechanization to bear upon the sorting and printing operations, however, we should be able to produce our three detailed finding aids within 2 weeks rather than a number of years.[2] The time necessary for actual sorting will be less than 10 hours. Printing (at 600 lines, about 5,400 words, per minute) will take another 12 hours.

The other factor to be considered is that of cost. One would assume that such advantages as those described above would be practical only in the case of extremely large units. A cost comparison of our clerical and data processing methods, however, suggests a saving, through data processing, on any project consisting of more than about 20,000 entries (lines). In our case, admittedly, we are utilizing a government data centre and therefore are not paying commercial rates, but by the same token we assume that clerical production of the same catalogues would be accomplished by our own clerical staff—not Office Overload! Assuming that paper costs would be the same (and then insignificant) for the finished lists in the case of clerical production and for cataloguing transcription sheets in the case of data processing, we can ignore these two costs for the sake of our comparison. Thus our cost for clerical

[1] The number of lines here refers to the number of lines produced in preparation for sorting. A letter with two subjects would require two lines, and one with three would need three lines, each of which would contain the same data under author, date, page numbers, etc. but would have a different subject.

[2] In the case of this particular project the actual indexing has now been completed, but the finding aids will not be ready for another few months. This is because, although production of cards by archivists began in 1961, our actual work on the electronic data processing phase did not begin until a year ago. The catalogues of the Macdonald papers are scheduled for completion by July 1, 1967, as a centennial project of the Public Archives.



Comparison of costs of producing catalogues for varying sizes of units, according to the number of lines of data, by clerical and electronic data processing techniques.

production can be estimated as one clerk's average monthly salary (about $310) per 1,000 lines. In the mechanized system we have only one constant cost—programming—which in our case amounts to $2,000, irrespective of the size of the operation.[3] We have computed the variables (typing the transcription forms, key-punching and verifying, cost of IBM cards, card to tape conversion, checking, eliminating errors, sorting, formating, and printing) to be approximately $200 per 1,000 lines. Using these two sets of costs we can produce a graph, as shown here, that should be of interest to small as well as large institutions.

Since the beginning of 1965 we have been busily feeding information into our programme. This we accomplish by first having the data that our staff archivists place on 3″ × 5″ cards transferred to cataloguing transcription sheets so designed that the various fields correspond to those on an IBM card. This transfer of information is a clerical operation; the form has been designed to fit a standard 12-characters-per-inch typewriter.

A fundamental problem—which arose the moment we entered the land of EDP (electronic data processing) and possibly is the one that deterred the Library of Congress from indicating subjects—

[3] Actually we can consider this cost for the one unit as being a fraction of the $2,000, as we are using the same programme for a number of other units as well as for the Macdonald papers. However, for the sake of comparison at this time we can imagine that the programme has been designed for the one unit only.

was that of standardization. When producing cards for a hand-sorted index, the archivist can afford to be a little less than a perfectionist in his indication of names and subjects, relying upon the clerk who is sorting the cards to do a certain amount of simple editing. The clerk will ensure, for example, that the subjects "Canadian Pacific Railway" and "Railways—Canadian Pacific" appear under a common heading in the final product. Unfortunately, however, and despite what anyone may hope to the contrary, our electronic marvel cannot think. The way must be prepared for the proper utilization of the system or else the system will produce a veritable monster. Thus we must take great care to have the same subject appear in exactly the same way every time it is used. To be precise, it is up to the archivist to make an earnest attempt at indicating subjects and names of authors consistently, and it is up to the clerks who place data on the cataloguing transcription forms to ensure that spacing and spelling are correct. The difference of a single space or letter means that an item will be sorted out of order. We have achieved the necessary consistency through the use of our standard transcription form and a set of absolutely rigid rules. Names always appear surname first, followed by initials and then title if necessary. Titles are kept to a minimum. "Sir," "Lord," "Lady," "Mrs.," or "Duke of" are acceptable; common civil or military distinctions (such as "Judge," "Doctor," or "Colonel") are not. A machine sorting references to Major, later General, Joe Doakes would do so as if he were two separate men. Take away Mr. Doakes' oft-changing rank, however, and the problem disappears.

In the subject sphere the same sort of standardization is achieved through the use of a master subject list. The archivist may, through consultation with his colleagues, add subjects to the list if he so desires, but he must never deviate from it in his description of the subject matter of the documents being indexed.

To indicate the dates of documents we use a "year-yearday" system of 5 numbers, 2 for the year and 3 for the day within the year. For example, in the Macdonald papers the date 1 July 1867 is coded as 67194. (Our reason for using such code numbers at this stage is that space on an IBM card is at a premium, and "67194" takes only half the space of "1 JUL 1867.") Needless to say our programme provides for a change of these 5 numbers into readable date designations on the final print-outs.

The standard indexing procedure entailing the use of square brackets to indicate that a date or an author's name has been sup-

plied by the archivist, from internal or external evidence, has no place in a machine-sorted system. The placing of a square bracket in the first space of any field, where normally the first letter or number would appear, will cause the computer to sort this line as if this first space were blank. The result would be that an entry beginning "[DOAKES JJ]" would be sorted out of order at the very beginning of the author catalogue. For this reason we do not use square brackets on our transcription sheets. In their place we have the clerk type an asterisk in the space immediately following either the author or date field (*i.e.*, space 21 or 68), whichever is appropriate.

The existence of a reply, enclosure, or the like is indicated through the use of abbreviations such as R and E in spaces 59–61, at the end of the subject field. (RO stands for "reply only," which in our terminology means that the document described is an outgoing letter. Other abbreviations used less frequently are M for memorandum and C for clipping.)

We have also built into our programme a means whereby we can provide subject and author cross-references. A subject cross-reference is a simple matter. We merely place the two parts totally within the subject field on our transcription form, for example:

SEPARATE SCHOOLS SEE EDUCATION – SEPARATE SCHOOLS
or DOM-PROV RELS SEE ALSO INTERPROV RELS

The other fields are left blank. As our computer has been instructed not to sort blank fields, subject cross-references will appear in the subject catalogue and nowhere else. Author cross-references create a problem. Because of the size of the author field (it consists of only 20 spaces) we cannot fit into it two normal author entries necessary for a cross-reference. The solution, however, is obvious: we place the first half in the author field, let the rest come in the subject field, and leave all other spaces blank. For example:

GREAT LAKES INV CORP		*SEE ALSO BROWN GT
OTTAWA LUMBER CO		*SEE BRONSON AB

The asterisk at the beginning of the subject field tells our computer not to sort these lines in the subject index. (One shudders at the thought of 20 pages filled with entries all beginning *SEE and *SEE ALSO!)

The completed cataloguing transcription forms are regularly

sent to the Taxation Data Centre, where key-punch operators transfer the data to IBM cards. From cards the information is transferred to tape, in order to facilitate faster conveyance of the data into the main-frame computer for final sorting. (One can appreciate the significance of this step if he reflects on the fact that we pay for the use of this main-frame computer by the hour, and the hourly rate is $150!) An immediate runoff of the information as it is received at the Data Centre comes back to the Archives for checking and correction of errors.

When deciding upon the arrangement of each of the final lists the archivist of course must remember that his first duty is to serve the needs of future historical scholars. The historian may be interested in a particular person, subject, or time period. He may also want some refinement within each of these broad groupings. We believe that the system we have devised will satisfy most scholarly needs. Our author list, for instance, breaks down the correspondence from each author first by subject, then by date within each subject. Thus in the author catalogue for the papers of Sir Robert Borden (Canada's Prime Minister during World War I) we shall be able to find references to all the correspondence from Sir Sam Hughes, Minister of Militia for much of World War I, on the controversial subject of the Ross rifle. Further, we can even narrow this down to all the correspondence from Hughes on the same subject during, say, 1916—or even March 1916. The subject indexes will be arranged so that references to correspondence on a given subject are subarranged first chronologically, then by author. Thus we shall be able to locate all the correspondence on the subject of the Ross rifle written during the month of March 1916, from Hughes or anyone else. To aid the researcher in finding specific documents we shall have our chronological catalogue broken down alphabetically by authors within each date.

Although I have been discussing our electronic data processing system in terms of the Macdonald and Borden papers, our programme is capable of taking a number of different units at the same time. All we have to do is provide a suitable designation for each different unit to be processed and, as far as the experts at the Data Centre are concerned, we can go on forever—or at least for as long as we have unit designations. Since January 1965 we have been putting into the system data for three units: the papers of Macdonald, Borden, and Arthur Meighen (Borden's immediate successor as Prime Minister). Eventually we hope to be able to process all our Prime Ministers' Papers in this way. Once the emphasis on the Macdonald papers is over, we shall find ourselves

involved in producing data for the papers of W. L. Mackenzie King, R. B. Bennett (Prime Minister, 1930–35), and Sir John S. D. Thompson (1892–94), as well as Borden and Meighen. Before a decade has passed the Public Archives probably will have finished work on all these projects and will be working on the catalogues of the papers of other major political figures. Such progress would be absolutely impossible without the advantages of electronic data processing.