Random Sampling Techniques: A Method of Reducing Large, Homogeneous Series in Congressional Papers

ELEANOR McKAY

As the twentieth century progresses, some archival institutions are becoming overstocked with masses of documentation from people and institutions important to the operations of our society. The archival profession no longer has sufficient staff and space to winnow this material, item by item. Even our most important leaders leave us with such a legacy of documentation that item-by-item weeding is generally no longer a feasible way to reduce bulk and produce a collection of manageable size. By any definition, the elected leaders of the U.S. government—the President, the senators, the representatives—are worthy subjects of research as they reflect and shape the operations and values of our society. This paper concerns a method for reducing the bulk of the papers of members of the legislative branch, thereby making the papers more usable for research.

Members of Congress are elected to represent local interests in the national forum and to make legislation applicable to the nation as a whole. Their papers contain information on issues and problems of both regional and national concern; consequently these papers are useful to scholars from a wide variety of disciplines and specialized interests. Archivists trained in history, as many of us are, must expand their horizons to be aware of the needs of many different kinds of historians as well as the needs of scholars from the full spectrum of the humanities and social sciences. Our interest in furthering research, however, must not blind us to the realities of archival space and staff limitations; not all the millions of pieces of constituent mail or congressional case files can be saved. Even compressed onto reels of microfilm or computer tape, this information will someday fill up the stacks and become too abundant for the research projects of individual scholars. An incredible amount of staff time and money will be spent on the preparation of thorough finding aids and on conversion to new forms of mechanical storage as the old forms become obsolete.[1]

It has been estimated that a single researcher, financed out of his or her own pocket, can handle a probability sample of about 2,000 cases; while a better-funded dissertation or faculty research project can use up to 20,000 cases.[2] To-

The author is curator of the Mississippi Valley Collection, Memphis State University. She prepared this paper for presentation to the annual meeting, the Society of American Archivists, at Salt Lake City, October 6, 1977.

[1] Meyer Fishbein, ed., *The National Archives and Statistical Research* (Athens: Ohio University Press, 1973), p. 75.

[2] James K. Benson, "Sampling Techniques in Archival Management and Quantitative Research" (Paper prepared for the Minnesota Historical Society, summer 1976), p. 2.

Downloaded from http://meridian.allenpress.com/american-archivist/article-pdf/41/3/281/2746528/aarc_41_3_t756k0723551xg3.pdf by India user on 25 August 2022

day there are 435 representatives and 100 senators generating and retaining papers at an enormous rate. Richard Baker, the historian of the U.S. Senate, indicates that although senators have free and unlimited microfilming services available, many choose either to destroy case files or to leave constituent mail to a successor.[3] A recent *SAA Newsletter* indicated the arrangements our current senators have made for the disposal of their papers.[4] The institutions selected to receive senatorial papers will soon have to grapple with awesome problems familiar to the staff of the National Archives and Records Service which has developed rigorous and systematic procedures for the retirement and disposal of governmental records.[5] Assuming that the papers of elected officials are really public property, a senator or member of Congress creates governmental records just like any appointed member of the federal government; and the techniques worked out by NARS for the treatment of federal records are applicable to these records which until recently have been considered private manuscript collections.[6]

For those not familiar with actual statistics on the staggering bulk of records of twentieth-century congressmen and senators, here are a few examples. In 1975 the State Historical Society of Wisconsin received more than 400 cartons from Senator Gaylord Nelson, representing his most recent decade in office. Another Wisconsin collection is the papers of Senator Alexander Wiley, which comprises 805 Hollinger boxes, 36 volumes, 12 tape recordings, 22 phonodiscs, and 2 films, mainly from the period 1938–62. The Minnesota Historical Society has accessioned the papers of Hubert H. Humphrey; included in additions since 1975 have been 156 cartons of constituent mail with form-letter responses from Senator Humphrey, 1974–77. The society also holds the papers of Congressman Clark MacGregor, including over 30,000 letters, telegrams, and petitions from Minnesota constituents on political issues in the decade of the 1960s. The Mississippi Valley Collection at Memphis State University received 152 cubic feet of papers from the single congressional term of George W. Grider, 1965–67. The papers of his successor, Dan H. Kuykendall, 1967–71, take up 104 cubic feet in the stacks. The Western Historical Manuscripts Collection at the University of Missouri in Columbia has processed 357 Hollinger boxes of the papers of Senator James P. Kem, 474 boxes from Senator Thomas C. Hennings, Jr., 145 boxes from Congressman Albert S. J. Carnahan, and 893 boxes from Congressman Thomas B. Curtis. This totals 1,869 Hollinger boxes, each 5 inches wide, and is in addition to the unprocessed papers of Senator W. Stuart Symington and three other elected federal officials who were in office in the 1950s and 1960s.[7] The list could go on and on.

[3] Richard A. Baker, Historian of the U.S. Senate, letter to Eleanor McKay, 25 April 1977.

[4] "Senators Tell Plans for Deposit of Their Papers," *SAA Newsletter* (November 1976), p. 14. See also U.S. Senate Office of the Secretary, Historical Office, "Senators' Papers—Locations of Collected Papers of United States Senators who left the Senate between 1947 and 1977" (prepared by Kathryn A. Jacob and Leslie Prosterman, April 1977), 11 pp.

[5] Ken Munden, ed., *Archives and the Public Interest: Selected Essays by Ernst Posner* (Washington, D.C.: Public Affairs Press, 1967), pp. 134–35.

[6] J. Frank Cook, " 'Private Papers' of Public Officials," *American Archivist* 38 (July 1975): 299–324; T. R. Schellenberg, *The Management of Archives* (New York: Columbia University Press, 1965), p. 65.

[7] Patrick M. Quinn, "Manuscripts Register to the Papers of Alexander Wiley, 1913–1967," State Historical Society of Wisconsin, 1970; Lydia Lucas, Assistant curator of manuscripts, Minnesota Historical Society, telephone interview with Eleanor McKay, 1 August 1977; James K. Benson, "Let-

Downloaded from http://meridian.allenpress.com/american-archivist/article-pdf/41/3/281/2746528/aarc_41_3_t756k0723551kg3.pdf by India user on 25 August 2022

Not every institution is burdened with large numbers of congressional collections, but just one or two collections can choke the regular procedures of an institution. As an example, consider the impact on my own institution if we succeed in our negotiations for the papers of a contemporary senator. If we get them, the university must allot to the archives one or perhaps two additional floors in the old library building to store the unprocessed collection. The senator may have a long career in Congress, and we will not process his collection until that career is completed or we have absolutely run out of dead storage space. When the Mississippi Valley Collection was established in 1964 it was organized as a rare book and historical manuscripts collection, and its procedures for arrangement and description of manuscripts were geared to small collections of individually valuable items. Like other archives and manuscripts repositories, our institution—with two congressional collections and the possibility of an accession of senatorial papers—has been compelled to develop different techniques for handling collections of great size and historical importance which contain few individual documents of specific research value or autograph interest in their own right.

Recognizing that this is the plight of many institutions acquiring twentieth-century collections, the Midwest Archives Conference sponsored a special seminar in August 1975 on the topic of subject access and control of large manuscript collections. At that seminar, archivists from six midwestern institutions[8] formed a task force to study the congressional collections at their institutions, to compare their findings, and to make suggestions for handling these collections cooperatively. No final report has yet been published. The task force wanted to determine if these collections were dominated by constituent mail on political and legislative issues or by case files requesting congressional aid in dealing with federal agencies of the executive branch. Another question concerned how traditional research techniques for examining individual documents could apply to bulky, homogeneous series of case files or constituent mail. Even though letters from constituents represent real, personal problems, future scholars simply would not and could not be vitally interested in the details of each case, or even the most significant or atypical ones. Finally, the task force needed to suggest procedures (1) to make enough of these records available to scholarship, (2) to present a true picture of the function of a typical congressman's office, (3) to serve the various secondary avenues of research from numerous disciplines, and (4) to make possible future research inquiries with interests unthought of today.[9]

In 1976 Lynn Wolf Gentzler at the Western Historical Manuscripts Collection of the University of Missouri at Columbia analyzed their holdings of seven collections of senators and representatives from Missouri in the 1950s and 1960s.

ters to Congressmen as Sources for Research" (Paper prepared for the Minnesota Historical Society, summer 1976), pp. 2–3; Robert D. Bohanan, "Manuscripts Register to the Papers of George W. Grider, 1964–66," Mississippi Valley Collection, Memphis State University; Lynn Wolf Gentzler, unpublished report on Midwest Archives Conference seminar on subject access and control of large collections with analysis of Western Historical Manuscripts Collection's papers of twentieth-century U.S. congressmen and senators, Western Historical Manuscripts Collection, University of Missouri, 28 April 1976, pp. 4–5.

[8] The institutions included the State Historical Society of Wisconsin, the University of Iowa, the University of Missouri, the Lilly Library at Indiana University, the Newberry Library, and the Minnesota Historical Society.

[9] T. R. Schellenberg, *Modern Archives: Principles and Techniques* (Chicago: University of Chicago Press, 1965), p. 156.

Downloaded from http://meridian.allenpress.com/american-archivist/article-pdf/41/3/281/2746528/aarc_41_3_t756k0723551xg3.pdf by India user on 25 August 2022

Four of these were processed; three were unprocessed. She found that the contents (mostly constituent mail) were similar and included (1) service cases in which the congressman intervened with a federal agency for the constituent; (2) requests for varied materials such as flags, publications, and autographs; and (3) correspondence on pending legislation and current issues. This material mainly related to public pressure on the elected official from the local constituents and was arranged in various ways within each collection. It was often interspersed with documents relating to the congressman's interaction with political colleagues, along with committee files, reference material, government documents, bills and related material, personal correspondence, and miscellany. Upon analyzing the breakdown of the four processed collections (see Table) totalling 1,869 Hollinger boxes, Gentzler identified about 60 percent of the documents as some sort of routine constituent correspondence. Within the collections, she identified legislative correspondence, service cases, requests, reading files, invitations, and service academy applications. Investigation of recent research use of the four processed collections indicated almost no use of service case files or of invitations, but it was found that several researchers had used the service case files in conjunction with legislative correspondence. There had been no research use of the service academy applications. Gentzler concluded that processing guidelines could be drawn up to weed major portions of these collections without endangering research potential.[10] It is true that future research may depend heavily on materials neglected in the mid-twentieth-century. But the problem here is that the great bulk of twentieth-century congressional files makes it almost impossible for researchers to find individually valuable items. These papers, if retained unweeded, will also take up stack space that could hold other manuscript material of more definite scholarly interest.

TABLE: Analysis of Four Processed Political Collections, 1950–69, held by the Western Historical Manuscripts Collection, University of Missouri at Columbia[11]

COLLECTION	*Legislative Correspondence*	*Service Cases*	*Requests*	*Reading File*	*Invitations*	*Academy Applications*	*TOTAL**
Senators:							
KEM (357 boxes)	113	42	11	60	4	5	235
HENNINGS (474 boxes)	145	87	5	43	16	6	302
Representatives:							
CARNAHAN (145 boxes)	64**		10	—	4	4	82
CURTIS (893 boxes)	222	142	11	175	15	6	571

* Total boxes of constituent-related material.

** Combination of legislative correspondence and service cases.

[10] Gentzler, report on Midwest Archives Conference seminar, pp. 4–7.
[11] Ibid., p. 5.

Downloaded from http://meridian.allenpress.com/american-archivist/article-pdf/41/3/281/2746528/aarc_41_3_t756k0723551xg3.pdf by India user on 25 August 2022

The problem of selecting material for retention is indeed a difficult one. Any decision is likely to be controversial. However, appraisal decisions form the essence of an archivist's professional responsibility. The archivist is responsible for the preservation of non-current documentation for legal, administrative, and scholarly uses. As Schellenberg wrote in 1965, "Archival principles and techniques should be developed with one basic objective in mind, i.e., to make the archivist more effective in servicing his material." This profession must develop its own discipline, its own standards, from which we will benefit by improving our efficiency of arrangement and description and by which we will contribute to the control of the documentary resources of this nation. The commonly offered excuse that we cannot yet select those documents of scholarly importance begs the question. Stacks glutted with constituent correspondence of questionable value preclude repositories from accessioning other collections of more certain research value.[12]

As part of the Midwest Archives Conference task force, Christine Rongone and I conducted an anlysis of the papers of federal, state, and local officials, papers in the custody of the State Historical Society of Wisconsin. Like Gentzler, we concluded that the bulk, similarity, and low research value of most documents in these collections obligated us to develop a method for weeding without ruining the collections. Several months followed of research and discussions with scholars from the history, political science, economics, and sociology departments at the University of Wisconsin; with other archivists and the History of Wisconsin project staff of the Historical Society; with Senator Nelson's staff; and with archival colleagues from the Midwest Archives Conference task force. Finally, Rongone and I decided to apply random sampling tecniques to weed large, homogeneous series down to 20 percent of their original volume. The statisticians and quantifiers assured us that this would preserve a large enough universe of information for most research uses.[13] The plans for this experiment have not yet been implemented.

Charles Palit, of the University of Wisconsin Economics Department, an expert in explaining statistical principles and applications to humanists, examined the papers of Senator Nelson, and several congressional collections, with Rongone and me. He concluded that the retention of as much as 20 percent of a large, homogeneous series like constituent mail would preserve a large enough universe for most statistical studies conducted in the future. Scholars doing a statistical study often need as little as a 2 or 3 percent sample from a total body of records, and Palit considered that they could obtain a sample valid for their studies from the 20 percent random sample we might do. He did caution archivists, however, to be very careful in the initial appraisal of a series and in making the decision to weed by random sampling. But once the decision has been made that a series is bulky, homogeneous, and repetitive, and that it offers little different information from document to document, he gave assurances that a 20 percent sample would be sufficient for most quantitative researchers.

[12] Schellenberg, *Management of Archives,* pp. 74–78.

[13] The sampling experiment that Christine Rongone and I designed was cancelled after I left the State Historical Society of Wisconsin for new duties in Memphis. I am happy to report that a similar experiment, designed with the advice of experts in sampling and statistical techniques, was undertaken at Memphis State University.

Downloaded from http://meridian.allenpress.com/american-archivist/article-pdf/41/3/281/2746528/aarc_41_3_1756k0723551xg3.pdf by India user on 25 August 2022

Many archivists and historians fear that wonderfully detailed letters from constituents, reflecting more than just the basic topic the constituent writes about, may be destroyed by weeding. Study of extensive runs of this correspondence indicates, however, that these meaty letters form a small portion of the total series. And a truly random sample allows 20 percent of such letters to be saved, just as it reduces the routine mail by a uniform 80 percent. Barbara Hinckley, of the UW Political Science Department, agreed with Palit's analysis. From her perspective as a quantitative researcher studying public opinion, she asked archivists just to prepare lists of how many letters were pro and con on a given issue; she had little interest in studying individual documents. Diane Lindquist, of the UW History Department, agreed with a colleague, a distinguished intellectual historian, that scholars would benefit from a reduction of bulky, modern series, even if some potentially valuable documents were destroyed. This historian, who prefers to remain anonymous, confessed to having once pulled files at random from the case files of a New Deal agency to document the generalizations that his previous research suggested about the impact of that agency on private citizens. The vastness of the series discouraged him from studying it at any length.

As a precautionary measure, Rongone and I decided to hold discarded material (80 percent of the total collection) for a further subjective sort for significant or atypical items. If random sampling proves a valid weeding tool, it should be possible to abandon this subjective sorting.

Our intended experiment was to apply the random sampling techniques to a large political collection at Wisconsin that was both bulky and homogeneous. Field service archivists had already judged the accessioned material to be of research value, so our task was to refine further their appraisal of certain series which had a low ratio of information content to bulk. Our definition of bulk was subjective; in the Gaylord Nelson case it included hundreds of cartons. By *homogeneous* we meant that the collection included material of similar nature with few or no extraneous items. Our definition of an item was a single document or group of documents fastened together by the congressman's staff for filing purposes. We would number all items in a container, usually a Hollinger box, in sequence from 1 to x. Using a table of random numbers from the Rand Corporation's *A Million Random Digits* (1955), we chose a number from which to begin selecting. Mathematics scholars have developed the random number table to assure that no subjective prejudice colors selection; randomly selected, every number has an equal chance of selection. By beginning the selection at a different randomly selected number, archivists can avoid the danger of destroying records reflecting cyclic trends, such as events occurring every two, five, or ten years.

Each page of this book contains lists of numbers which have been mathematically selected to follow one another in a truly random pattern, one which reflects no subjective quirk in their relationship. Although the lists give five-digit numbers, one can use the list on a problem containing numbers with fewer than five digits. For this experiment, we decided to use only two-digit numbers, using the last two numbers from the right side of the five digits. Further, one can select a number from any point on a page from this book, then select the next numbers in a vertical, horizontal, or diagonal pattern.

Many statisticians advise that the sample begin with the decision about what

Downloaded from http://meridian.allenpress.com/american-archivist/article-pdf/41/3/281/2746528/aarc_41_3_t756k0723551xg3.pdf by India user on 25 August 2022

portion of the whole body of records to retain. For example, if only 10 percent of a given collection can be housed in the stacks, design the sample around that fact. But to preclude using weeding as an administrative convenience, Rongone and I intended to select series for bulk and homogeneity, and then weed them down by 80 percent by random sampling techniques. The following possibilities for weeding were suggested by Lynn Wolf Genzler:

Form requests from organizations, for copies of bills.
Requests from individuals for copies of bills.
Library of Congress lists of books borrowed by staff members.
Unsolicited membership directories of organizations.
Constituent requests for information on state problems, especially if the member of Congress expressed no opinion.
Stationery Room accounts.
Plane tickets.
Telephone and telegraph bills.
Service academy appointments.
Reading file.
Clippings.
Meeting notices of committees.
Statements before committee hearings if essentially the same as printed record.
Miscellaneous files—"A Miscellaneous . . .," etc.
Army-Navy football game ticket requests.
Consent calendar, private calendar.
Whip notices.
Office equipment requests.
Form letters sent to newly naturalized citizens.
Flag requests, autograph requests, etc.
Post-card campaigns and other bulk mailings of constituent opinion.

This would preserve for researchers an adequate universe of information for most research purposes.[14]

Weeding by means of random selection allows the retention of a typical run of homogeneous records in a large series. An item-by-item selective sample generally saves atypical or significant documents, but this type of sample must be done by a professionally trained archivist familiar with the subject field and other collections; and it presents researchers with a skewed picture of the actual operations of the creator or creating office. However, once the archivist has made the decision concerning those series to be weeded, and has determined what percentage of the whole will be retained, a paraprofessional or student assistant can easily be trained to do the actual weeding using random sampling techniques. That person can also make a list of all folder titles, whether material from an individual folder is retained or not. Until the profession feels comfortable with random sample weeding, the archivist with the subject knowledge and the experience with this sort of collection can examine the 80 percent discards and selectively retain atypical or significant items for permanent retention. The researcher can readily discern these items, from those randomly selected, by a special identifying mark on the subjectively selected items. In addition, the find-

[14] Paul Lewinson, "Archival Sampling," *American Archivist* 20 (October 1957): 308–9; R. Michael McReynolds, "Statistical Sampling of Archives" (Paper presented at the SAA annual meeting, Philadelphia, 2 October 1975), pp. 13–15; Gentzler, report on Midwest Archives Conference seminar, p. 9.

Downloaded from http://meridian.allenpress.com/american-archivist/article-pdf/41/3/281/2746528/aarc_41_3_t756k0723551txg3.pdf by India user on 25 August 2022

ing aid must describe clearly the procedure of the random sample and describe the identifying marks on the later, subjectively selected items.

After leaving the State Historical Society of Wisconsin for the curatorship of the Mississippi Valley Collection, I decided to apply the technique of random sample weeding to the papers of former-Congressman George W. Grider, who represented the Memphis, Tennessee, area from 1965 to 1967. He had donated to the MVC 60 cartons totalling 152 cubic feet of congressional files, and he specifically permitted us to discard duplicate and weeded materials. Processed in 1976, the Grider collection was maintained as the former congressman's staff had arranged it, although some material which had been misfiled or simply dumped by the office staff was removed to a more logical file. The processed collection includes five series:

(1) Washington office files, arranged by year and then alphabetically by subject, 11 cu. ft.
(2) Closed case files, arranged alphabetically by constituent's last name, 2 cu. ft.
(3) Pending case files, arranged by year and then by constituent's last name or by government agency, 3 cu. ft.
(4) 1964 and 1966 campaign files, arranged by year and alphabetically by subject, 12 cu. ft.
(5) Miscellaneous legislative files, arranged alphabetically by subject concerning Grider's opinions, actions, and research on pending legislation, 5 cu. ft.[15]

In the judgment of the MVC staff, the pieces of documentation in this collection did not have great individual research value; and so each series was weeded using random sampling techniques to retain only 20 percent of the whole. Analysis of the 80 percent selected for discard revealed few unique and valuable items. In fact, none of the 80 percent was retained following the subjective analysis.

Now, what if every archival institution did reduce certain series and record groups by 80 percent? Researchers might find that the different collections retained at various archives offered few new insights from collection to collection. Archivists and manuscript curators might want to consider cooperative retention policies concerning certain series of types of files in congressional papers containing information mainly important in the aggregate. After careful study, institutions could decide on cooperative collection development of particular records of national or more local significance, thereby freeing other institutions to discard their series on that topic. For example, it may not be necessary to save requests for bills or academy applications in every congressman's records.[16] Libraries have already pioneered inter-institutional collection development in such areas as rare books, special subject areas, and periodical holdings.

However, this will be an area for the highest professional judgment. For instance, applications for appointments to post office positions may be of little value on the surface, yet they may reveal much about the unique local patronage scene. We may be tempted to discard all invitations declined by congressmen and senators, yet a study of trends may reveal that changes in the acceptance of invitations from various special interest groups is significant. At one time certain officials would not bother to attend functions of Blacks, or of women, but that time has changed.

[15] Bohanan, "Manuscripts Register to the Papers of George W. Grider, 1964–66," pp. 1–4.
[16] Gentzler, report on Midwest Archives Conference seminar, p. 6.

Downloaded from http://meridian.allenpress.com/american-archivist/article-pdf/41/3/281/2746528/aarc_41_3_t756k0723551xg3.pdf by India user on 25 August 2022

We must be careful not to be too dogmatic about our appraisal decisions, but we must make them. As Margaret Norton wrote, "It is comparatively easy to select records of permanent value, relatively easy to decide on those of no value. The great bulk of records are borderline."[17] However, as Lynn Wolf Gentzler suggests, we must consider action as a profession to keep our stacks and our time from being wasted on preserving records of little or no individual value. We must develop professional standards to improve the efficiency of arrangement and description and to provide better control for the documentary resources of the nation. Scholars having a realistic view of limitations of interest, energy, and funds will thank us for having the courage to make these difficult decisions.

[17] Thornton W. Mitchell, ed., *Norton on Archives* (Carbondale: Southern Illinois University Press, 1975), p. 240.

Downloaded from http://meridian.allenpress.com/american-archivist/article-pdf/41/3/281/2746528/aarc_41_3_t756k0723551txg3.pdf by India user on 25 August 2022