The Application of Automated Techniques in the Management and Control of Source Materials

By FRANK G. BURKE

Library of Congress

MAN is an inveterate adapter. His history is one of the application of narrow concepts to general needs. In the technological age such application has meant the proliferation of uses of a single invention far beyond what the inventor intended. The steam engine, developed to pump water from mines, became a power plant for locomotion, a means for creating electrical energy by driving generators, and, as a byproduct, a force to blow a penetrating whistle that replaced the steeple bell for mass communication of simple messages. The electric relay system, perfected to make possible telegraphic communication, developed into such diverse uses as the doorbell and the remotely controlled rail switch, and it became the mechanical predecessor of the electronic computer. The computer itself has followed this pattern of adaptation. Originally designed as a machine to aid in solving mathematical problems, it has been employed to address envelopes, to determine when magazine subscriptions will lapse, to control automobile traffic, to aid in the launching and progress of space flights, and to permit the designing of an apartment building with 167 different floor plans.[1]

By applying the computer in archival and manuscript work, the professions can attack problems that long ago were abandoned as prohibitively expensive. The computer is making possible information retrieval on a massive scale. The problem of retrieving information is not unfamiliar to archivists and manuscript curators. It existed even half a century ago, when collections generally were small, new accessions were few and far between, and the scholarly world was a rather tightly knit community of researchers and teachers who usually knew each other and knew what each was

This paper was written by Mr. Burke while still serving as Head of the Preparation Section, Manuscript Division, Library of Congress. His paper is a somewhat expanded version of one read before the Society of American Archivists at Atlanta, Ga., on Oct. 7, 1966. On Jan. 23, 1967, Mr. Burke transferred to the National Archives and Records Service as a member of the staff of the Office of Administration and Technical Services.

[1] Watergate East. Paul Amer, "What Will the Computer Do Next?" in *New York Times,* Apr. 24, 1966, sec. 11, p. 17.

doing. This academic Elysium, however, was shattered by several blasts: the population explosion, the education explosion, and—as a chain reaction from the two—the information explosion.[2] Lincoln's contention that "the dogmas of the quiet past are inadequate to the stormy present" is a universal and timeless assertion, no less applicable to the solution of archival problems than to Civil War politics.

Many segments of the academic and scholarly world are reacting to this "stormy present" by computerizing their processes. The most obvious, and the first to fall, so to speak, were fiscal units of universities. The physical sciences were quick to follow, and often the computer used for administrative fiscal purposes was shared by mathematicians and physicists and other scientists in an effort to spread the cost of the machinery. Recordkeeping in the registrar's office of some institutions then intruded itself into the system, and other administrative processes were adapted to the speed and efficiency of the computer. Between 1957 and 1964 the number of computers on American campuses increased from 40 to 400.[3]

Farsighted librarians were not slow to consider the possibilities of the new equipment for their own purposes. An indication of the future came with the appointment of a specialist in electronic information retrieval as head of the Graduate Library School at the University of Chicago. Concurrently a special committee, headed by Gilbert W. King, was appointed in 1961 to study the possibilities of automation at the Library of Congress; it made its report in 1963.[4] Since then, work has progressed on the development of an automated approach to bibliographic control of book material in the Library. Some special libraries, such as the National Library of Medicine, have already produced automated catalogs and computer-generated data about their holdings. The profession is moving rapidly toward a consensus on standards and methods.

Although great strides are being made in the evolution of a system for automating fiscal and book-related processes, the application of automated control to nonbook material in library collections stands at the threshold. As the King report states: "The Manuscript Division, Map Division, Music Division, and Prints and Photographs Division [of the Library of Congress] were considered

[2] See, for example, Julian P. Boyd, "A Modest Proposal To Meet an Urgent Need," presidential address before the American Historical Association, Dec. 29, 1964, in *American Historical Review,* 70:342–343 (Jan. 1965).

[3] American Council of Learned Societies, *Newsletter,* vol. 17, no. 4, p. 1 (Apr. 1966).

[4] *Automation and the Library of Congress; a Survey Sponsored by the Council on Library Resources, Inc.,* submitted by Gilbert W. King *et al.* (Washington, Library of Congress, 1963).

outside the scope of this report because their collections involve materials which differ markedly from the central library collections."[5] The assessment is correct: nonbook materials are different and should not (indeed, can not) be incorporated item-for-item into the core of an automated system designed for books.

This exclusion, however, does not imply that nonbook materials are not susceptible to automated control. In fact, there is likely to be more significant retrieval of information from *X* number of manuscript or archival collections than from *X* number of books, if generally accepted standards of description for each are used. The need for better retrieval of information has been evident for many years, but the profession has been fighting a losing battle against increased accessions and use without proportionately increased staff to process and analyze material. It would seem, in fact, that staff needs by old standards have become unrealistic. At the turn of the century the Manuscript Division of the Library of Congress contained 36,000 documents and had four staff members, for a ratio of one staff member to each 9,000 items. Had that proportion been maintained, the Division would now need a staff of over 3,000. To process materials today in the manner used half a century ago would thus require massive manpower. One of the factors has to change: either objectives or method. No curator wishes to sacrifice excellence to expediency, yet few have funds for the staff necessary to achieve excellence as heretofore understood. The factor that seems most susceptible to change, therefore, is method.

At this writing, machine processing has been applied to several aspects of manuscript and archival work. The ones to be discussed here are among the most prominent systems in the country, and, conveniently for discussion, they cover the full spectrum of bibliographic control in manuscript and archival operations.

At the most intense level is subject analysis of individual items within a collection. To state that a letter is from Mr. Y to Mr. Z is one thing, but to add that it concerns subjects *A*, *B*, and *C* is something else. If this could be done for each letter or document in each collection in a repository, it would be possible to produce a list of subjects that would answer most reference questions. Another approach would be to store the information and retrieve it on an individual-request basis. Once the information is properly put into the system, there would be little difficulty retrieving it on command. The great problem here is the amount of time spent in analyzing

[5] *Ibid.*, p. 38.

each document and translating the information it contains into machine-readable form.

An alternative to omnibus subject analysis is to analyze only specific subjects from a selected group of documents. This, in essence, is what has been done experimentally by the Winterthur Museum. Such selective analysis may be based on the subject orientation of the repository. A medical collection may index only references to medical affairs; State or local repositories may wish to concentrate on State or city references; and so on.

The criteria for such analysis are defined boundaries of subject interest and relatively small holdings. The staff of a repository that concentrates on historical documents of the colonial period of American history, knowing the value to researchers of specific subjects and names, could examine each document for subjects and names and code these into a retrievable system. Recently a project was undertaken cooperatively by the Henry Francis du Pont Winterthur Museum and the Graduate School of Library Science of the Drexel Institute of Technology whereby the Jonker Optic-Coincidence System was used to index a variety of materials in several institutions. The materials indexed included furniture and other artifacts, prints, manuscripts, books, public records, microfilm, and other forms of material. The optic-coincidence system is neither electronic nor fully automatic. It does not produce a printed page as an end product of its operation; it consists of a file of cards coded to indicate whatever it is desired to index. By means of a grid system of perforations, one can retrieve from the group of cards all those that are alike in any one of many features.[6]

The advantage of the Jonker (Termatrex) system over electronic computer equipment is that the cost, both initially for equipment and continuously for use, is far less than that of an electronic computer and associated equipment. Electronic machines could easily be used for this method of indexing, but, to be practical, a great amount of information would have to be desired, with a concomitant enlargement of input. Under certain circumstances Termatrex or other comparable systems may be considerably more efficient than electronic equipment if "real time" of machine use is considered. To be efficient, a computer should be operating almost continuously, but the Jonker equipment can sit quietly in a corner for hours or days and still be economically feasible for use in a small repository.

[6] Elizabeth Ingerman Wood, *Report on Project History Retrieval; Tests and Demonstrations of an Optic-Coincidence System of Information Retrieval for Historical Materials* (Philadelphia, Drexel Institute of Technology, 1966).

The size of a collection limits the feasibility of subject indexing in depth because of the costs of item analysis, transcription, and translation into machine-readable form. One limitation of the optic-coincidence system is that it provides only information retrieval and has no facility for printing the data once they are retrieved. It is therefore useful for answering queries directed to a repository by a researcher who either presents himself personally or makes a request by mail, but it does not permit the automatic preparation of lists of information that has been gleaned from the retrieval system. All the other retrieval methods discussed in this paper have that capability.

If it is not practical for an institution to undertake deep subject analysis in its collections, one may use another system, which is less detailed but still provides considerable information about a repository's holdings. This system is item indexing by machine, currently used in a number of institutions, including the Public Archives of Canada and the Library of Congress in its Presidential Papers program. The latter project was begun in 1958. The 23 Presidential collections in the Library's Manuscript Division have historically received heavy use both by researchers visiting the Library and by those outside the Washington area desiring photocopies of one or many documents from them. Because such use threatened wear and tear on the documents, the Library decided several years ago to make a master negative microfilm of each collection. Even earlier a project was underway to improve and standardize the existing card indexes to the Presidential collections in the Manuscript Division. With the advent of the microfilming program the two projects were merged and definitive arranging and indexing of the material was done before the filming. The index thus became the key to the microfilm.

Item indexing was not a development of the Presidential Papers program. It is, in fact, probably the oldest approach to correspondence in manuscript collections. The Manuscript Division, since its establishment in 1897, had itself prepared item indexes to many important collections. By the mid-1940's, however, the method had almost been abandoned as a bibliographic tool. The basic deterrent was cost. To index by item a 500- or 1,000-piece collection of colonial materials might have been feasible, since the size was not overwhelming and the content of each document might have warranted a separate index entry. But when collections became more and more the office or professional files of an individual, important though he might be, the bulk of the whole—and the relative insignificance of many individual items—led to limited returns from a massive indexing project.

The Presidential collections posed a dilemma in traditional manuscript terms. Collections of 20th-century Presidents were massive, and yet each item had actual or potential research significance greater than that of similar materials in large, non-Presidential collections. This fact, coupled with the desire to provide some approach to the microfilm to be made for each collection, led to a program for item indexing the 23 collections. The decision to prepare alphabetical indexes to hundreds of thousands of items led, in turn, to a consideration of electronic data processing for the production of printer's copy.

Not all collections can justify the effort necessary for item indexing, and yet there is no one arrangement of material that will satisfy all research demands. Though a chronological arrangement of papers is suitable for the chronological study of historical events, it does not readily reveal correspondence between the person around whom the collection is formed and specific correspondents. Nor does a chronological arrangement show the compartmentalized division of the careers of certain individuals, such as John D. Rockefeller, Julius Rosenwald, Harold H. Swift, or Bernard Baruch, careers which are an intricate pattern of many activities taking place simultaneously and quite often not closely related. To arrange such papers chronologically without indexing would destroy the patterns of activity of the individual. To arrange the papers by activity only would obscure the interrelationships of each. (Most collections of the papers of men of diverse activities are already, however, in subject order. If so arranged, they should probably be left that way.) To make a multiple card file for each possible index entry would be to proliferate card files beyond the normal capacity of repositories, and any addition to a collection so indexed would require a prodigious amount of interfiling and cross-referencing. A number of solutions to the problem have been posed, using electronic data processing. One of these is folder indexing.

This method, in at least one instance, is based on association of like materials and very general subject indexing by keyword. As done in the archives of the IBM corporation for the papers of T. J. Watson, the material is left in its original order, and unity and organization are gained by means of the index. This, in essence, eliminates processing, or the arrangement of material in a logical pattern. In large measure it is the true archival approach, since it respects the original order of material. Where the original order is totally chaotic, however, an archivist would feel it necessary to impose some rational system on the papers.

The folder indexing procedure begins with the numbering of each folder in a collection. This number is the index factor for all subjects, names, and dates within each folder. The data are then transcribed to punched cards in a prescribed format and translated into machine-readable form. The role of the computer then is to provide rapid sorting and listing (and sometimes instantaneous retrieval) of the information. Because indexing is provided only to the folder in which the material is located, the system borders on item indexing but is less specific. In the IBM project, subject headings from the items in the folders or the folder labels are run through a KWIC (Key Word In Context) program to reveal all the possible indexing terms in a folder caption. Thus, an entry titled "Research in Machine Retrieval" would automatically be indexed under the words "Research," "Machine," and "Retrieval."

This system was designed for, and is best suited to, answering questions directed to the repository from distant researchers, but it does not greatly facilitate research in the material itself. In most repositories the arrangement of manuscript and archival collections (particularly the former) is done with the needs of future researchers in mind. The concept of consanguinity of material is an important one to researchers working directly with the documents, and it is honored by most curators.

A system now being tested in an experimental project at the Herbert Hoover Archives in the Hoover Institution on War, Revolution, and Peace at Stanford University combines folder indexing with respect for material relationships. The Hoover Archives system, like that of IBM's T. J. Watson project, allows the machine sort-and-list to provide affinity of like material. It goes a step further, however, in that it provides for arrangement of material in some logical pattern before indexing, and this arrangement can be of assistance to the researcher who wants to work directly with the collection rather than query the finding aids. The Hoover method is, however, a combination of subject, item, and folder indexing, for presumably every item in each folder must be analyzed in order to indicate in the index the contents of the folders. The system is flexible in that obviously unimportant routine material, such as bills and monthly financial statements, can be amassed without subscripts into a single folder and a number and descriptors can be assigned to the folder only.

When the Manuscript Division of the Library of Congress approached the question of automation, several considerations were influential. One was the nature of the questions asked by researchers, another was the nature of the Division's supporting files and records for its collections, and the third and most important was the size of the Library's manuscript holdings.

Size was the dominant determination for automation. The Manuscript Division contains over 3,000 collections totaling some 30 million items—in more than 100,000 containers on about 9 miles of shelving. The questions to the staff range from the very simple to the very complex. The simple inquiry, received daily, is whether the Division has a specific collection. The staff then needs to know where the material is in the *ad hoc* shelving arrangement. Other basic questions concern size of collections, existence of finding aids, restrictions on use or access, physical condition of the material, and the like. More complex inquiries might concern literary rights, provenance, subject content of large collections, name (correspondent) analysis, chronological span of the material, or the relation of one collection to other holdings in the Division.

The answers to most of these questions, though available in the Division's files and records, are scattered in card indexes, title catalogs, typewritten finding aids (registers), printed guides, case files, and accession records. The question facing the Manuscript Division is essentially the same question facing librarians, archivists, and others interested in retrieving information: how does one convert many passive data files into one active one? In library terms, the problem could be solved by the conversion of all the information on a catalog card into rapidly retrievable data. In manuscript terms the problem entails the conversion of information from catalogs, guides, registers, case files, and accession records into a single record with retrievable elements.

An answer to this problem has been sought for many years. The question as related to the researcher does not limit itself to the papers or collections in one repository alone. An interim solution has been to provide the researcher with indexed guides to the holdings of many repositories. The most comprehensive such guide thus far produced is the *National Union Catalog of Manuscript Collections* (NUCMC), which lists collections of many repositories, giving information about the chronological span, the form of material, source, existence of finding aids, and restrictions on use, and providing other information needed by a researcher surveying the field. In addition, NUCMC provides, for each collection listed, a "scope-and-content note," which includes a qualitative analysis of the papers and lists the major subjects covered and major correspondents represented in the papers. These scope-and-content notes are then indexed, providing the researcher with an overview of his subject or with the names of persons in whom he is interested. In some repositories the NUCMC printed catalog cards are filed in a card file, and in some cases all of the index entries are also filed. This act in itself, however, does not create an active file of easily retrievable information. In essence it creates a multitude of passive files, which represent only part of the material in a repository. The reasons for this are clear.

The aim of NUCMC is to lead the researcher to the place where the material he seeks is most likely to be found. The *Catalog* "is intended to aid the scholar in his quest for manuscripts that may substantively advance his research";[7] it is not intended to be a definitive list of the contents of each collection reported to it. It cannot claim to be a replacement for finding aids and other devices at the repository that will direct the researcher to the parts of particular collections where a search might be most fruitful. The *Catalog* was not designed to be definitive or specific at the repository level, and one cannot evaluate it on these points. It is, in reality, a publication of abstracts of registers, inventories, indexes, calendars, and other finding aids. Its index, too, includes only the information abstracted from such aids, not all the information available about the collections at a repository. The index is thus at least twice removed from the actual contents of the collection itself. Nor does NUCMC, through its index, attempt to provide control over provenance or give other information, such as donors' names, specific restrictions on use and their time limitations, status of ownership by the repository, and occupations or professions of those around whom collections are formed. The existence of NUCMC does not relieve repositories of the responsibility for providing these facts, tailored to their own needs, or providing in-depth analysis of their holdings.

This latter feature—in-depth analysis—is what the three or four other automation projects in manuscript and archival collections mentioned here have been attempting to provide. They aim beyond the generalities of NUCMC for the specifics, gleaned from the material itself. But they do not aim at making passive files active in relation to descriptive information, and their approach to subject analysis is to go into the material and wrest from the sources a mass of specific facts—names, dates, and in some cases subjects.

[7] Library of Congress, *The National Union Catalog of Manuscript Collections, 1959–1961,* p. v (Ann Arbor, Mich., J. W. Edwards, 1962).

The Library of Congress Manuscript Division almost immediately dismissed this approach as impractical except for its Presidential collections and possibly a limited number of other collections of the papers of comparably eminent Americans. Not that the Winterthur, IBM, Hoover, and Presidential Papers programs did not have merit. Rather it was felt that the scope of the Division's holdings precluded the attainment of such specific goals in the foreseeable future. One could not easily attempt a reanalysis of 30 million items in over 3,000 collections. To go into the Division's 100,000 manuscript containers and to item index or even folder index them while doing the same for current acquisitions seemed highly impractical. Yet NUCMC and local card files were not answering the needs of researchers or the Division's staff. A solution became mandatory.

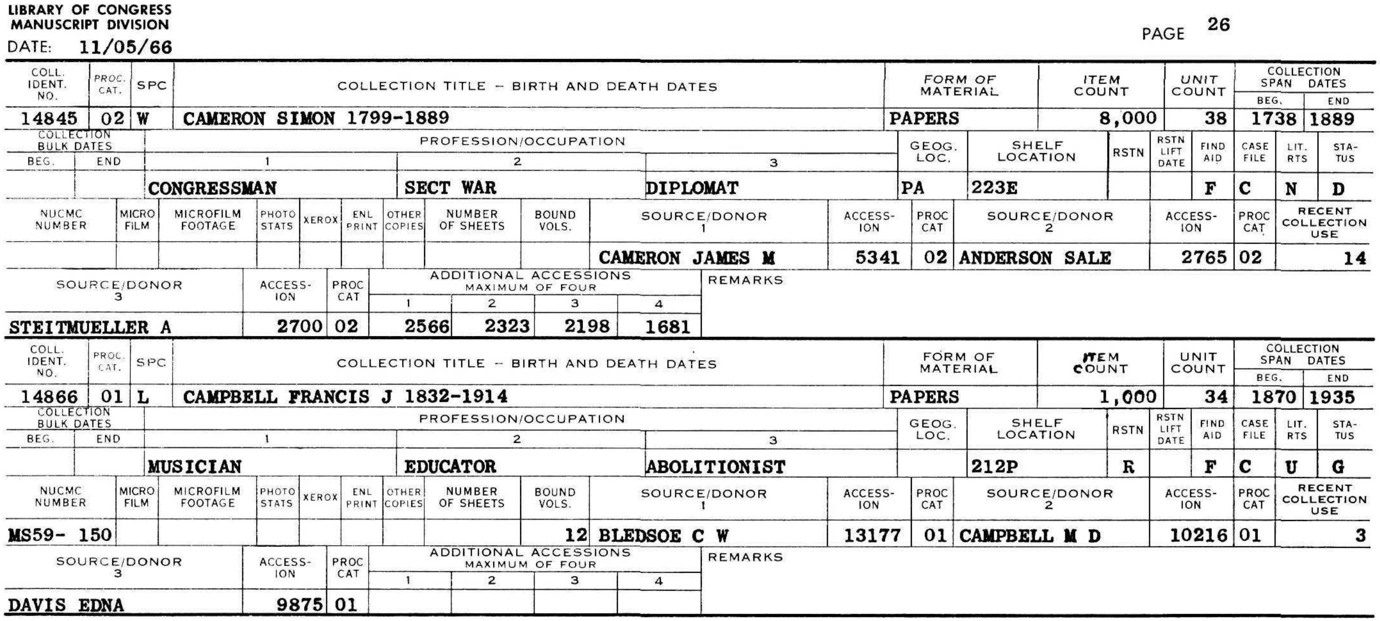
That solution, now at hand, assumes a number of forms. There is no quick, easy way to solve the problem, but there are a number of methods, each simple in conception, which form an integrated system of some complexity. The Division has begun a program that has proceeded by progressive stages—much like the Space Program, if the analogy will be forgiven. A number of suborbital shots were made first, and the experience of each successful (or unsuccessful) mission was cumulated into ever larger and more sophisticated projects. The Division had, as its aim, the development of two programs—one for automating its descriptive control devices (catalogs and records about collections) through data processing, and the other seeking an automated approach to information retrieval. Since the initial planninge stage, a third and a fourth goal have been added, which are really enlargements upon the initial goals: to include in the control program the data for statistical analysis of Divisional functions and operations and to develop an intermediate program for handling groups of subcollections (less than 50 items in a discrete unit) in a compatible system.

To accomplish these objectives, two computer programs were used: a sort-and-list program and a "Selective Permutation Index." The sort-and-list program developed over a 6-year period from a simple checklist of the Library's manuscript holdings. The checklist, a "title" list of the 3,000-odd collections, became the nucleus of the "master record" approach to machine use. Initially most of the elements of the NUCMC data sheet were coded into punched cards, so that the demand for information for NUCMC was the first impetus for automating these data. Utilizing a tape format instead of the card format avoided the limitations on length-of-line entry normally associated with punched cards. Discussion and planning led to the development of 98 separate items of information about each collection that it was desirable to bring together in the record. These items include collection title, dates, size, shelf location, occupation or profession of the person or corporation, processing status, restrictions, existence of finding aids, microfilm or other copies, use by readers or for other reasons, provenance, accession numbers, and NUCMC card numbers. Of the 98 items in the record, 75 are retrievable, 23 are statistically cumulative, and some are both. Gathering these data was time consuming but not difficult, since they were gleaned from existing records, cumulated onto forms, and key punched.

In the early stages of the project a print-out was produced every month, while cumulation was still progressing, and each print-out was edited for accuracy and updating. Since the most essential information about each collection was put into the record first, the monthly print-outs became progressively more useful tools within the Division (see Exhibit 1). For everyday staff use the print-outs soon supplanted the card catalog as a source of basic information about collection titles, shelf location, and size. Their use soon spread beyond the Manuscript Division; copies of specific lists are sent to other divisions in the Library for their information and use. For instance, a list of the Division's scientific collections is available in the Science and Technology Division. The lists are used by the Photoduplication Department as a means of cutting down the time necessary to search out manuscript materials for which there are photoduplication requests. The Exchange and Gift Division is regularly supplied with updated computer-produced lists of donors with whom the Library has entered into negotiations for the receipt of additional manuscript materials.

As part of this program the Manuscript Division has developed an automated call slip for manuscript material, which is clearer, more accurate, and easier for the researcher to use than the old paper call slip (see Exhibit 2). The call slip, as part of the master record, provides a quick approach to statistics on collection use; these were previously so difficult to cumulate that the practice of providing certain figures on use was abandoned several years ago.

The "master record" concept employed in the Division's programs is not new; it is new only to manuscript collections. It was borrowed from the "Master Employee Record" used by the Library since it began to operate its own computer in 1964. The employee record uses the employee as the central figure around whom some



LIBRARY OF CONGRESS
MANUSCRIPT DIVISION
DATE: 11/05/66

PAGE 26

COLL. IDENT. NO.	PROC. CAT.	SPC	COLLECTION TITLE – BIRTH AND DEATH DATES	FORM OF MATERIAL	ITEM COUNT	UNIT COUNT	COLLECTION SPAN DATES	
							BEG.	END
14845	02	W	CAMERON SIMON 1799-1889	PAPERS	8,000	38	1738	1889

COLLECTION BULK DATES		PROFESSION/OCCUPATION			GEOG. LOC.	SHELF LOCATION	RSTN	RSTN LIFT DATE	FIND AID	CASE FILE	LIT. RTS	STA-TUS
BEG.	END	1	2	3								
		CONGRESSMAN	SECT WAR	DIPLOMAT	PA	223E			F	C	N	D

NUCMC NUMBER	MICRO FILM	MICROFILM FOOTAGE	PHOTO STATS	XEROX	ENL PRINT	OTHER COPIES	NUMBER OF SHEETS	BOUND VOLS.	SOURCE/DONOR 1	ACCESS-ION	PROC CAT	SOURCE/DONOR 2	ACCESS-ION	PROC CAT	RECENT COLLECTION USE
									CAMERON JAMES M	5341	02	ANDERSON SALE	2765	02	14

SOURCE/DONOR 3	ACCESS-ION	PROC CAT	ADDITIONAL ACCESSIONS MAXIMUM OF FOUR				REMARKS
			1	2	3	4	
STEITMUELLER A	2700	02	2566	2323	2198	1681	

COLL. IDENT. NO.	PROC. CAT.	SPC	COLLECTION TITLE – BIRTH AND DEATH DATES	FORM OF MATERIAL	ITEM COUNT	UNIT COUNT	COLLECTION SPAN DATES	
							BEG.	END
14866	01	L	CAMPBELL FRANCIS J 1832-1914	PAPERS	1,000	34	1870	1935

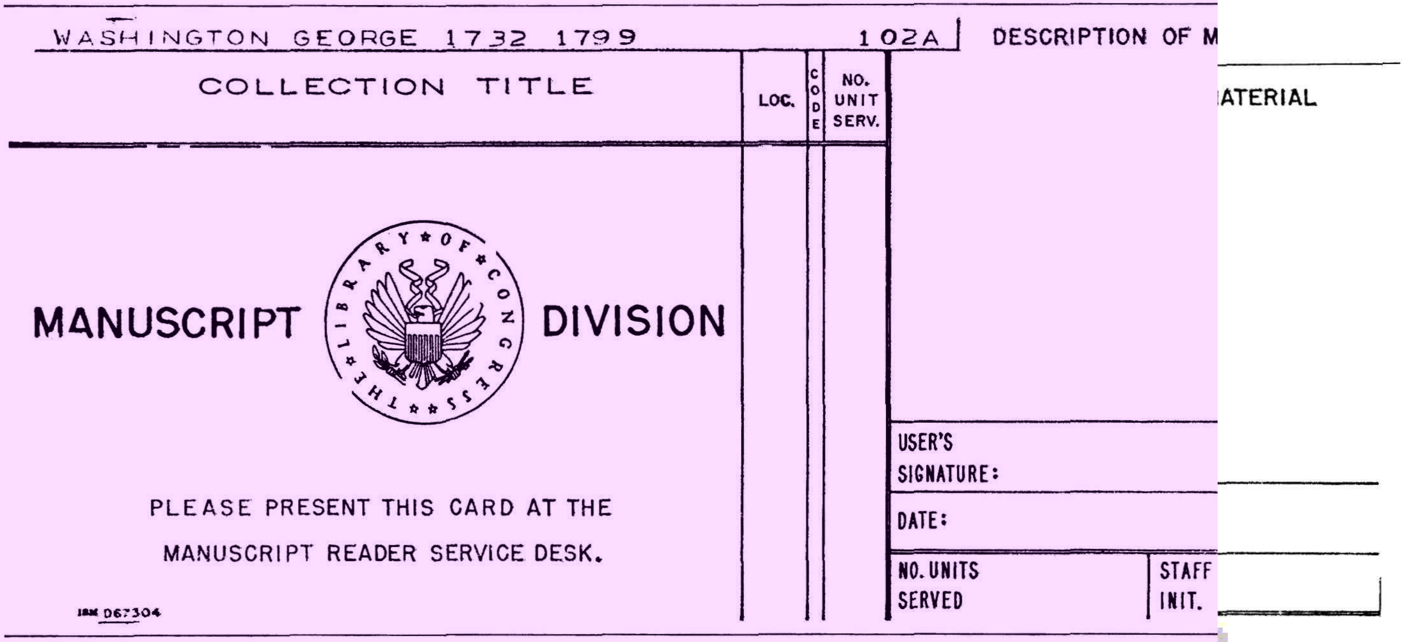
COLLECTION BULK DATES		PROFESSION/OCCUPATION			GEOG. LOC.	SHELF LOCATION	RSTN	RSTN LIFT DATE	FIND AID	CASE FILE	LIT. RTS	STA-TUS
BEG.	END	1	2	3								
		MUSICIAN	EDUCATOR	ABOLITIONIST		212P	R		F	C	U	G

NUCMC NUMBER	MICRO FILM	MICROFILM FOOTAGE	PHOTO STATS	XEROX	ENL PRINT	OTHER COPIES	NUMBER OF SHEETS	BOUND VOLS.	SOURCE/DONOR 1	ACCESS-ION	PROC CAT	SOURCE/DONOR 2	ACCESS-ION	PROC CAT	RECENT COLLECTION USE
MS59- 150								12	BLEDSOE C W	13177	01	CAMPBELL M D	10216	01	3

SOURCE/DONOR 3	ACCESS-ION	PROC CAT	ADDITIONAL ACCESSIONS MAXIMUM OF FOUR				REMARKS
			1	2	3	4	
DAVIS EDNA	9875	01					

EXHIBIT I

A sample of two collection entries as they would appear when extracted from the master record of manuscript collections, on a form carrying 44 of the 98 elements for each collection listed.



WASHINGTON GEORGE 1732 1799 | 102A | DESCRIPTION OF MATERIAL

COLLECTION TITLE	LOC.	CODE	NO. UNIT SERV.

MANUSCRIPT DIVISION

THE LIBRARY OF CONGRESS

PLEASE PRESENT THIS CARD AT THE
MANUSCRIPT READER SERVICE DESK.

USER'S SIGNATURE:

DATE:

NO. UNITS SERVED | STAFF INIT.

IBM 067304

Exhibit 2

Obverse of a prepunched call slip for the papers of George Washington, showing collection title (with birth and death dates) as well as shelf location (102A). The card is filled out by researchers when requesting manuscript material; the back of the card filled out by staff members of the Manuscript Division, serves as the actual chargeout and return record.

50 classes of information are gathered, including payroll number, social security number, department designation, wage rate, leave rates for annual and sick leave, deductions schedule for insurance, taxes, and bonds, and other pertinent information. In its application to manuscripts, substitutions were made so that the central figure is the person around whom the collection is formed (*e.g.*, Thomas Jefferson) ; around this person some 98 classes of information are gathered, including collection number, NUCMC number, shelf location, size, and amount of use.

Although the sort-and-list program for all these descriptive and statistical items (a total of over 270,000 items of information for all 3,000 collections) provides much information about the physical characteristics, use, and care and handling of the material, it does not provide for retrieval of content data—subjects and correspondents. In order to bring out this information, the Division settled on a combination of two programs. Fundamental to both was the decision that the methods employed for retrieving information should not initially require the reprocessing or even the reanalysis of any collection already processed for which a finding aid has been prepared. The next consideration was to avoid adding any further work that would slacken the processing pace of an already committed processing staff.

The Division conceives of processing to consist of four major steps, each dependent upon the other, sometimes with two or more being taken simultaneously. These steps are analysis, selection, arrangement, and description. A collection is analyzed before processing to determine if it is in disarray and, if so, to decide what arrangement is best suited to the papers. Analysis continues during arrangement with an eye to writing a description of the finished collection. Selection consists of separating forms of material (maps, photographs, printed works, etc.) as well as multiple copies (near-print or carbon) of documents for disposition or transfer to appropriate custodial divisions within the Library. Arrangement consists of assuring the proper grouping of material in series or subseries that seem natural for the collection, reestablishing disarranged chronology, or merely imposing order on chaos. When analysis, selection, and arrangement are complete, the collection is described in a register (roughly equivalent to the preliminary inventories issued by the National Archives). The register in its present form at the Library has developed from the format established around 1952 by Katherine E. Brand, then Head of the Recent

Manuscripts Section of the Manuscript Division.[8] It consists of a preface (or note on provenance), a brief biographical sketch of the person whose papers it describes, a scope-and-content note, a series description, and a container list. The register is to a collection what the preface, introduction, and table of contents are to a book. It indicates what one can expect to find and in some cases what will not be found in a group of papers. The Division currently has registers for more than 500 of its collections. Although this represents only one-sixth of the number of collections, the 500 cover almost half of the 30 million items in the Division. The register became the key to the second phase of the automation project.

The description of each collection progresses from the specific to the general. Depending on the arrangement of the papers, each folder carries a label describing its contents in general terms. These are cumulated into a container which, in the register, bears a title describing all the folders within it. Containers of associated material are gathered into a series; and a number of series, described together, make up the basis for the scope-and-content note. Thus, if one were to analyze each container of a collection, the scope-and-content of the collection would be known. Since they follow standard format, all of the more than 500 registers in the Manuscript Division are composed of these elements: provenance note, *curriculum vitae,* scope-and-content note, series description, and container list.

It was decided that these registers would become the basis for indexing all processed collections for which registers have been prepared. Experimentation with a number of programs showed that the most rewarding approach to subject- and name-oriented container lists was the application of a modified KWIC program, which has been named the SPINDEX (Selective Permutation Index). The mechanics are simple. A key-punch operator takes each register and translates the container lists verbatim into punched-card format. Subject listings receive one code number (*2* in this instance), while names receive another (*1*). Listings of form only ("correspondence, 1920–1925") receive a third code. A contraction of the collection title, collection identification number, sequential card number, and container number are also punched in the card (see Exhibit 3). Employment of this method entails little or no editing of the registers and no going back to the collections to analyze them. This means that the speed with which all 500 collections that have registers can be indexed depends only on the speed of a

[8] Katherine E. Brand, "The Place of the Register in the Manuscripts Division of the Library of Congress," in *American Archivist,* 18:59–67 (Jan. 1955).

```
PRATT ADMIRAL                                            002TAYLOR 50062  1006
ADMIRAL PRATT - INFORMAL LETTERS AND REPORTS             002TAYLOR 50062  0005
JAPANESE AND CHINESE FIGHT FOR CONTROL OVER SHANGHAI     001TAYLOR 50062  2004
TAYLOR#S OUTLINE OF ACTION DURING SHANGHAI INCIDENT      001TAYLOR 50062  2003
CHIEFS OF NAVAL OPERATIONS - REPORTS TO                  001TAYLOR 50062  2002
CORRESPONDENCE AND REPORTS 1931-1933                     001TAYLOR 50062  0001
```

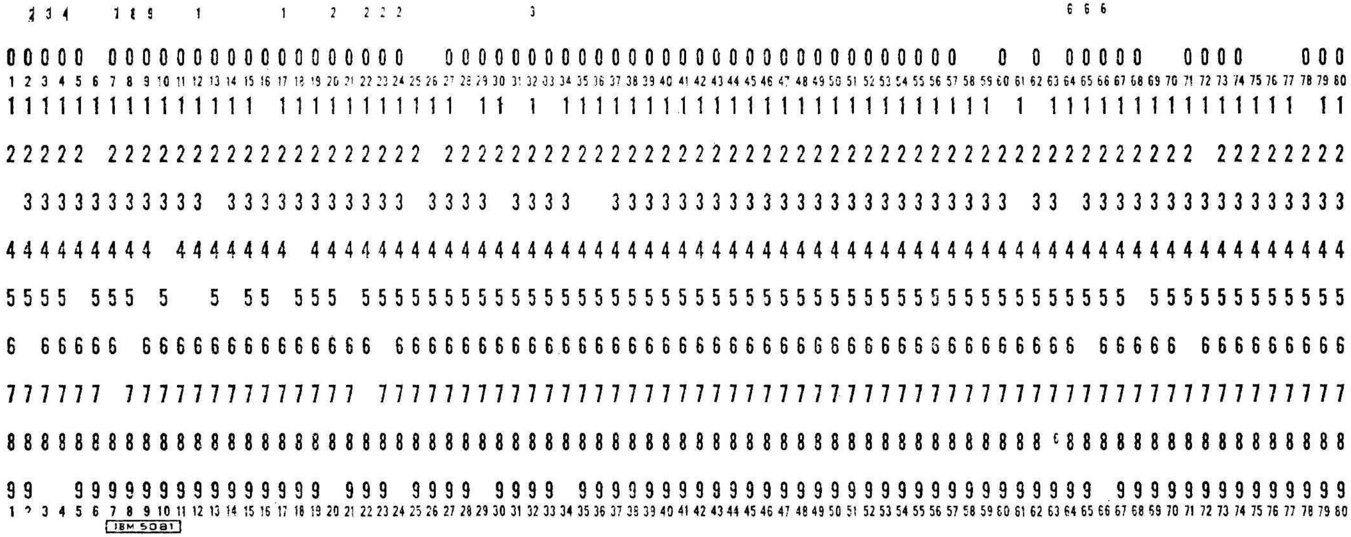


EXHIBIT 3

The first six cards, as key-punched, from the papers of Admiral Montgomery Meigs Taylor, indicating entry, container number, collection name code, collection identification number, code for form of material (o), name (r), or subject (2), and sequential card number.

key-punch operator plus the speed, almost incidental, with which the computer system can process the information.

The SPINDEX program provides for indexing each word in a title (unless one or more of them have been "stopped" purposefully). Container 6 of the records of the Daniel Guggenheim Fund for the Promotion of Aeronautics, for instance, contains a number of folders, including one labeled "Byrd Antarctic Expedition." This entry, from the container list, is translated into machine-readable form through the preparation of a punched card, and all the cards from the collection are subjected to the program. The SPINDEX program prints each keyword in its context, followed by the container number, abbreviation of the collection title, and the collection number (each of the Division's collections has been assigned a 5-digit number). In this instance, none of the words are stopped. The print-outs appear in alphabetical order along with other keywords:

```
ANTARCTIC
    BYRD ANTARCTIC EXPEDITION                 6 GUGGEN 17699
BYRD
    BYRD ANTARCTIC EXPEDITION                 6 GUGGEN 17699
EXPEDITION
    BYRD ANTARCTIC EXPEDITION                 6 GUGGEN 17699
```

If a few hundred registers are so indexed, the references to the Antarctic and Antarctica, to expeditions and explorations, and to Admiral Byrd will all file together or in proximity to each other, indicating collection title and container numbers. An extract from a sample of this program produced the following entries under the word AIRCRAFT:

```
                                                          [collection
                                    [container   [title    identification
                                       no.]   contraction]     no.]
AIRCRAFT
    AIRCRAFT DEVELOPMENT AND           228      ARNOLD        11189
        HISTORY
    AIRCRAFT INDUSTRIES ASSOCIA-         8      EAKER         19331
        TATION
    EUROPEAN AIRCRAFT MANUFAC-           8      GUGGEN        17699
        TURERS
    GUIDED MISSILES AND PILOT-         254      ARNOLD        11189
        LESS AIRCRAFT
```

Although incomplete, the above sample is an indication of the type of indexing received when the container lists for the papers of

Henry H. ("Hap") Arnold, Ira C. Eaker, and the Daniel Guggenheim Fund were subjected to one program.

A portion of one page printed from the SPINDEX program, given below, illustrates the intermixing of names and subjects from 10 registers of collections relating to aeronautics, indicating keyword, keyword in context, container number, collection name code, and collection identification number for the papers of Washington I. Chambers, the Daniel Guggenheim Fund for the Promotion of Aeronautics, Inc., Alfred Hildebrandt, and the Wright Brothers:

```
EXHIBITIONS                    CONTINUATION
  AVIATION EXHIBITIONS—GHENT, BELGIUM—1913  032  HILDEBR  25916
  AVIATION EXHIBITIONS—ILA BERLIN  1928     031  HILDEBR  25916
  AVIATION EXHIBITIONS—OLDENBURG, GERMANY
    1927                                    032  HILDEBR  25916
  AVIATION EXHIBITIONS—PRAGUE, CZECHOSLO-
    VAKIA 1927                              032  HILDEBR  25916
EXPEDITION
  ACCIDENTS—AIRSHIP ITALIA ON NORTH-POLE
    EXPEDITION                              001  HILDEBR  25916
  ARCTIC EXPEDITION PHOTOGRAPHS             044  CHAMBER 50799
  BYRD ANTARCTIC EXPEDITION, 1928–1929      006  GUGGEN   17699
  EXPEDITION TO ANTARCTICA                  003  WRIGHT   46706
  GREELEY RELIEF EXPEDITION                 037  CHAMBER 50799
  GREELEY RELIEF EXPEDITION                 044  CHAMBER 50799
  NICARAGUA CANAL SURVEY EXPEDITION         041  CHAMBER 50799
  NICARAGUAN SURVEY EXPEDITION IN 1884 & 1885 037 CHAMBER 50799
  SURVEY EXPEDITION IN NICARAGUA            005  CHAMBER 50799
  UNIV OF MICHIGAN GREENLAND EXPEDITION,
    1926–29                                 012  GUGGEN   17699
```

With similar information from a few hundred registers in the computer system, a number of possibilities present themselves. One is that the computer can reconstitute container lists from the index, so that a container list of only the Guggenheim Fund papers could be called for. This will be especially convenient when an "on-line" computer system is installed in the Library, allowing remote use of the computer record. Another possibility is paragraphing, and thus composing a very rough scope-and-content note, in those cases where only a container list has been compiled. If the Guggenheim Fund container list reads as follows (with appropriate continuation):

```
  AIRSHIP INSTITUTE                      4   GUGGEN  17699
  FOG FLYING RESEARCH                    5   GUGGEN  17699
  WESTERN AIR EXPRESS                    6   GUGGEN  17699
  EUROPEAN AIRCRAFT MANUFACTURERS        6   GUGGEN  17699
  DYNAMIC METEOROLOGY                    6   GUGGEN  17699
```

it is possible to eliminate box number and collection codes, precede the paragraph with a predetermined general statement, and print out:

THE MAJOR SUBJECTS TO WHICH THESE PAP

ANDERSON, DR. S. HERBERT
ASSOCIAZIONE ITALIANA DI AEROTECNICA
6. BYRD ANTARCTIC EXPEDITION
CALIFORNIA INSTITUTE OF TECHNOLOGY
[etc.]

[from master manuscript record]

PROCESSED BY LRG. SPECIALIST: SCIENCE
COLLECTION NO. 17699. NUCMC NO. MS62–697
SHELF LOC. 245P

All of this information could then be re-formated by the computer to produce a workable catalog entry (in card or page form) that would conform to the Library of Congress rules for cataloging manuscript collections, with some variations for local cataloging practices and needs (such as addition of shelf location, accession numbers, etc.). The catalog format might appear as follows:

[from master manuscript record]

245P 21 [*i.e.*, shelf loc. & container count]
1S MS62–697 [processing stage, Specialist, NUCMC card no.]
17699 [collection ident. no.]
DANIEL GUGGENHEIM FUND FOR THE PROMOTION OF AERONAUTICS, INC.
RECORDS [1926–1930]
6020 ITEMS
AERONAUTICS
PHOTOSTATS—150 SHEETS

[from master index record]

GENERAL CORRESPONDENCE, FUND RECORDS, LEGAL AND FINANCIAL PAPERS, R. H. MAYO MATERIAL

THE MAJOR SUBJECTS TO WHICH THESE PAPERS REFER ARE: AIRSHIP INSTITUTE, FOG FLYING RESEARCH, WESTERN AIR EXPRESS, EUROPEAN AIRCRAFT MANUFACTURERS, DYNAMIC METEOROLOGY [etc.]

PERSONS REPRESENTED AS SUBJECTS OR CORRESPONDENTS INCLUDE: ANDERSON, DR. S. HERBERT; LEVINE, CHARLES A.; LINDBERGH, CHARLES A.; CLARK, VIRGINIUS C.; WOOLARD, PROFESSOR E. W.; YOUNGER, CLARENCE M.; [etc.]

[from master manuscript record]

PUBLISHED REGISTER. CASE FILE. LITERARY RIGHTS DEDICATED. PROCESSED 1962. GIFT H. F. GUGGENHEIM. AC. 12863.

SEE ALSO: COLL. 24315, THE PAPERS OF HARRY FRANK GUGGENHEIM.

These formats are only hypothetical, but the ability to provide them automatically is not. In sum, from the many individual items of information supplied from a multitude of files and other sources, the computer is capable of synthesizing and formating a rough register and catalog entry. These would win no Pulitzer Prize for literature, but they are working tools and can be the basis for a more literary product for publication. The production of rough registers and catalog entries is only one aspect of a comprehensive automation program. Others include special bibliographies for subject areas or chronoligical periods (*e.g.*, the American Revolution), or even a complete handbook of the Division's holdings. This means that the only thing necessary for the production of registers or other descriptive forms for the Division's 2,500 unregistered collections is the information from the master manuscript record and the container lists.

The Manuscript Division is not item indexing all its collections, nor is it folder indexing them. Rather, it is preparing an automated index to its registers. This imposes no change on the methods of arranging material. The processor has nothing to do with the automation project, except that he is instructed to make the registers that he prepares on new collections as representative of the true nature of the collections as possible and not to stint in his description of the material that he processes. As a result, better registers are being written.

The one problem now facing the automation program in the Manuscript Division is posed by the chronologically arranged collections. The container lists for these provide merely the inclusive chronology for each container. The only subject and correspondent analysis is in the scope-and-content note of the register. Here again two solutions are possible for gaining at least minimal control over the collection contents. One is to have an editor indicate in each scope-and-content note the subjects and names by underlining them, and then have the key-punch operator translate the information into punched cards for the SPINDEX program, omitting any reference to container numbers since such information is not applicable. This solution would at least file the subjects among their counterparts in the master index and would refer the researcher to the collection. From that point, however, he would have to resort to traditional research methods to find the material in question. An alternative to this method would be to adapt some of the automatic abstracting/indexing programs (such as IBM's SYNTRAN) to the diversities of scope-and-content note language. This method might be difficult to initiate, but once perfected it could be used on any scope-and-content note in any register in any repository. The Manuscript Division is currently exploring the practical uses of this method.

The Division plans to include collections of corporate records in the automation program. Such records, arranged according to hierarchically archival standards, pose no problem, since the machine process is capable of operating at any level, from broad control of all holdings to item or subject indexing. It will probably be desirable, in some archives, to bring indexing down to the series level rather than the container level, with only certain series receiving more intensive indexing. This is common practice at the National Archives, as reflected in the format and contents of that agency's many preliminary inventories. Slight modifications of the Manuscript Division's program could produce a machine-generated association of all related records within a repository (or a number of repositories), and this procedure could easily develop relationships from agency to agency or trace the responsibility of one function (*e.g.*, public health) over a long chronological span. The computer thus bridges the gap between the stratified physical structure of records and the idiosyncratic intellectual structures imposed or desired by the researcher.

All the programs discussed above have been tested or are functioning in the Manuscript Division of the Library of Congress, and they are all compatible with one another. The master manuscript record provides the descriptive characteristics of a collection; the machine then matches (through the collection identification number) this information with the subject-correspondent analysis of the master index record. The two can then be printed jointly in any desired format. After all collections are processed, the machine follows through and indexes all this information, referring to the collection title and number and, when desired, to the series or container number for each indexed name or subject, including the miscellaneous items or groups of papers in the Division. It is planned that most print-outs from these programs will be selective, according to a subject area or chronological period, but a full run of all of the programs would result in over 50,000 computer-produced pages of information about the Division's 3,000 collections.

If one wants a list of finding aids to collections in the Manuscript Division, the program can provide one. If a researcher desires to see all the collections relating to Revolutionary War Army officers or to 20th-century explorers, the information can be retrieved through this program. If the Librarian wishes to know how often the George Washington or Abraham Lincoln papers were removed from the shelves during a specific period, and for what purposes, the program can provide the statistics. If a journalist is writing about Jenkin Lloyd Jones, the program can provide a list of all the collections in which Jones' name appears in the register and can indicate in what context Jones is mentioned and in which container in each collection the material will be found.

Furthermore, if a staff member wants to determine as of this week how many of the 100,000 containers on the Division shelves need new labels (either as a total figure or broken down by collection title), the information is available on request; and the computer can be instructed to print the correct number of labels, with collection title and box number for each. If a European archivist asks how much material the Library has copied from the Austrian Haus-, Hof-, und Staatsarchiv in its foreign copying program, the record can supply him with the number of pages of hand transcripts, sheets of photostats, and reels and footage of microfilm and can give information to show what portion of the Archiv has been copied.

None of this information has been created solely for the computer program; it has always been available in one form or another in the Division's files, catalogs, or indexes. It was sometimes difficult to find, and locating 10 items of information about a collection might have necessitated consulting as many internal sources. This complexity is now being simplified, and heretofore unrealized relationships among collections are being revealed. As new collections are added to the Division's holdings, information about them will automatically be integrated with that about older holdings in the master manuscript record and the master index record, and the records will thus be continually updated. As current files are brought under complete control, a future stage of the project may call for reanalysis of selected collections and closer indexing—perhaps at the folder or item level.

The purpose of the Manuscript Division's program, and of the other programs in this field, is solely to free both researchers and staff members from tedious searching for material about which a record has already been made and to establish relationships among collections or materials within collections. It is hoped that removing the mechanical tedium from the researcher's shoulders will free him to spend his time more valuably, in reading and analyzing documents pertinent to his goals. As the record of information grows, and as the ease of access to it is still further improved by technological improvements in the computer system, automation will supply the keys necessary to unlock more and more compartments in the Nation's greatest treasure house of manuscript sources for the documentation of American life and letters.

Most of the systems described in this paper were designed to answer the specific needs of the repositories using them. The variety of approaches, both at different levels and within each level, however, indicates the versatility of automation as applied to manuscript and archival work. The adaptations to local conditions are endless; the universality of concepts is obvious. It is neither too early nor too late for archivists and manuscript curators to begin serious consideration of the adaptation of the electronic machine to their work methods and research problems.